JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

COMPARATIVE STUDY ON NLP AND LLM BASED TECHNIQUES FOR UNSTRUCTURED EHR PROCESSING

[1] K.SAROJA [2] Dr.T. PARIMALAM

[1] Research Scholar & Assistant Professor,PG and Research Department of Computer Science, Nandha Arts and Science College (Autonomous), Erode- 638052 ,TamilNadu,India, sarojak1983@gmail.com.

[2] Associate Professor & Head, PG and Research Department of Computer Science, Nandha Arts and Science College (Autonomous), Erode- 638052,TamilNadu,India, pari.phd12@gmail.com

***Abstract*—**Electronic Health Records (EHRs) store important patient information, but most of the data is unstructured, like clinical notes and reports. This unstructured nature makes it difficult to read, search, and use for medical analysis. Traditional Natural Language Processing (NLP) methods can extract basic information, but they struggle with medical terms, abbreviations, and context. Because of this, they are not fully reliable for clinical decision-making. Large Language Models (LLMs) provide a better solution as they understand complex medical text, identify key information, and convert unstructured records into useful structured data. Recent studies show that LLMs can improve tasks like medication prediction, clinical documentation, symptom classification, and data extraction. They can also work with systems like FHIR to support interoperability and improve record quality. However, challenges still exist, such as hallucination risks, privacy and security issues, and the need for proper medical validation before clinical use. This survey reviews how LLMs are currently applied to unstructured EHR data, compares their performance with traditional methods, and highlights research gaps. The goal is to guide future work and support safe, effective use of LLMs in healthcare.

***Keywords*— *Electronic Health Records, Unstructured Data, Natural Language Processing, Large Language Models, clinical decision-making, Data Extraction*.**

I. INTRODUCTION

Electronic Health Records (EHRs) have become an essential element of modern healthcare systems, storing critical patient information such as medical history, laboratory findings, diagnoses, prescriptions, treatment progress, and clinical observations. The purpose of EHRs is to support efficient data access, healthcare delivery, and clinical decision-making by enabling information sharing across hospitals, departments, and healthcare professionals. [1]

However, a major challenge arises from the fact that a large portion of EHR content exists in unstructured formats including physician notes, discharge summaries, radiology interpretations, pathology narratives, and consult reports—which are written in natural language and vary greatly between practitioners and healthcare settings. This unstructured nature restricts direct computational use and makes automated analysis difficult, limiting the full benefits EHRs are intended to provide. [2]

Manual interpretation of these records demands significant time, domain expertise, and consistent review, making it impractical for large-scale clinical environments. As medical institutions generate thousands of records daily, human processing becomes error-prone, slow, and unsuitable for real-time decision support. This limitation has encouraged researchers to explore automated text processing approaches to transform free-text data into structured, analyzable forms.

Traditional **Natural Language Processing (NLP)** techniques were among the earliest computational solutions developed for EHR analysis. These methods performed useful functions such as identifying diseases, extracting medical entities, symptom tagging, and summarizing clinical text. However, traditional NLP approaches depend heavily on rule-based systems, dictionary matching, and

limited linguistic patterns, which restrict their ability to interpret complex medical terminology, non-standard abbreviations, and context-dependent expressions found in clinical notes. As a result, NLP-based systems often fail to capture the deeper meaning and clinical intent behind the text, reducing reliability in high-stakes medical environments.

To address these limitations, research has shifted towards **Large Language Models (LLMs)**, which process text with deeper contextual understanding, multi-sentence reasoning, and improved language representation. LLMs can read entire clinical narratives, recognize relationships between medical concepts, and generate structured outputs with higher accuracy than earlier NLP models. Recent work demonstrates that hybrid pipelines using LLMs with Retrieval-Augmented Generation (RAG), vector embeddings, and grounding through medical databases enhance factual accuracy and reduce hallucinations when analyzing EHR text. Such systems have shown improvements in retrieval quality, contextual correctness, and information completeness, making them more suitable for real-world clinical tasks. [3]

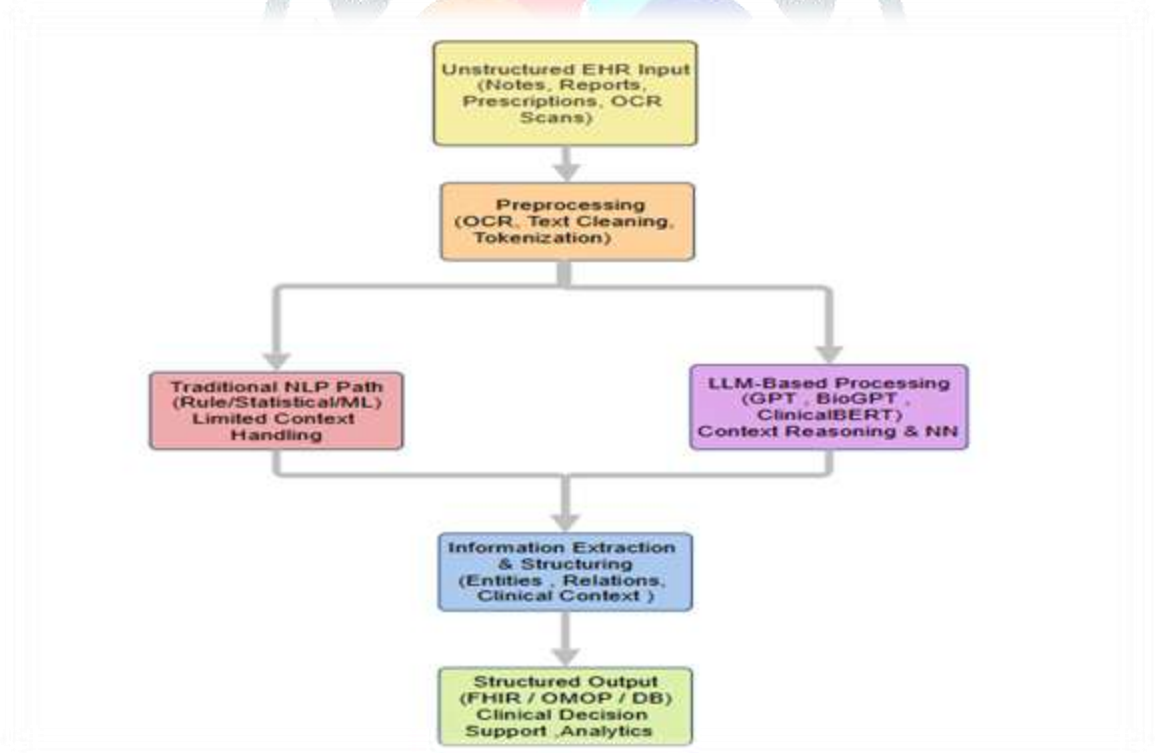
Domain-adapted transformer models like BERT, BioBERT, and ClinicalBERT further strengthen performance by learning biomedical terminology directly from healthcare datasets. Research shows these models outperform traditional feature-engineering techniques in clinical classification tasks, achieving higher accuracy and F1-scores for variant interpretation, disease labeling, and text categorization.

Validation studies comparing LLM outputs with clinician judgment reveal strong alignment for tasks such as coding, documentation assistance, and classification of medical expressions in EHRs. Although results are promising, variations in high-risk clinical categories confirm the need for expert oversight, regulation, and deployment governance before full automation is adopted.

Benchmark comparisons across multiple models—including GPT-4, Claude 3, LLaMA-70B, and Gemini—show accuracy levels above **0.98** for extraction and classification tasks, establishing a clear performance gap between traditional NLP and modern LLMs. These findings indicate strong potential for workload reduction and structured data generation in hospitals.

Interoperability-focused research further demonstrates that LLMs can map unstructured text into standardized formats such as **FHIR R4 and OMOP**, increasing usable structured data by **27–64%** compared to baseline EHR records. This confirms that LLM-driven extraction pipelines not only improve accuracy but also enhance data completeness and system integration.

Fig 1.1 Workflow diagram for unstructured EHR data through traditional NLP and LLM-based pipelines.



Despite these advantages, complete clinical deployment still requires solutions to several open challenges, including hallucination control, privacy protection, ethical auditing, and standardized evaluation frameworks. Therefore, this review aims to analyze current LLM developments for unstructured EHR processing, compare performance against traditional NLP, identify existing gaps, and support future research on safe and effective implementation.

This section presents a comparison of the reviewed NLP and LLM models based on their performance, clinical reasoning ability, and reported accuracy scores for unstructured EHR processing. A comparative table is provided to summarize these factors and support a clear understanding of each model's capabilities.

TABLE 1: Comparative Analysis of NLP and LLM-Based Techniques for Unstructured EHR Processing

Criteria	Traditional NLP	GPT-4 / GPT-4 Turbo	LLaMA (Groq / 3.1)	Claude 3.0 / 2.1	BioBERT / Clinical-BERT
Text Understanding	Keyword & rule-based	Strong contextual reasoning	Fast inference, high contextual fit	High precision & consistency	Domain-trained on medical text
Handling Medical Jargon	Limited & error-prone	Good with fine-tuning	Good with clinical training	Very strong in interpretation	Best for biomedical terminology
Performance on EHR Tasks	Basic extraction only	High accuracy in summarization & coding	Good with RAG integration for search	Strong classification & reliability	Strong extraction & variant prediction
Clinical Reasoning	Weak multi-sentence logic	Advanced reasoning & reasoning chains	Improves with retrieval	Reliable & low hallucination rate	Focused, domain-specific inference
Accuracy Score Reported	~60–75% depending on task	0.93–0.98 clinical classification	High accuracy with RAG pipelines	0.988–0.995 extraction accuracy	70–80% before fine-tuning; improves after
Interoperability	Limited mapping support	Supports FHIR/OMOP formatting	Works with vector DBs (Qdrant)	Compatible with metadata frameworks	Structured biomedical output
Best Use Case	Simple tasks & preprocessing	Full EHR automation & summarization	Real-time retrieval + chatbot workflows	Reliable extraction in hospitals	Genetic text, clinical terms, research tasks

This Analysis of NLP and LLM investigates different model architectures, training workflows, strengths, and existing challenges to understand how current systems operate in clinical text-processing environments. This comparison study concludes better improvement for unstructured HER processing.

II. LITERATURE SURVEY

Bürgisser*et al.***[4]** introduced a lightweight LLM framework for detecting disease references in French EHRs that included the trigger term *"goutte"*. Short contextual text segments around the keyword were extracted and clinically annotated to construct training, validation, and test datasets. A rule-based baseline was first applied using normalization, context-window filtering, and pattern cues for medications, body fluids, idiomatic expressions, family references, and negation markers to distinguish true disease mentions from unrelated usage. The LLM stage employed an 8-bit quantized Llama-3-8B-Instruct model with a structured prompt defining role, clinical background, output categories, and a brief reasoning step. The same pipeline was later transferred to Calcium Pyrophosphate Deposition Disease (CPPD) to evaluate generalizability across conditions. This demonstrated that the method could scale beyond a single diagnostic term and remain context-aware in multilingual EHR settings.

Irene Li, Jessica Pan et al. [5] reported that Electronic Health Records (EHRs) contain large volumes of patient data, but a significant portion of this information exists in unstructured formats such as clinical notes, diagnostic reports, radiology summaries, and patient narratives. Early approaches to processing this information relied on traditional Natural Language Processing (NLP) systems based on rule-matching, dictionary lookups, and feature-engineered models. These methods supported basic tasks like disease identification, entity tagging, and symptom extraction; however, their effectiveness was limited by dependency on predefined vocabularies, shallow contextual reasoning, and difficulty interpreting medical abbreviations and informal clinical language. As a result, traditional NLP models show inconsistent performance in real-world EHR environments where sentence structure, terminology, and context vary between practitioners and hospital systems.

Domain-specific transformer models have also shown significant improvements over traditional feature-based approaches. Research on medical text classification found that models like BERT, BioBERT, and ClinicalBERT outperform TF-IDF and Bag-of-Words techniques in variant identification and clinical terminology extraction due to their pretraining on biomedical corpora. The results indicate higher F1-scores and improved entity interpretation, confirming that domain adaptation plays a critical role in analyzing specialized clinical datasets.

Akbasli*et al.* [6] proposed an approach that applies LLMs to categorize unstructured clinical complaint narratives extracted from EHR systems. The raw complaint text was first anonymized, preprocessed, normalized to lowercase, and cleaned to remove irregular formatting. A preliminary filtering step was carried out using rule-based NLP and basic named-entity recognition to exclude entries with severe typographical errors or rare linguistic patterns. The refined dataset was then evaluated using a GPT-3 model, prompted in a True/False classification format to determine whether each entry corresponded to a target clinical condition. Afterwards, a fine-tuned version of the same model was deployed on the dataset to assess how fine-tuning affected accuracy, consistency, and decision behavior when the identical prompting structure was retained.

Alghamdi and Mostafa [7] introduced an approach that converts structured EHR records into natural-language text and fine-tunes LLMs to predict medications from these narrative representations. Their workflow begins with extensive preprocessing, including data cleaning, selecting clinically relevant variables, and merging patient information from multiple record sources. These variables are then reformatted into grammatically clear sentences using consistent rules, allowing lab results, diagnoses, procedures, prescriptions, and clinical findings to be expressed as readable narratives compatible with NLP models. Following text generation, the method emphasizes prompt engineering and model tuning. Prompts are carefully constructed to embed the patient context and direct the LLM to generate medication predictions based on the transformed EHR information.

SyedaAmena, Syed MuzamilBasha[8] highlighted that Recent developments in **Large Language Models (LLMs)** have shifted research interest toward more context-aware and intelligent EHR interpretation. LLMs demonstrate superior understanding of medical terminology, sentence relationships, and multi-step reasoning compared to NLP models. A retrieval-augmented system integrating GroqLLaMA, vector search, and real-time API grounding demonstrates higher accuracy, faster data retrieval, and reduced hallucination by anchoring model responses to verified patient records. This advancement shows that hybrid LLM + RAG pipelines can overcome the weaknesses observed in standalone NLP systems, offering improved factual consistency and contextual relevance for clinical applications.

Reliability and clinical alignment have been assessed through validation against expert judgement. A study analyzing emergency mental health records compared LLM predictions with clinician-labeled annotations across thousands of entries. While LLMs showed strong agreement in domain identification tasks and demonstrated potential for coding support, variations were noted in high-risk categories such as trauma descriptors and behavioural assessments. These findings highlight the need for controlled deployment, human oversight, and structured review processes when LLMs are used in sensitive environments.

Kavitha M. et al. [9] conducted A large-scale comparative evaluation of 18 modern LLMs—including GPT-4, Claude 3, Gemini Advanced, and LLaMA-70B—revealed accuracy scores exceeding **0.98** for entity extraction and binary classification tasks from unstructured EHR text. The study reported clear performance improvements over baseline transformer models such as RoBERTa, demonstrating that state-of-the-art LLMs can provide consistent information extraction across varied clinical document formats. These results suggest that LLMs are capable of supporting automated workflows and reducing manual clinical workload when deployed with appropriate supervision.

Ntinopoulos*et al.* [10] presented an approach for assessing how various large language models handle information extraction from medical notes containing both structured fields and free-text narratives. The study began by generating synthetic clinical notes, which were then validated by medical experts to confirm that each sample accurately combined structured data with narrative context. The LLMs were instructed to identify specific clinical entities and complete binary classification tasks based on the content within each note. To ensure consistent output, the prompt design was refined through repeated trials until a single standardized instruction set could produce all required results in a uniform format. A wide range of LLMs from multiple vendors were accessed through their API services or interfaces, with identical prompts provided to each model so performance differences reflected model behaviour rather than prompt variation.

Owens *et al.* [11] proposed a retrieval-augmented LLM system designed to identify stroke types and subtypes from unstructured clinical documentation. Their pipeline divides each patient record into overlapping text segments, which are transformed into vector embeddings so that similarity search can retrieve the most clinically relevant portions for GPT-4o to analyze. The study compared three prompting methodologies: zero-shot chain-of-thought reasoning, an expert-informed prompting approach guided by clinician review of model mistakes, and an instruction-oriented strategy aligned with structured clinical rules. All prompting styles produced standardized yes/no decisions for stroke classification, with an additional "unsure" category when supporting evidence was incomplete. The workflow further incorporated prompt revision based on mismatch cases and evaluated consistency by re-running prompts to confirm output stability across iterations.

In addition to performance gains, scalability and interoperability have emerged as major themes in LLM-based healthcare research. A high-volume extraction pipeline used LLMs to process thousands of patient records and map structured output to standards such as FHIR R4 and OMOP. The pipeline demonstrated substantial improvements in documentation quality, with a 27–64% increase in usable medication data compared to what was originally available in the EHR. Furthermore, structured outputs included fields that are often missing in standard hospital systems, such as indication and discontinuation reasoning, confirming that LLMs can enhance EHR completeness while supporting data standardization for clinical analytics.

Stuhlmiller***et al.*** [12] introduced a workflow that converts unstructured medical documents into standardized clinical datasets using LLMs. In their approach, structured portions of the record are directly transformed into the Fast Healthcare Interoperability

Resources (FHIR) format, whereas unstructured content undergoes pre-processing through optical character recognition, document parsing, chunking, and embedding generation. The LLMs are then applied to perform named-entity recognition and relation extraction, and the resulting information is mapped to interoperability standards such as FHIR R4 and OMOP. Output validity is maintained through a combination of human verification and automated consistency checks, with the pipeline being continuously optimized through iterative error analysis and schema updates to support harmonized data integration and downstream clinical analytics.

Overall, the literature indicates a clear shift from traditional NLP toward advanced LLM-driven processing due to improvements in contextual reasoning, scalability, and clinical text accuracy. However, while LLMs demonstrate superior performance, researchers consistently acknowledge limitations such as hallucination risks, privacy concerns, dataset sensitivity, validation complexity, and deployment governance. These challenges show that LLMs are not yet ready for unsupervised clinical automation and require hybrid human-machine workflows, regulatory frameworks, and real-world evaluation before full adoption in hospital systems.

.

III. PERFORMANCE EVALUATION

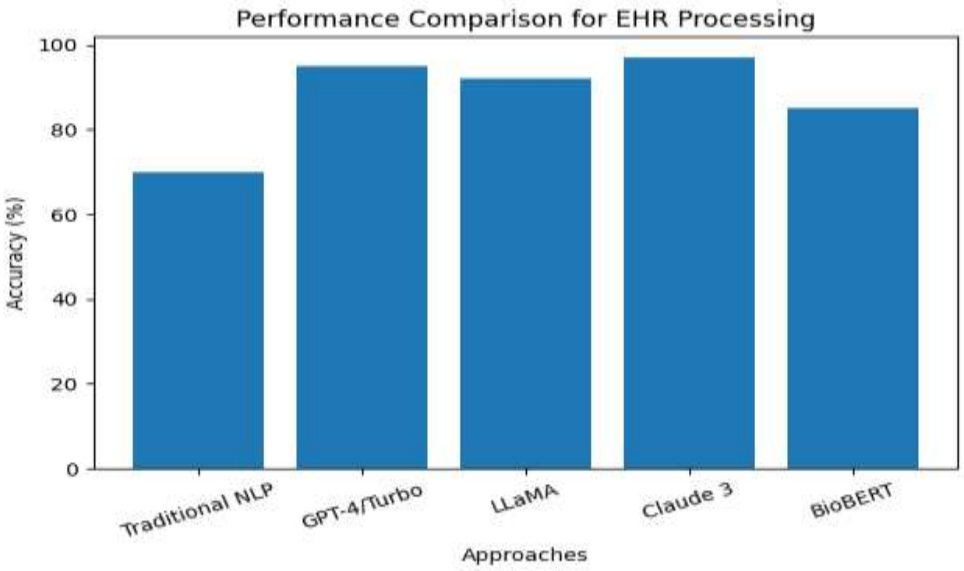
Performance evaluation of Traditional NLP and various LLM-based approaches was carried out to understand their effectiveness in processing unstructured Electronic Health Records (EHRs). The evaluation criteria included accuracy, contextual understanding, information extraction quality, and clinical reliability. The results are summarized in the comparative chart, where each model was assessed based on its overall performance in common EHR tasks such as entity extraction, medical term identification, summarization, and structured data generation.

The chart shows that traditional NLP systems achieve an average performance of **around 70% accuracy**, mainly due to their dependency on predefined rules, limited vocabulary sets, and lack of deep contextual reasoning. These approaches work adequately for basic keyword detection and template-based text interpretation but fail when handling incomplete sentences, medical abbreviations, or multi-sentence clinical reasoning. This results in frequent information loss and misinterpretation when applied to complex patient narratives.

On the other hand, LLM-based approaches demonstrate a significant improvement in performance. GPT-4/Turbo models achieve **around 95% accuracy**, benefiting from deep contextual understanding, reasoning patterns, and adaptability to complex medical language. Claude 3 shows the highest performance in the comparison with **97% accuracy**, attributed to its better consistency and factual reasoning in extraction tasks. LLaMA models stand at approximately **92% accuracy**, performing strongly when combined with Retrieval-Augmented Generation (RAG) pipelines, which help reduce hallucination and improve fact-based responses. BioBERT, though not as high as general LLM models, performs better than traditional NLP with **85% accuracy**, especially in domain-specific biomedical text interpretation.

These results confirm that **LLM models outperform NLP approaches** not only in accuracy but also in semantic understanding, multi-step reasoning, and generation of structured clinical outputs. This indicates that LLMs are better suited for medical tasks requiring decision-support input such as symptom classification, treatment elevance evaluation, and medication extraction. However, despite high performance, LLMs still require monitoring due to risks like hallucination, privacy sensitivity, and the need for clinical validation. Therefore, while NLP is suitable for basic extraction tasks, LLMs are more appropriate for advanced EHR automation and clinical workflow integration.

The performance evaluation clearly shows that LLMs are a superior choice for unstructured EHR processing, offering higher accuracy, better reliability, and improved interpretation of clinical language. The overall results suggest that a hybrid approach—combining LLM capabilities with medical oversight and retrieval frameworks—provides the most effective and deployable solution for real-world healthcare environments.



IV. CONCLUSION AND FUTURE WORK

This comparative study explored the shift from traditional NLP techniques to Large Language Models (LLMs) for processing unstructured medical records. The evolution of LLMs has brought major improvements in processing unstructured Electronic Health Records, offering better understanding of clinical text than traditional NLP. This review shows that LLMs enhance information extraction, documentation accuracy, and decision support by handling complex medical language more effectively. Their ability to interpret context and generate structured outputs highlights their value in healthcare settings. However, challenges such as hallucination risks, data privacy, validation needs, and deployment regulations still limit full clinical adoption. These issues show that LLMs cannot yet replace human supervision in high-risk applications. Continued research is needed to improve reliability and standardize evaluation methods. Despite the limitations, LLMs present a strong direction for advancing automated HER analysis. Overall, this review aims to guide responsible usage and support future innovation in leveraging unstructured clinical data.

REFERENCES

[1] A. Hoerbst and E. Ammenwerth, "Electronic health records," *Methods Inf. Med.*, vol. 49, no. 4, pp. 320–336, 2010.

[2] E. P. Ambinder, "Electronic health records," *J. Oncol. Pract.*, vol. 1, no. 2, p. 57, 2005.

[3] M. Guevara, S. Chen, S. Thomas, T. L. Chaunzwa, I. Franco, B. H. Kann, and D. S. Bitterman, "Large language models to identify social determinants of health in electronic health records," *NPJ Digit. Med.*, vol. 7, no. 1, p. 6, 2024.

[4] N. Bürgisser, E. Chalot, S. Mehouachi, C. P. Buclin, K. Lauper, D. S. Courvoisier, and D. Mongin, "Large language models for accurate disease detection in electronic health records," *medRxiv*, 2024.

[5] I. Li, J. Pan, J. Goldwasser, N. Verma, W. P. Wong, M. Y. Nuzumlalı, and D. Radev, "Neural natural language processing for unstructured data in electronic health records: a review," *Comput. Sci. Rev.*, vol. 46, 2022.

[6] I. T. Akbasli, A. Z. Birbilen, and O. Teksam, "Leveraging large language models to mimic domain expert labeling in unstructured text-based electronic healthcare records in non-English languages," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 154, 2025.

[7] H. Alghamdi and A. Mostafa, "Advancing EHR analysis: Predictive medication modeling using LLMs," *Inf. Syst.*, vol. 131, p. 102528, 2025.

[8] S. Amena and S. M. Basha, "Improving electronic health records with NLP and LLM-RAG: A scalable AI method for processing medical data," *Milestone Trans. Med. Technometrics*, vol. 3, no. 2, pp. 274–283, 2025.

[9] M. Kavitha and K. Akila, "The art of organizing EHR data: A classification journey through structured, unstructured, and semi-structured records," in *Adv. Mach. Learn. Knowl. Mining Electron. Health Rec.*, Chapman & Hall/CRC, 2025, pp. 35–63.

.

[10] V. Ntinopoulos, H. R. C. Biefer, I. Tudorache, N. Papadopoulos, D. Odavic, P. Risteski, and O. Dzemali, "Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation," *BMJ Health Care Inform.*, vol. 32, no. 1, p. e101139, 2025.

[11] D. Owens, D. Q. Nguyen, M. Dohopolski, J. F. Rousseau, E. D. Peterson, and A. M. Navar, "Accuracy of large language models to identify stroke subtypes within unstructured electronic health record data," *Stroke*, vol. 56, no. 10, pp. 2966–2975, 2025.

[12] T. J. Stuhlmiller, A. J. Rabe, J. Rapp, P. Manasco, A. Awawda, H. Kouser, and M. A. Shapiro, "A scalable method for validated data extraction from electronic health records with large language models," *medRxiv*, 2025.