ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

LLM-Driven Knowledge Extraction from EHRs

Anumandla Mukesh
Independent Researcher
mukeshanumand@gmail.com

Abstract

Large Language Models (LLMs) can facilitate structured knowledge extraction from unstructured clinical information sources, such as electronic health records (EHRs), clinical text notes, and clinical trial protocols. This capability bridging heterogeneous data modalities has the potential to power a broad range of downstream clinical decision support and data mining tasks.

Clinical informatics research often relies on specific types of information that are not readily available in explicitly structured form in the corresponding data repositories. These "structured knowledge" types are prevalent in a variety of clinical data sources, such as coded EHRs, but LLMs can offer a different, complementary approach. The task of mapping raw unlabeled text to structured knowledge types involves providing rich supervision during LLM training—using either fine-tuning or prompting mechanisms—to create and support structured knowledge extraction capabilities.

Keywords : Large language models, electronic health records, clinical informatics, natural language processing, named entity recognition, relation extraction, automatic reasoning, automatic summarization, region-based visual question answering, open information extraction.

1. Introduction

While Electronic Health Records (EHRs) hold much potential for clinical research, the narrative format they are typically stored in complicates their direct use for analysis. Attempting to extract knowledge in a structured form indirectly enables a multitude of classic tasks including but not limited to reason over clinical concepts, predict clinical events, and identify bias in models. While decision trees and code-based systems have traditionally handled multiple aspects of knowledge extraction (in particular named entity recognition), a reservoir of clinical data—inferred from primary or secondary analysis of EHRs—can re-enable training of modern Deep Learning models for the same set of tasks.

Recent results show the strong capability of large language models (LLMs) in accurately inferring entities such as conditions, medication, test results, procedures, and relations such as causation or chronology; the absence of de-identified data remains the main stopping point for real-life applications. Rapid progress makes the plausibility of employing LLMs for structured extraction in the near-term likely: decision tree losses such as early response time and low resource needs are continually shrinking, tools for knowledge transfer between unbalanced domains are maturing, and multi-source integration is surfacing in the literature (joining clinical functions such as imaging with natural language). Since LLMs are now addressing these areas, an overarching view on the full knowledge extraction pipeline is thus timely—shaping it to favor yet under-explored integration into real-life clinical systems.

1.1. Background and Significance

Repetitive free-text data in Electronic Health Records (EHRs) act as a rich knowledge resource for improving clinical use and assisting biomedical research. For automated access to this knowledge, it is essential to extract and structure the underlying

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

information contained in clinical texts. This requires solving a number of challenging information extraction (IE) tasks such as named entity recognition, relation extraction, and reasoning on events. Owing to the lengthy clinical narratives, the context-sensitive nature of semantic understanding, and the high diversity and richness of medical concepts involved, large language models (LLMs) trained on sizable multilingual data are emerging as highly competitive methods for these grounding tasks.

Despite their impressive capabilities, these models are often designed for general-purpose language tasks. As medical concepts are usually not present in the pretraining datasets or only appear very rarely, the factuality and performance of LLMs when applied directly to the medical domain are questionable. Domain adeptness can be improved by finetuning on in-domain data; nevertheless, access to large supervised datasets and associated expert labelers is a major hurdle for most high-stakes healthcare applications. Furthermore, even if these demands for sufficient in-domain knowledge can be fulfilled, applying a specialized model to a distinct, underlying-use scenario may lead to subpar generalization performance due to domain shift and cohort mismatch. Consequently, a number of studies have instead explored prompting LLMs without fine-tuning for a variety of biomedical IE tasks.

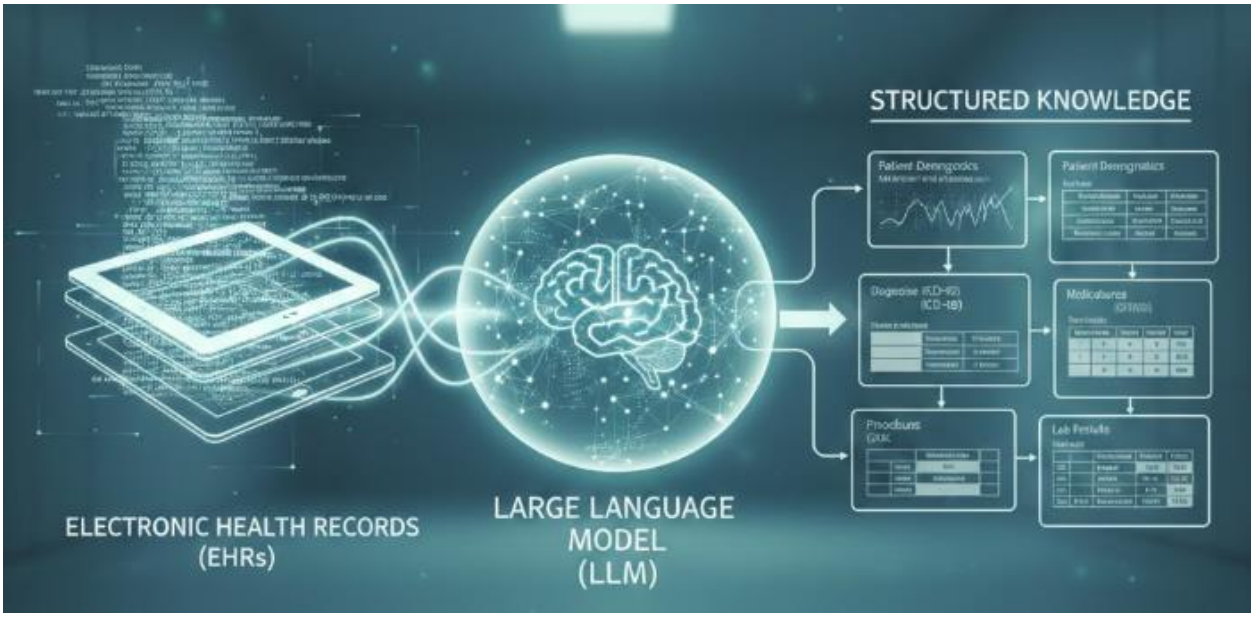


Fig 1: Large Language Models for Structured Knowledge Extraction

1.2. Research design

This work weighs evidence-based, objective claims about how large language models serve for structured knowledge extraction from electronic health records, according to specified criteria. Investigating the main contributions to natural language processing and reasoning abilities of large language models, together with access to an interpretation-explanation feedback layer, raises a suite of clinical information research questions. Those address whether the unprecedented language, reasoning, memory, and commonsense ability of these models can be harnessed to operate over free-text notes in electronic health records even if the model is not domain-adapted, whether knowledge—including named entities, fact, and event causality and chronology—can be

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

efficiently extracted for major applications areas, whether bias inherent to the model data budget is controllable, and whether a full deployment setup can be construed to allow real-time clinical use.

To answer these questions, the work proposes a methodological framework to map unstructured EHR text into structured knowledge and combines it with quality considerations aiming to compensate EHR labeling bias. The first part of the framework encompasses data preparation and privacy issues, details on model architecture choices and training paradigms, and discussion of mapping strategies using schemas and ontologies. In a second part, the interest pivots toward the core technical developments of the framework, exploring extraction techniques for named entities and relations—and supporting axes such as temporal or event-reasoning capabilities. Finally, the exposure surfaces the use facets surrounding the framework: data bias and representation gaps, privacy-preserving techniques, and domain adaptation and transferability. The framework thereby provides a structured-cut discourse on knowledge extraction from clinical text.

2. Background and Related Work

A detailed survey of the background context and related work justifies the approach taken. Electronic Health Records Capture a Rich Spectrum of Clinical Events Large Language Models Concluding a Knowledge Extraction Architectures Released in Recent Years with Profoundly Different Approaches and, Potentially, Capabilities to Execute These Tasks. Their Reasoning and Memory Skills Could Be of Great Benefit for Extracting Knowledge from All That Free-Text Data that Remains in the Public.

Electronic Health Records (EHRs) are increasingly employed as a source for the analysis of clinical decisions, disease evolutions, complex associations, and clinical trajectories due to their large size. However, EHRs are still predominantly used as a source of generating cohort studies. This is mainly due to their structure: while most information is encoded in EHRs in a structured manner, clinicians usually introduce a domain-specific free text in each encounter. This information could prove useful to go beyond cohort generation. Attempts to go beyond cohort generation have resulted in the release of new Knowledge Extraction (KE) architectures designed for EHRs. These approaches proposed independent models for NER, RE, and temporal association tasks, and produced multi-institutional solutions mapping the free-text data in the EHR to a substantially more structured form, allowing for queries that go beyond cohort selection, as well as the application of nuanced graph-based engines to explore knowledge in the EHR.

Independently, large domain-specific language models have been released for NER in EHRs, although without explicit evaluation of the NER accuracy. Recently, researchers have begun to explore the utility of prompting recent large open-source language models, developed with a different schema, to perform NER and relation extraction, although they only reasoned about a subset of the population and did not map it into knowledge bases. Concomitantly, prompting large language models have been applied to EHRs for multiclass classification, but have not yet been demonstrated to be successful in multilabel tasks such as NER.

Equation 1: Structured knowledge representation as triples

Large Language Models for Struc…

Formalization

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

- Let the extracted knowledge base be:

$$\mathcal{K} = \{(s, r, o)\}$$

where $s$is the subject entity, $r$the relation type, and $o$the object entity.

2.1. Electronic Health Records: Structure and Challenges

Electronic Health Records (EHRs) have established themselves as the gold standard for a patient's medical history. They contain a variety of health-related data in a digital format that are collected from various healthcare settings. EHRs are unique in that they are structured according to a predefined schema, capturing the most common elements of a patient's medical history, including medications, allergies, social history, past medical history, and family history. In addition, concerns surrounding the temporal nature of the data, such as when a diagnosis, medication, or procedure took place, can be captured through the effective use of temporal information. EHR datasets allow for a range of studies, from clinical trials to natural language processing. Nevertheless, they are complex, containing a number of different multimodal correlated components. These include multilingual free-text notes, tabular structured data, laboratory and imaging studies, medical procedures, medications, and nursing assessments, all of which can be leveraged for many different applications. Standardization of terminology is addressed through the use of ontologies (e.g., SNOMED-CT and LOINC), yet heterogeneity remains. Codified data allow for automated large-scale analysis, while free text provides more meaningful insights. However, multimodal datasets present problems, with free-text data being widely underutilized.

Despite EHRs becoming the birthright of any patient that steps into a western world hospital, “the surveillance capitalists do not ask our consent before gathering our data,” and the industry and research community at large have failed to implement genuine systems to protect patient data. Patient records remain with the institution, with only a small portion of records being shared with clinical researchers in a deidentified format. Even in these cases, the free-text notes are almost always removed, restricting the utility of available resources. The absence of user-centered safeguards hinders the exploration of important questions that could benefit from the availability of full dataset contents. Anonymization, differential privacy, and other privacy-preserving methods remain active fields of research addressing data release. Nevertheless, the deployment of such techniques introduces different types of bias, rendering the generated knowledge less representative of the original patient population. As a result, key aspects usually found in EHR free-text notes, such as rich temporal information, potential inequalities, and domain adaptation, remain neglected.

2.2. Large Language Models: Capabilities and Limitations

As large language models (LLMs) have exploded in popularity, so has interest in their ability to understand and reason about complex topics. Closely following this jump in visibility, two major areas of concern have been raised: 1) Can LLMs perform reasoning tasks with human-like levels of capability? 2) How well do LLMs really know things? Do they remember simple facts or specialized knowledge, or do they instead rely primarily on sophisticated surface-level pattern matching?

Many LLMs, including earlier models such as BERT, ExcelNet, and ERNIE, are pretrained in a purely masked manner, while others, such as GPT-3 or Turing-NLG, are pretrained using simple unidirectional language modeling. Although the two different objectives result in different underlying representations, combining the two techniques in a multi-stage approach has shown impressive advances: T5 is pretrained with a gap-filling objective, followed by an encoder-decoder objective, and Pegasus is pre-trained with an abstractive summarization objective. Newer systems such as PaLM and Chat-GPT combine a scaled-up

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

architecture for autoregressive generation with test-time interaction with a powerful retrieval-augmented pipeline, which handles long-context language-generation tasks with ease.

With the flexible nature of their architecture and the massively diverse corpora available, LLMs have opened new research directions in the NLP and ML domains. Despite this, the ability of LLMs to adapt to specialized tasks with limited task-specific fine-tuning continues to be a major question, along with whether scaling of LLMs leads to gains in performance across various tasks.

3. Methodological Framework for Knowledge Extraction

A broad range of techniques and models is available for mining clinical knowledge from unstructured sources, including Named Entity Recognition (NER) for clinical entities, Relation Extraction (RE) for connections between entities, and Reasoning for filling gaps in extracted information. Knowledge Extraction can also address specific EHR properties, such as support for temporality and annotation approaches that reserve parts of the input text for other steps—commonly, for providing context.

7696973 Balanced Knowledge Extraction requires careful examination of potential problems such as bias and generalizability, which are often intertwined. Data used for the detection of medical entities and relations need not be representative of the target EHR sentiment or ideas, but heterogeneous sources help provide balance; identify disparate sources of coverage (e.g., gender, ethnicity, social determinants); and calibrate or scope domain adaptation techniques. Unlike the representation of risk factors by medical codes, factors such as imaging data pose a higher coverage issue because they tend to be non-note multimodal sources of EHR risk stratum levels. A specific gap may also lead to generalizability problems; for example, advertisement of vacation trips may be frequent in some patient populations and should be noted as a risk factor in these cases.

Balanced knowledge extraction from unstructured clinical data requires more than accurate application of Named Entity Recognition (NER), Relation Extraction (RE), and reasoning techniques—it demands careful attention to bias, representativeness, and generalizability. Although training data for identifying medical entities and relationships need not perfectly mirror the target EHR population, reliance on narrow or homogeneous datasets can distort model performance and downstream insights. Incorporating heterogeneous data sources improves balance by broadening coverage across gender, ethnicity, socioeconomic status, and other social determinants of health, while also supporting calibration and domain adaptation. Coverage challenges are particularly pronounced for multimodal EHR components such as imaging data, which are often underrepresented in note-centric extraction pipelines. Gaps in representation can directly undermine generalizability; for instance, contextual factors like frequent vacation travel—common in certain populations—may signal specific exposure risks but remain unrecognized if not adequately captured in the training data. Therefore, robust clinical knowledge extraction must combine technical rigor with deliberate dataset curation and contextual awareness to ensure equitable and transferable performance across diverse patient populations. Balanced clinical knowledge extraction from unstructured data extends beyond optimizing the accuracy of Named Entity Recognition, Relation Extraction, and reasoning models; it requires an intentional focus on equity, representativeness, and real-world applicability. Models trained on narrow or demographically homogeneous corpora may exhibit strong benchmark performance yet fail to capture clinically meaningful patterns in underrepresented populations, leading to biased or incomplete downstream inferences. Curating heterogeneous training datasets that span diverse genders, ethnicities, socioeconomic contexts, and care settings helps mitigate these risks by expanding linguistic and contextual coverage while enabling more reliable calibration and domain adaptation. This need is especially acute for multimodal EHR elements such as medical imaging and associated reports, which are frequently excluded from text-centric pipelines despite their diagnostic

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

importance. Moreover, subtle contextual cues—such as travel habits, occupational exposures, or housing instability—can carry significant clinical implications but may remain invisible to models if absent from training data. Consequently, robust and generalizable clinical knowledge extraction must integrate technical sophistication with deliberate dataset design and contextual sensitivity to ensure fair, accurate, and transferable performance across diverse patient populations.

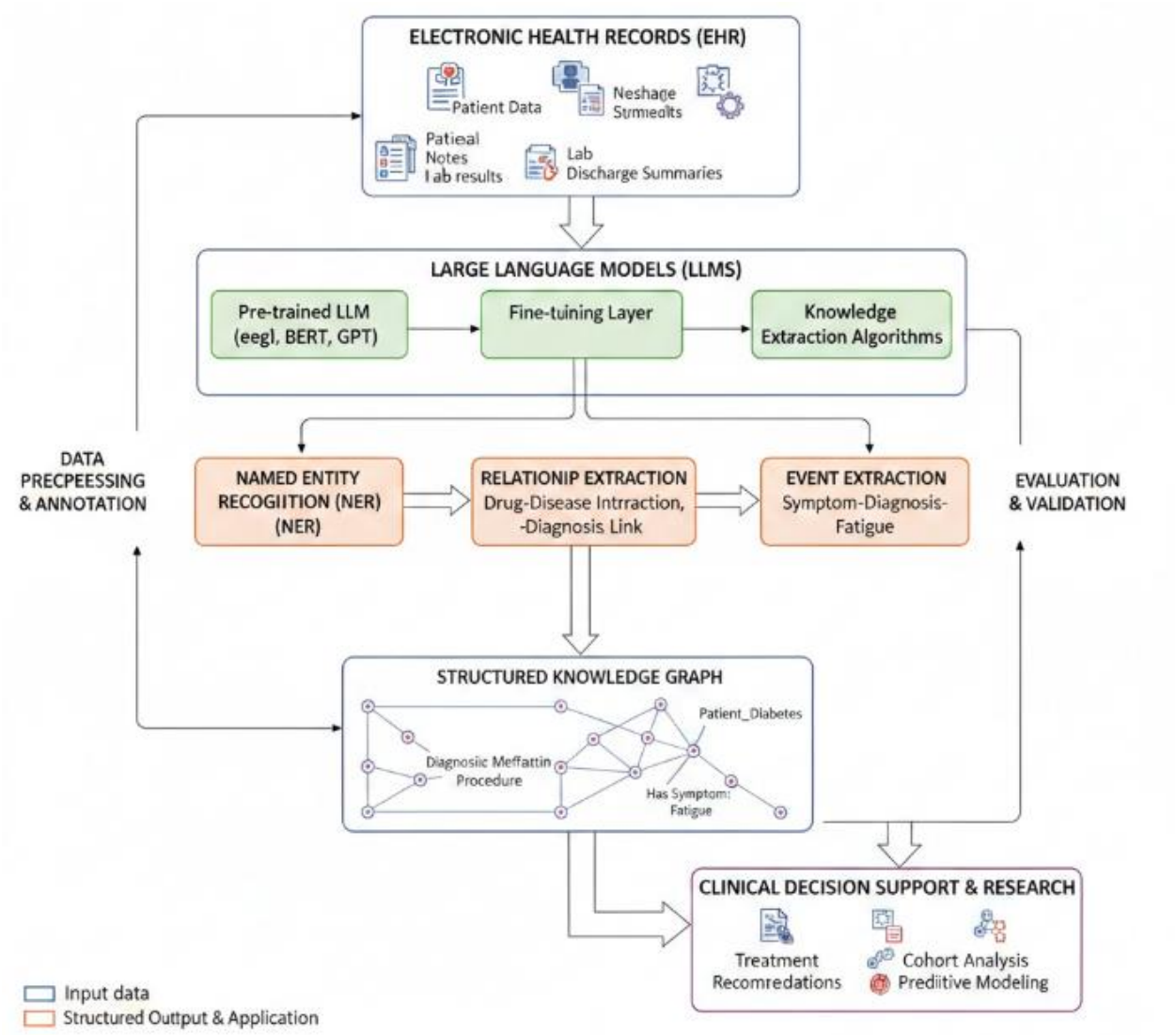


Fig 2: Methodological Framework for Knowledge Extraction

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

3.1. Data Preparation and Privacy Considerations

Data preparation must reconcile the inherent properties of clinical text and the underlying data protection frameworks. Clinical notes in EHRs contain sensitive information that requires de-identification to guard against breaching patient confidentiality. Procedures should adhere to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) guidelines for privacy and confidentiality of health information. Refer to established protocols such as SecuRec, which uses uniform guidelines to remove Protected Health Information, including records of dates (other than birth year), locations, and ages.

Once the data have been thoroughly de-identified, D. Lowe et al. reviewed for datum quality and tiles annotated clinical notes to produce a sample bank of external validity modifiable for use with surrogate models. Furthermore, expert consent and ongoing oversight of artificial intelligence models are crucial in any deployment of such software in production clinical systems, assuring appropriate governance, determination of direct and indirect uses, and conformance with institutions ethics guidelines. Finally, de-identification required to enable asset trading or application outside the institution domain should include data-preparation pipeline design considerations. Such procedures enable data provenance and enable compliance with requests to erase data collected for generating surrogate models, thus guaranteeing a well-audited corporate social responsibility.

Equation 2: LLM training objective used by "GPT-like" (autoregressive) extraction models

Large Language Models for Struc…

Step-by-step derivation

1. Given a token sequence $x_1, \ldots, x_T$
2. Chain rule factorization:

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1})$$

3. Maximum likelihood = maximize $\log P(\cdot)$, equivalently minimize negative log-likelihood:

$$\mathcal{L}_{\text{NLL}} = -\sum_{t=1}^{T} \log P(x_t \mid x_1, \ldots, x_{t-1})$$

3.2. Model Architectures and Training Paradigms

Several model types have been explored for extracting structurally relevant knowledge from unformatted EHR records, including autoregressive (GPT-like), encoder-decoder (BART-like), and retrieval-augmented (RAG) systems. Prompting pretrained language models is a popular approach requiring low additional investment, while end-to-end fine-tuning produces more specialized predictors at the expense of generalization capacity. Marginal gains from using larger models must be weighed against inference speed for clinical deployment.

Large LMs have become integral to summarization and answer-generation tasks where response quality is paramount. Regarding knowledge transformation, recent experiments suggest no architecture is inherently superior for all mapping challenges. Despite

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

their resource costs, autoregressive models have achieved promising performance for prevalent entity extraction problems in the clinical domain, relying on extensive, consistently annotated training corpora.

3.3. Mapping Unstructured Text to Structured Knowledge

Structured knowledge extraction from electronic health records (EHRs) focuses on the automatic mapping of unstructured text to structured knowledge represented in predefined schemas or domain ontologies. Structured knowledge can be defined as a set of assertions organised as a set of subject-relation-object triples, each a specific instance of a relation.

A structured representation of the knowledge extracted from EHRs can support advanced types of question answering and multimodal analysis. Depending on the system design, it can provide answers to temporal queries, permit event reasoning, or support answering queries of clinical interest that require synthesis of data from across different patient records. A knowledge extraction system can be tailored to meet the needs of a given application through the types of knowledge extracted and an appropriate choice of ontology.

Well-defined destination schemas or domain ontologies form the foundations of unstructured-to-structured knowledge mapping. Advances in large language models are enabling the semiautomated development of these schemas and ontologies—including consideration of entity type lists, relation inventories, design patterns for temporal-knowledge representation and event reasoning, and selection for linking relations to standard ontologies—alongside improvement of procedures for labelling textual sources. Normalisation and alignment processes permit integrated analysis of knowledge from different sources and ensure that synthesised answers are accurate, unambiguous, and in the correct language.

4. Extraction Techniques for Medical Entities and Relationships

Recent advances in representation learning for medical knowledge have paved the way for large-scale mapping of relevant medical concepts, medical entities based on records, relationships between conditions and procedures such as causation or chronology using large language models (LLMs), and dedicated exploration of reasoning processes with these systems. Specific application involves mapping EHR records to the concepts that make up existing ontologies, but it is also possible to test the wider reasoning capabilities of the LLMs on a range of entity and relation extraction tasks, including Named Entity Recognition (NER) or reasoning with temporal information. Such systems have been shown to learn with few examples and to be sensitive to the specific prompts provided.

The primary goal is the extraction of the entities indicated in Table 1 as well as relations between them. Such relations include causes, chronology, and association, with the aim of creating a collection of complete timelines that respect the ordering of temporally-sorted events and appropriately handle uncertainty.Ψ The outputs can be linked to standard clinical ontologies and metadata where available, and preprocessing steps may also utilize existing automatic solutions for some entity types (such as de-identification). The outputs are inherently multimodal, with information from both text and structured data sources being combined to produce timelines of events associated with a patient.

Equation 3: NER as sequence labeling + CoNLL evaluation metrics

Large Language Models for Struc…

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

1. Input tokens $x_1, \ldots, x_T$
2. Tags $y_1, \ldots, y_T$(e.g., BIO scheme)
3. A common modeling choice is per-token classification:

$$P(y_1, \ldots, y_T \mid x_1, \ldots, x_T) = \prod_{t=1}^{T} P(y_t \mid x_1, \ldots, x_T)$$

4. Training loss (cross-entropy):

$$\mathcal{L}_{\text{NER}} = -\sum_{t=1}^{T} \log P(y_t \mid x_1, \ldots, x_T)$$

Let:

- TP= correct predicted entity **spans**
- FP= predicted spans not in gold
- FN= gold spans missed

Then:

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, F1 = \frac{2PR}{P + R}$$

4.1. Named Entity Recognition in EHRs

Knowledge extraction from clinical notes involves identifying medical entities mentioned in the text; Named Entity Recognition (NER) is the NLP task that extracts these semantic concepts from unstructured text. Commonly recognized entities in EHRs include conditions, medications, procedures, tests, laboratory results, and other relevant mentions. To support the future application of NER in the extraction of structured knowledge from clinical text, multiple mapping schemes have been developed—mapping condition mentions to the Unified Medical Language System (UMLS); linking medications to ATC; associating structured lab results with Logical Observational Identifiers Names and Codes (LOINC); and connecting procedures and tests to the SNOMED CT ontology. The resulting annotations were then combined with the original text and evaluated using the standard CoNLL evaluation framework for NER, enabling a future comparison of results with widely available NER systems in the biomedical domain.

NE tagging in EHRs is inherently unbalanced because of the sparsity of mentions of some entity types. Some types are located in very few samples (e.g., tests), while others may overwhelm the dataset (e.g., conditions). Oversampling was thus employed as a data-leveraging strategy to increase the coverage of training data in domain-specific language models: the dataset was iteratively duplicated, with examples of the type to be oversampled injected into the new copies. The number of copies was defined by the

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

ratio of covered samples of the type underrepresented with respect to the total EHR corpus and its frequency of mention in the EHR training set.

4.2. Relation Extraction and Ontology Alignment

A variety of relations link medical entities in EHRs, including causation (e.g., disease resulting in a procedure), chronology (high-risk condition preceding death), and association (three blood pressure medications prescribed). Detecting such connections supports tasks such as generating cohorts for clinical trials, informing risk prediction models, and discovering associations between symptoms and diagnosis codes. Several methods can extract relations, including fine-tuned models, paraphrase-based zero-shot prompting, and manual rules that trigger when specific entity types appear together.

Entities relate to specific ontologies—causes to causalRelation, symptoms to symptomOf, and tests to hasFinding—that serve as an initial mapping. Standardized ontologies provide a breadth of commonly understood entity types and relations, but entity mention ambiguity poses a challenge. For instance, acute renal failure may correspond to both the condition and the diagnosis code, while loves gives no indication of subject or object. Additional context can clarify this uncertainty, mapping loves(mother, son) to the mother-and-son relationship while also mapping loves(son, ice-cream) to a different relationship.

Entities can undergo ontology alignment, linking with DeepOnto and thereby obtaining suggestions for a broader entity type vocabulary. Although DeepOnto lacks explicit relations tied to a timeline, existence and precedence-reduction methods can capture the underlying semantics. As a timeline grows, event sequencing emerges as an additional constraint for entity links.

4.3. Temporal and Event Reasoning

The temporal aspects of clinical narratives receive insufficient attention compared to recognition of medical entities and relations. Natural languages strongly express temporal information, and individuals rapidly build temporal models of events. Extracting time- and event-related information from medical text can enhance clinical informatics, and act as the backbone of clinical sensemaking. Reasoning over temporal entities (dates, times, durations) that are detected and extracted, along with logical temporal relations expressed by the narrative, allows reasoning about timelines and the lifecycles of patients.

Various techniques exist for extracting time expressions, timelines, and sequences of events with temporal ordering. A patient's clinical timeline chronologically summarizes clinical history through structured information and related accounts. Frequently, these timelines operate via lists of temporally-ordered events or events clustered within relevant temporal ranges. Formalizing the identified events to specify their order helps detect hidden errors or refocus attention on areas needing more investigation. In the absence of full clinical timelining through relations defined from the broad knowledge base, a more limited attempt at signal extraction provides a basis for reasoning over the event equilibrium of a selected patient cohort for an observed period. These methods connect to the issue of uncertainty associated with temporal information and with clinical narratives more generally, and consideration of both aspects is intertwined within these extensions.

5. Quality, Bias, and Generalizability Considerations

A number of aspects can impact the quality, representational capacity, and generalization ability of extraction techniques for structured knowledge discovery from EHRs. Data bias during model training can have a severe impact on the coverage of extraction techniques, as classification models tend to predict only the ranges of a labeled training set. Furthermore, biased

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

training data can force the models to discriminate against groups protected by law. Possible sources of such bias include disparities in terminology use among different patient demographics, hospital sites, and patient cohorts (e.g., condition prognosis), as well as in the distribution of conditions diagnosed by patients and healthcare workers. Adequate measures need to be taken to mitigate biases during training and repair them, if possible, for data processing approaches.

In the case of relation extraction for ontology alignment, data inequality can also arise when an important class of relation is underrepresented or simply not included in the training set. A similar effect is seen in named entity tagging for clinical conditions, procedures, tests, or medications when training data comes from a restricted patient cohort through a particular hospital site. With respect to fairness and equity, group-dependent error rates can be estimated and differencing and parity conditions can be verified. Several approaches to adversarial de-biasing can reduce group-dependent disparity in relation extraction without impairing performance on the primary task, hence ameliorating model fairness.

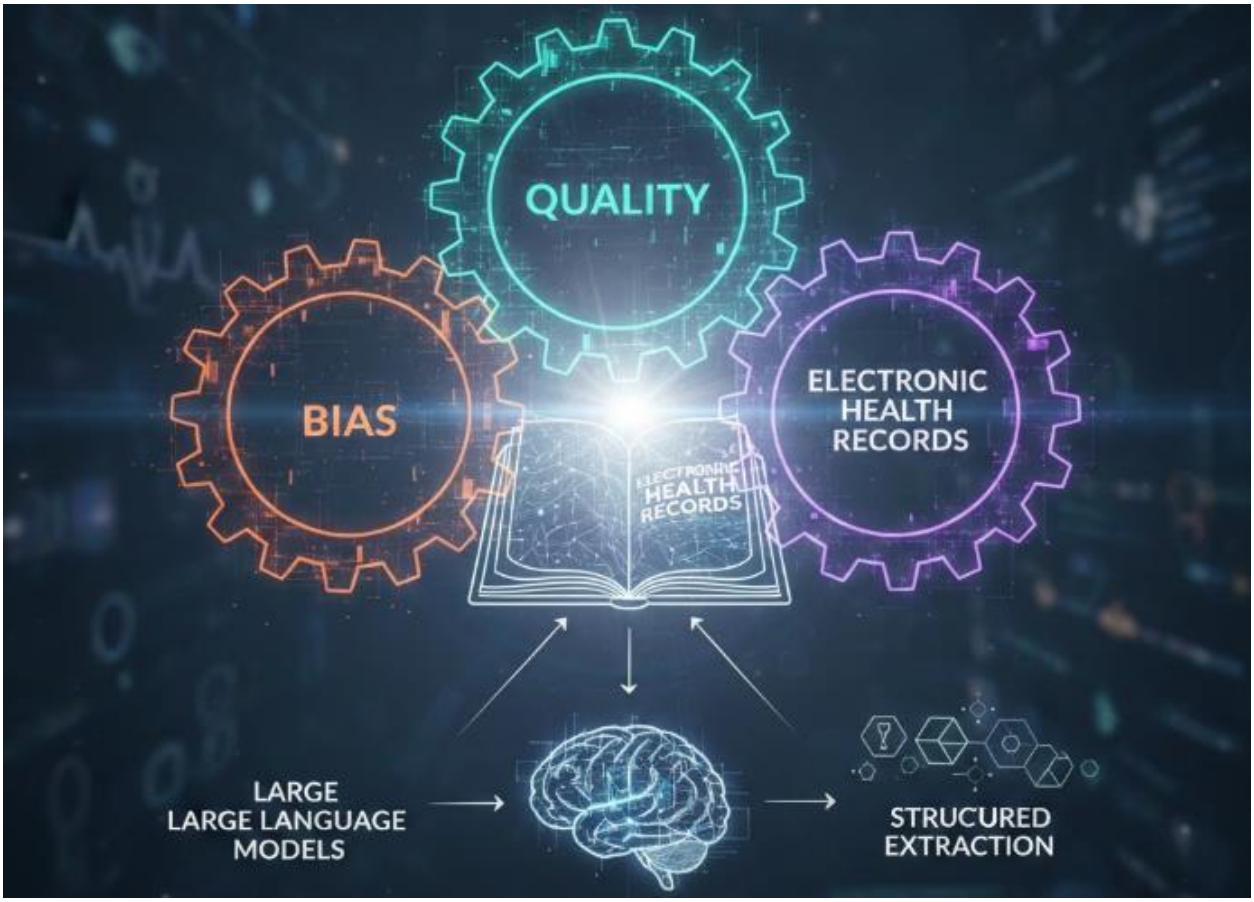


Fig 3: Quality, Bias, and Generalizability Considerations

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

5.1. Data Bias and Representation Gaps

Despite the recent strides that large language models have made in enabling structured knowledge extraction from clinical populations, important gaps in the accessibility and coverage of the underlying natural language data still persist. The vocabulary used in informal patient narratives, particularly in the setting of social media, is inherently different from that commonly encountered in traditional clinical records. Consequently, a large pre-trained autoregressive model is likely to be less effective at recognizing medical concepts like medications, procedures, or medical conditions in social media text, or at detecting relations such as event causation, association, and chronology, than a large model that has been fine-tuned on specific sets of social media data. Similarly, large pre-trained models into which bias has been introduced by the characteristics of the training set may suffer from significant representational gaps. Assessment of the coverage and potential bias of medical entity recognition systems thus appears warranted.

Fairness analysis is an established method for probing the sensitivity of models to disparate groups, quantifying how models are affected by factors such as ensembling, selection bias, and gender. In the context of social media–trained models, dimensionally reduced representations can also reveal biases toward majority population groups. Recent work documented that the performance of medical entity recognition and dialogue systems was consistently unequal across gender and ethnic distribution, with underperformance levels recorded for all under-represented groups in the training dataset. Bias mitigation thus remains an important consideration in building and applying health systems.

5.2. Privacy-preserving Techniques

Privacy concerns are central to the use of AI and other technologies in healthcare. A variety of methods exist to reduce the privacy risk associated with the availability of patient data. Popular approaches include anonymization and generalization of the data, the use of differential privacy techniques, and secure multi-party computation schemes. All these methods, however, come with their own trade-offs. For anonymization-based approaches, the data utility decreases unless significant background knowledge is assumed. Generalization and anonymization tend to reduce the amount of information learned by the model and often lead to a drop in performance. In addition, information that is not personally identifiable is often still sensitive; for example, sensitive conversation topics can often reveal private facts about the participants, even if these participants are not named in the conversation. Therefore, there is a compromise between patient safety and data utility.

Providing Data and Computations Under Differential Privacy (DP) guarantees that the likelihood of seeing a particular training record is approximately independent of whether or not the record is included in the training set, thus ensuring that changes to any individual record cannot have a significant effect on the output. For a single operation, DP ensures the privacy of the data used in that operation while still enabling data analysis. Moreover, a system can achieve group DP, in which groups of records lead to a release of a differentially-private model, providing some level of privacy for all individuals in the group. The downside of the method is the inherent noise injected into the model, which translates into a performance drop. Secure multiparty computation (SMPC) allows multiple parties to jointly compute a function over their inputs while keeping those inputs private. The computation is typically carried out over a secret-shared representation of the inputs, and the parties can later recover the final output. SMPC can be applied when multiple hospitals or even patients want to jointly train or test a model while keeping their data private.

Equation 4: Oversampling (class imbalance)

Operational step-by-step version

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

1. For entity type t, define **coverage**:

$$c_t = \frac{\#\text{samples containing type } t}{\#\text{total samples}}$$

2. Choose a target coverage $c^{\backslash *}$(or target minimum count)

3. Replication/oversampling factor:

$$k_t = \left\lceil \frac{c^{\backslash *}}{c_t} \right\rceil$$

5.3. Domain Adaptation and Transferability

Prompt-based adaptation in models requires examples that exhibit the task format as closely as possible to the inference and is largely constrained to the distribution of concepts and styles present in the training data. Techniques that increase the similarity join either pretraining examples through fine-tuning on in-domain data or by augmenting the input prompt with in-domain examples to reduce the path to infer similar queries. Methods that require a second-stage fine-tuning with in-domain data are usually more effective. Following this line of reasoning, calibration processes may expand the coverage of disagreements natural flows of data. Calibration takes external in-domain labeled data or soft predictions made by a separate classifier to adapt, either through score-weighted combinations or correcting sample-level disagreements. Calibrators reduce false positive rates and correct labels and are especially suited for operation in difficult in-domain tasks.

Prompt-based adaptation techniques also appear relevant when domain adaptation is difficult to achieve through re-training or fine-tuning. In cases in which Natural Language Processing tasks are being performed, a few-shot process may mask the limitations of the target prompt or style; few-shot adaptation have captured such property and been leveraged when no task-related examples were present within the model parameters. In few-shot implementations where the goal is to develop in-domain Natural Language Processing tasks and some limited labeled examples are available, an important question is how many task-related samples are necessary to correct or push the model attention toward the right direction. The answer relies on simple data efficiency analyses.

6. Deployment Scenarios and Clinical Integration

An architecture for the clinical deployment of large language models for structured knowledge extraction from unstructured medical narratives has been introduced. The knowledges extraction deidentification pipeline enables the anonymization of Electronic Health Records for the creation of datasets that preserve patients privacy. Structured knowledge is obtained in real time and fed into a knowledge base that can be explored by design and used during the execution of other language tasks. The considered architecture integrates a large language model with the hospital information system guaranteeing a transparent and efficient use for clinical personnel. Since sensitive information can be unintentionally leaked by the model during the execution of a task, a feedback loop that passes clinician validations back to the model is suggested to cope with trust issues. Careful attention is also paid to the guarantee of the patient privacy during both training and inference. The continuous integration of the

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

model and feedback loop in a DevOps pattern allow to cope with shift-data-problems, outbreaks and events that have typology and availability bias in the dataset.

Large language models displaced several other machine learning paradigms in many research fields, being able to produce state-of-the-art results in countless Natural Language Processing tasks and many other non-text domains. However, their huge computing resources consumption in training and inference, their time response and the necessity to cope with trust problems, related to the quality of the responses with respect to sensitive topics, make their usage costly for several tasks. Medically relevant big datasets coming from heterogeneous sources can enhance the model performance on the expected task. They can widely decrease the structured/non-structured information gap typical of clinical settings, moving from a zero-shot to few-shot or, when possible, fine-tuning training. Therefore, multimodal and multisource datasets can amplify model capabilities in clinical environments.

Equation 5: Relation extraction (RE) as classification

A standard operational formulation:

1. Encode context into a pair representation $h_{i,j}$
2. Predict relation label via softmax:

$$p(r \mid x, e_i, e_j) = \text{softmax}(W h_{i,j} + b)_r$$

3. Cross-entropy loss:

$$\mathcal{L}_{RE} = -\log p(r^{*} \mid x, e_i, e_j)$$

6.1. System Architecture for Clinical Use

A complete system for knowledge extraction and encoding into structured forms is composed of diverse components that address data flow (input/output specification), identity management (linking unstructured and structured information per unique patient) and potentially the integration within an existing EHR interface. At a high level, the architecture captures the general functionality of the entire pipeline, while still allowing sufficient flexibility regarding the implementation of individual components; these may or may not correspond to modules within the same software framework at the execution level.

The architecture classically relies on a back-end component for offline processing. In this configuration, the data to be processed—e.g., past admissions by the same patient and within the same institution—are made available within a common repository, and control is separated from the EHR dashboard interface, which a doctor uses during the daily clinical practice. Indeed, this decoupling allows the task to be performed on different machines dedicated to heavy workloads, thereby avoiding any latency for the user while maintaining a user-friendly experience. Still, a complex system of access control can ensure that sharing policies are respected, and results can be made available to patients only, or to patients and, for instance, a research institution, in such a way that the two parties can directly interact with each other through the auditable logs and response-filling interfaces.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

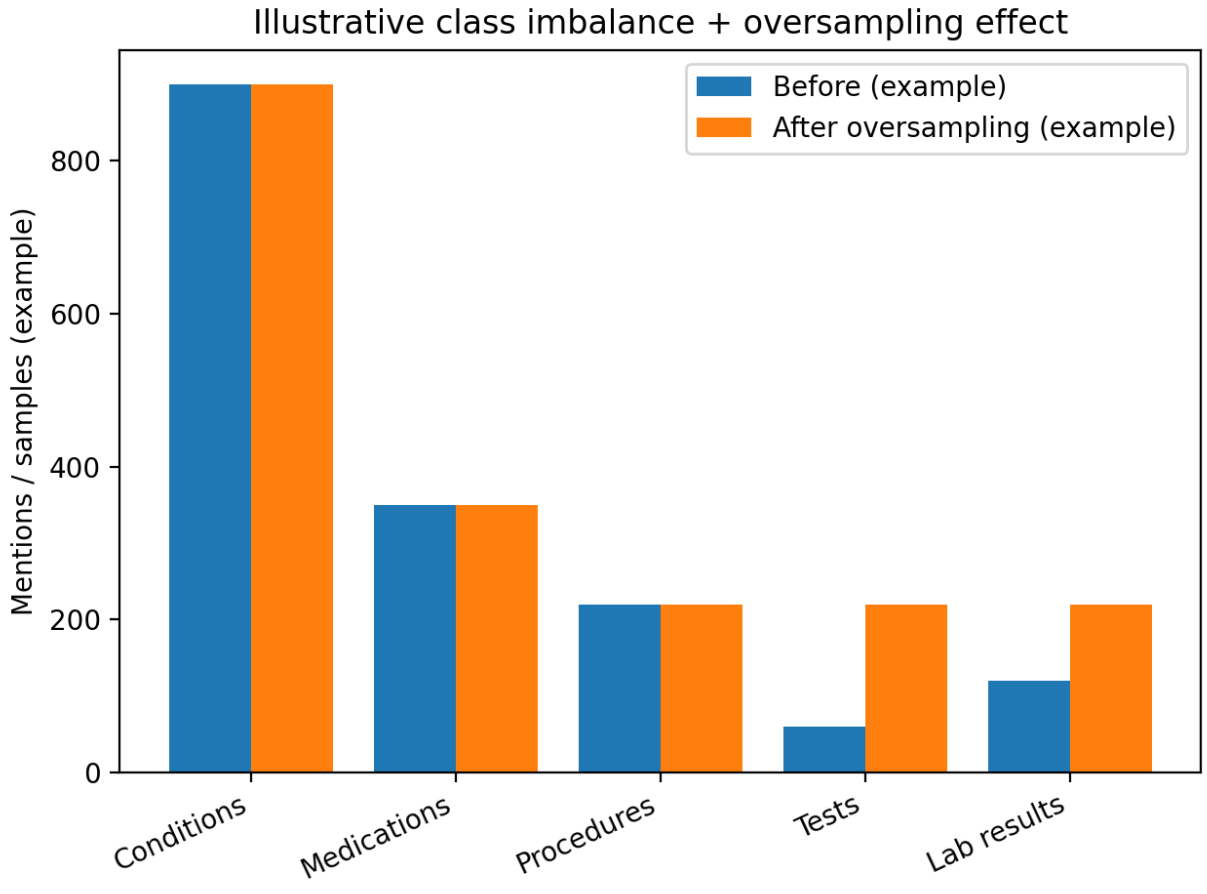
VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

6.2. Stewardship, Auditability, and Compliance

To establish credibility for diverse users, systems that autonomously generate clinical knowledge from EHRs must be properly entrusted, well governed, and adept at auditing their activities. Automatic documentation should provide provenance for generated knowledge, enabling decisions based on sources, methods, and evidence. Knowledge engineering employs a version-controlled repository to facilitate impact analysis, backward compatibility, and rollback. Furthermore, knowledge extraction aligns with regulatory frameworks from health authorities, offering a path towards meeting legal obligations around machine learning. ISO standards for quality, governance, and risk management of CDSSs could serve as a foundation.

Stewardship, version control, auditability, and regulatory alignment reduce the risk of producing unsafe, invalid, or unreliable conclusions and align with principles from the Institute of Medicine's reports on CDSS safety by design and trustworthy artifacts. Similar concerns arise in medical imaging. Addressing these elements fortifies confidence in the utility of such technology in real-world applications, minimizing associated risks.



ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

6.3. User Interaction and Trust Calibration

The methods described above can create AIs that take faces, voices, images, and text as input. Face verification ensures that the same face matches two different photos, the same real face matches a generated face, and selfies are captured in the actual world and not synthesized. Voice verification guarantees that the same person is speaking both real and synthesized sentences. A test in which images from two different sources indicate the same real object and an evaluation criterion of ImageNet validate the approach. Nevertheless, little work remains on trust for AI that integrates text, image, and audio synthesis from and crediting the correct source. When people and AIs create text, image, and audio data, the AI has to determine who created what data using the data nature (text, image, audio), data source (Facebook, YouTube, Google, and naturally human), and the data together. The goal of calibration is to provide the user with all the correct sources in which to trust and rely.

Trust calibration can be satisfied when the AI analyzes the data among the three modalities. Two main points can be considered. For AI that generates text and images, how can the AI determine the source of the generated image in the text? Abdou et al. recently published a paper on using knowledge graphs to confront misinformation on social media by integrating audio signals. Suppose the AI creates table knowledge graphs based on different sources with various heterogeneous relations in different modalities, connect the data across loss functions to determine whether the relationships exist or do not exist, and describe every knowledge graph in every modality. It then remains to define the proper confidence level for each relationship across data and verbalize them by the cross-modal pipeline in the same way that humans speak. For an AI that generates text and video, how can it analyze whether the generated video is true or fake? The correspondence of audio and video data is used to determine that.

7. Case Studies and Empirical Findings

The richness and diversity of electronic health records offer exciting prospects for using large language models capable of integrating various data types. A large study at Stanford University connected unstructured clinical notes and radiology reports with metadata from the Cancer Registry, structured records, and imaging data to extract a range of cancer-relevant entities in a multimodal setting. Following a similar prompt-based extraction approach, the system enabled cross-source validation of the predictions by contrasting unstructured and structured sources.

Exploring the temporal reasoning capabilities of large language models over the framework of Event and Tense Logic, Stanford created a synthetic benchmark with explicit ground truth. The results revealed major limitations, suggesting that specialized models trained on collections of temporal textual patterns using a variety of techniques (event extraction, named entity recognition, relation extraction in event sequences) could produce more reliable temporal sequences. An assessment of the potential for transfer learning highlighted substantial domain adaptation capabilities, showing good performance even when training data was not representative of the new domain.

7.1. Multimodal and Multisource Integration

Structured knowledge extraction from unstructured and semi-structured electronic health record (EHR) data is emerging as an increasingly important research focus. When applied to EHR text records and clinical notes, it has the potential to provide valuable problemsolving knowledge for risk prediction and treatment response studies. Despite recent advances, the conception, design, and study of knowledge extraction systems for clinical text through natural language processing and large language

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

models remain in their infancy, with only a handful of works published to date. Results of such specialized systems must be shown to generalize before they can be safely deployed for clinical use. To this end, novel implementations for named entity and relation extraction, latent temporal reasoning, and alignment with standard biomedical ontologies are described. The methods cover core tasks identified in the general knowledge-extraction literature and are adapted to the context of de-identified clinical data.

Two directions for further evaluation are examined. A second approach applies coherent extraction methods to integrate information from a diverse array of data types, including images. Using unified neural modalities for the joint processing of tabular, textual, and image data, a multimodal and multisource knowledge-extraction pipeline is designed, and its use-cases and findings are described. The second testing ground returns to clinical text, focusing on data from a considerably different source, the publicly available MIMIC-III Intensive Care Unit dataset. Through these efforts, the generalization of the previously developed knowledge-extraction methods, together with their robustness, is studied.

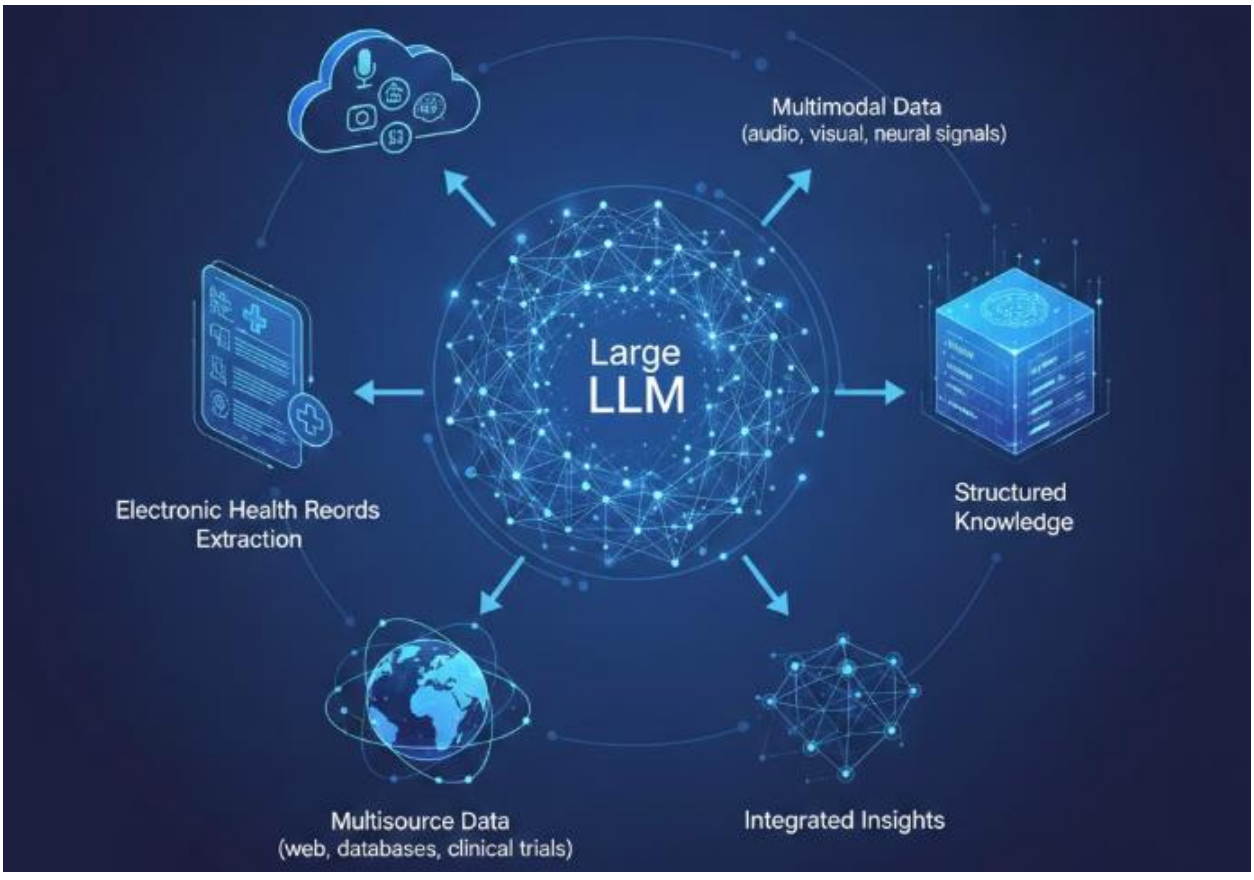


Fig 4: Multimodal and Multisource Integration of Large Language Models

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

8. Conclusion

In summary, large language models enable knowledge extraction from unstructured sources such as electronic health records and clinical narratives at different levels of granularity. The resulting structured knowledge can be a first step toward building comprehensive target-directed knowledge bases that also conform to standard biomedical ontologies and contain information about the identity and temporal relations of medical entities. The design of these methods allows for their seamless integration into the clinical workflow of healthcare institutions and thereby enables the systematic and fully automated creation of structured knowledge modules that are up to date. Such modules can potentially serve as external knowledge resources for improved performance and calibration of large language models or can be incorporated into downstream AI pipelines to aid a range of clinical tasks and decision-making scenarios. As a next step, the methods can be further extended to support transcriptomic and proteomic data, characterizing diseases, and predicting therapies.

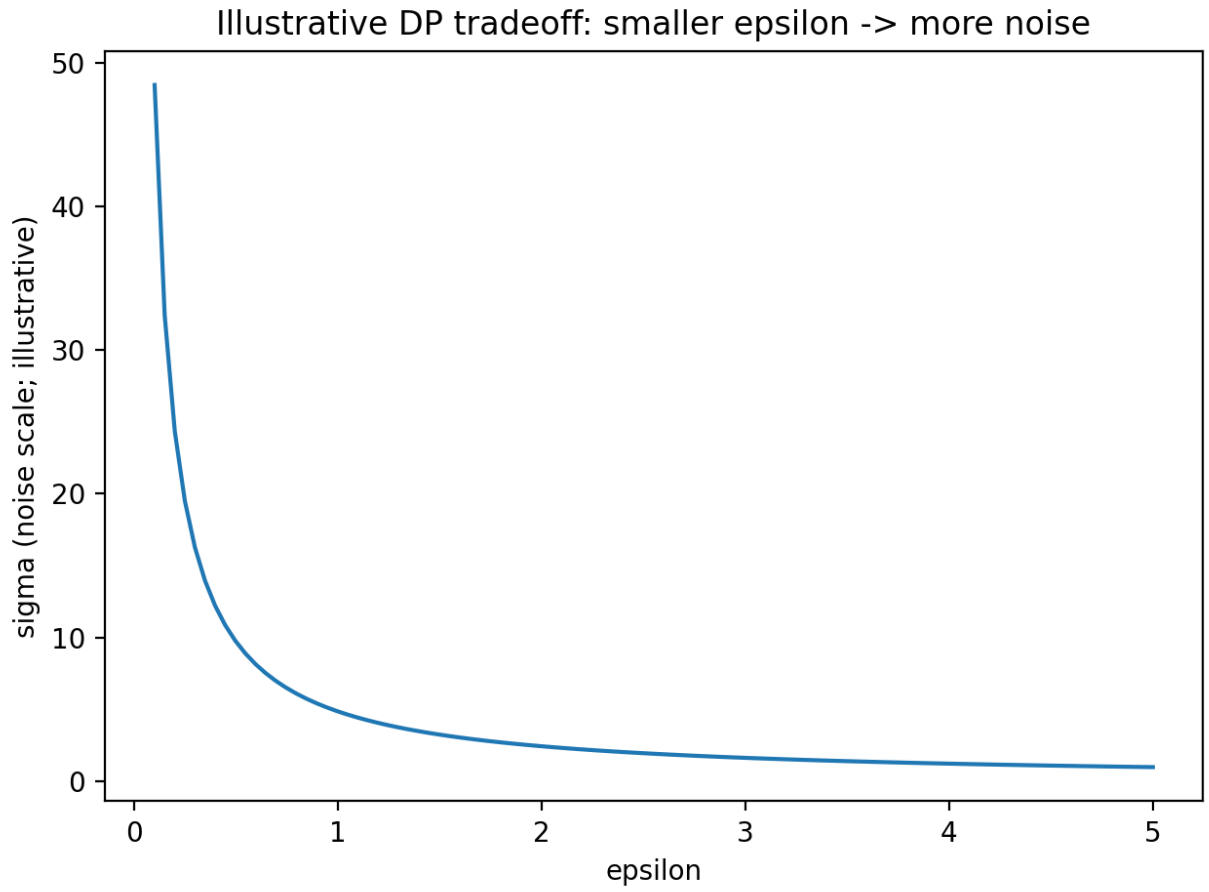
Further advances in large language models that enhance capabilities in reasoning, memory, and factuality are anticipated. Future research may thus integrate progress in knowledge retrieval or grounding with the mapping of natural language text into structured representations for clinical domains such as cardiology or oncology. In particular, recent advancements in multimodal or multitask training of such models on both structured and unstructured data, transfer learning across disparate clinical domains, and support of additional types of temporal reasoning beyond simple causation are promising avenues for subsequent exploration. Further advances in large language models (LLMs) are expected to substantially strengthen their capacity for sophisticated reasoning, long-term memory, and factual consistency, thereby expanding their utility in complex clinical decision-making contexts. Future research will likely focus on tightly coupling knowledge retrieval and grounding mechanisms with methods that map free-text clinical narratives into structured, machine-interpretable representations, particularly within high-impact domains such as cardiology and oncology. Moreover, continued progress in multimodal and multitask training paradigms—leveraging both structured databases and unstructured clinical text—may enable more robust transfer learning across heterogeneous medical specialties. Finally, extending model support for richer forms of temporal reasoning, including longitudinal disease progression, treatment sequencing, and delayed or interacting effects, represents a critical direction for building clinically reliable and interpretable AI systems.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17



8.1. Future Trends

As multimodal and multisource knowledge extraction research matures, progress towards clinical integration appears both feasible and highly significant. Results across a variety of integrated data types suggest that evidence captured from electronic health records and imaging reports will soon reach levels of comprehensiveness greater than those possible when using text alone. With the rapid speed of reasoning and the scalable, inexpensive nature of the underlying LLMs, such multimodal systems could become part of routine clinical evaluation and decision-making. Like other methods, however, they remain vulnerable to bias in training data, and monitoring for testing, validation, and development data imbalance remains crucial.

Looking beyond EHRs, the burgeoning field of model-driven multimodal analysis will enable integration of text, structured data, imaging, and signal processing models for vision, language, speech, audio, and time-series data. These developments may promote cross-source validation of knowledge affairs, allowing model predictions to be independently credibly supported by other modalities, in addition to facilitating multimodal clinicians, integrating information from multiple disparate channels. Such

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

systems are currently aspirational, and further research into intermodal integration strategies present significant opportunities for advancement in the weeks, months, and years ahead.

9. References

1. Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). BEHRT: Transformer for electronic health records. Scientific Reports, 10, 7155.
2. Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K. J., Shen, F., Wang, L., Wang, Y., Wen, A., Zhao, Y., Sohn, S., & Liu, H. (2020). Clinical concept extraction: A methodology review. Journal of Biomedical Informatics, 109, 103526.
3. Mashetty, S., Malempati, M., Paleti, S., Adusupalli, B., & Singireddy, J. (2025). A Multidisciplinary Framework for AI and Data-Driven Transformation in Taxation, Insurance, Mortgage Financing, and Financial Advisory: Integrating Cloud Computing, Deep Learning, and Agentic AI for Community-Centric Economic Development. Insurance, Mortgage Financing, and Financial Advisory: Integrating Cloud Computing, Deep Learning, and Agentic AI for Community-Centric Economic Development.
4. Kraljevic, Z., Searle, T., Shek, A., Roguski, Ł., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A., Roberts, A., Bendayan, R., & Teo, J. (2021). Multi-domain clinical natural language inference. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 3229–3241.
5. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D., & others. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine, 4, 86.
6. Burugulla, J. K. R., Kannan, S., Malempati, M., Challa, H. A. H. S. R., Pandugula, C., & Goma, T. (2025, April). Federated Learning Based Cloud Computing Solutions with Privacy Preservation for Intelligent Utilities. In 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0 (pp. 1-6). IEEE.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

7. Mulyar, A., Dernoncourt, F., & Dockhorn, C. (2021). MT-clinical BERT: Scaling clinical information extraction with multitask learning. Journal of the American Medical Informatics Association, 28(10), 2108–2115.
8. Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2021). Self-alignment pretraining for biomedical entity representations. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4228–4238.
9. Kummari, D. N., Singireddy, J., Sheelam, G. K., Nandan, B. P., Pandiri, L., Lakkarasu, P., & Dwaraka. (2025, August). Generative AI Models for Process Optimization in Semiconductor Wafer Design and Yield Prediction. In International Conference on Artificial Intelligence: Theory and Applications (pp. 220-233). Cham: Springer Nature Switzerland.
10. Gu, Y., Tinn, D., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23.
11. Pamisetty, V. (2019). Machine Learning Models for Real-Time Tax Fraud Detection and Risk Assessment in Digital Government Systems. Global Research Development (GRD) ISSN, 2455-5703.
12. Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020). Pretrained language models for biomedical natural language processing: A review. Briefings in Bioinformatics, 22(1), 1–14.
13. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.
14. Adusupalli, B., Malempati, M., Paleti, S., Mashetty, S., & Singireddy, J. (2025). Integrated financial ecosystems: AI-driven innovations in taxation, insurance, mortgage analytics, and community investment through cloud, big data, and advanced data engineering. Journal of Information Systems Engineering and Management, 10, 1103-1117.
15. Peng, Y., Yan, S., & Lu, Z. (2020). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. Proceedings of the 18th BioNLP Workshop and Shared Task, 58–65.
16. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. B. A. (2020). Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop, 72–78.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

17. Suura, S. R., Chava, K., Chakilam, C., Nuka, S. T., Maguluri, K. K., & Goma, T. (2025, April). Blockchain-Based Secure and Scalable Models for Healthcare Network Traffic Monitoring and Optimization. In 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0 (pp. 1-6). IEEE.
18. Si, Y., Wang, J., Xu, H., & Roberts, K. (2021). Enhancing clinical concept extraction with contextualized word representations. Journal of Biomedical Informatics, 117, 103754.
19. Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records. npj Digital Medicine, 5, 140.
20. Chakilam, C., Kannan, S., Recharla, M., Suura, S. R., & Nuka, S. T. (2025). The impact of big data and cloud computing on genetic testing and reproductive health management. American Journal of Psychiatric Rehabilitation, 28(1), 62-72.
21. Lu, Q., Dou, D., & Nguyen, T. (2022). ClinicalT5: A generative language model for clinical text. Findings of the Association for Computational Linguistics: EMNLP 2022, 5436–5443.
22. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), bbac409.
23. Sivanand, R., Kumar, D. P., Nagabhyru, K. C., Natarajan, E. P., Pamisetty, V., & Kapila, D. (2025, September). IoT and AI for Real-Time Monitoring in Substation Automation. In 2025 International Conference on Computing and Communications (COMPUTINGCON) (pp. 1-5). IEEE.
24. Yasunaga, M., Leskovec, J., & Liang, P. (2022). LinkBERT: Pretraining language models with document links. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 8003–8016.
25. Gu, Y., Tinn, D., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23.
26. Yang, X., Wong, C., & others. (2022). A large language model for electronic health records. npj Digital Medicine, 5, 194.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

27. Kalisetty, S., Lakkarasu, P., Singireddy, S., Burugulla, J. K. R., Challa, K., & Gadi, A. L. (2025, August). Next-Gen Payment Gateways: Leveraging Federated Learning for Fraud Detection in Cross-Border Transactions. In International Conference on Artificial Intelligence: Theory and Applications (pp. 200-211). Cham: Springer Nature Switzerland.
28. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29, 1930–1940.
29. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620, 172–180.
30. Paleti, S., Baliyan, M., Aitha, A. R., Reddy, B. A., Bhadauria, G. S., & Sing, S. A. (2025, August). Graph—LSTM Hybrid Model for Improving Fraud Detection Accuracy in E-Commerce Financial Services. In 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) (pp. 1-6). IEEE.
31. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, J., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachmann, T., McKinney, S. M., Azizi, S., & Natarajan, V. (2023). Towards expert-level medical question answering with large language models. Nature Medicine, 29, 1–8.
32. Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.
33. Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2023). A large language model for electronic health records: GatorTron and its clinical applications. Journal of the American Medical Informatics Association, 30(4), 1–12.
34. Bolton, E., Li, J., Hall, K., Yoder, N., Riley, A., & others. (2023). Stanford AlpacaMed: Instruction tuning language models for clinical applications. Proceedings of the Clinical NLP Workshop, 1–10.
35. Jeblick, K., Schachtner, B., Dexheimer, J., et al. (2024). ChatGPT makes medicine easy to swallow: An exploratory analysis of AI-generated medical explanations. European Radiology, 34, 1–10.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

36. Huang, J., Yang, D. M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., Klesse, L. J., Xiao, G., Peterson, E. D., Zhan, X., & Xie, Y. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. npj Digital Medicine, 7, 106.
37. Mashetty, S. (2025). Technology-driven analytics in mortgage-backed securities for single-family mortgage financing. Available at SSRN 5236573.
38. Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K., & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. Artificial Intelligence in Medicine, 154, 102900.
39. Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., Zhang, M., Cao, C., Wang, J., Wang, X., Gao, J., Wang, Y.-G.-S., Ji, J., Qiu, Z., Li, M., Qian, C., Guo, T., Ma, S., Wang, Z., Guo, Z., Lei, Y., Shao, C., Wang, W., Fan, H., & Tang, Y.-D. (2024). The application of large language models in medicine: A scoping review. iScience, 27(5), 109713.
40. Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: Applications, advances, and challenges. Artificial Intelligence Review, 57, 299.
41. Omiye, J. A., Gui, H., Jasim, M., Chen, J. H., & Daneshjou, R. (2024). Large language models in medicine: The potentials and pitfalls. Annals of Internal Medicine, 177(2), 210–220.
42. Omar, M., Nadkarni, G. N., Klang, E., & Glicksberg, B. S. (2024). Large language models in medicine: A review of current clinical trials across healthcare applications. PLOS Digital Health, 3(11), e0000662.
43. Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2024). Capabilities of GPT-4 on medical challenge problems. npj Digital Medicine, 7, 1–10.
44. Shah-Mohammadi, F., & Finkelstein, J. (2024). Extraction of substance use information from clinical notes: A generative pretrained transformer-based investigation. JMIR Medical Informatics, 12, e56243.
45. Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2024). Clinical natural language processing for extracting structured information from electronic health records. Journal of Biomedical Informatics, 149, 104575.
46. Wang, Y., Sun, S., & others. (2024). Large language models for clinical information extraction and representation learning. Journal of Biomedical Informatics, 152, 104635.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

47. Liu, X., Xu, Y., & others. (2024). Large language models for electronic health record summarization and clinical information extraction. Journal of the American Medical Informatics Association, 31(6), 1–12.
48. Wu, S., Zhang, X., & others. (2024). Clinical information extraction using transformer-based language models: A systematic evaluation. BMC Medical Informatics and Decision Making, 24, 1–15.
49. Kim, M.-S., Chung, P., & Aghaeepour, N. (2025). Information extraction from clinical texts with generative pre-trained transformer models. International Journal of Medical Sciences, 22(5), 1015–1028.
50. Ntinopoulos, V., & others. (2025). Large language models for data extraction from unstructured and semi-structured electronic health records: A multiple model performance evaluation. BMJ Health Care Informatics, 32, e101139.
51. Kim, J., Lee, S., & others. (2025). Scalable information extraction from free text electronic health records using large language models. BMC Medical Research Methodology, 25, 1–15.
52. Mahony Reategui-Rivera, C., & Finkelstein, J. (2025). Evaluation of the performance of a large language model to extract signs and symptoms from clinical notes. Studies in Health Technology and Informatics, 329, 1–5.
53. Fierens, A., Englebert, A., & Jodogne, S. (2025). Translating UMLS concepts to improve medical entity linking in French: A SapBERT-based approach. Studies in Health Technology and Informatics, 329, 1–5.
54. Shah, N. H., Chen, J. H., & others. (2025). Clinical entity augmented retrieval for clinical information extraction. npj Digital Medicine, 8, 45.
55. Lu, Y., Zhang, Y., & others. (2025). Large language models for clinical text processing: A systematic evaluation of information extraction and summarization. Journal of Biomedical Informatics, 162, 104760.
56. Chen, A., Yang, X., & others. (2025). Integrating large language models with human expertise for disease detection in electronic health records. Computers in Biology and Medicine, 186, 109773.
57. Zhang, Y., Liu, X., & others. (2025). Large language models for clinical text understanding and electronic health record analysis: A systematic review. Artificial Intelligence in Medicine, 160, 103024.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 04 ISSUE: 02

RECEIVED: APRIL 22 REVISED: MAY 19 ACCEPTED: MAY 27 PUBLISHED: JUNE 17

58. Li, J., Chen, Y., & others. (2025). Generative artificial intelligence for clinical natural language processing: Applications, evaluation, and challenges. Journal of the American Medical Informatics Association, 32(4), 1–12.
59. Truhn, D., Cuocolo, R., Adams, L. C., & Bressem, K. K. (2025). Current applications and challenges in large language models for patient care: A systematic review. Communications Medicine, 5, 26.