Extraction of Patient Subtypes using LLM Generated Knowledge Graphs Integrated With a Transformer Architecture

Benjamin Holmes[1], Cogan Shimizu[1]

[1]Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435

Abstract

Extracting patient subpopulations (clinically relevant cohorts of individuals who share overlapping symptoms, risk factors, or diagnostic criteria) from unstructured medical notes is an ongoing challenge due to the variability of clinical language and the complex nature of patient conditions. We demonstrate a pipeline that combines named entity recognition (NER), transformer embeddings, guided dimensionality reduction, and LLM-mediated knowledge graph integration to enhance patient extraction. The approach begins with NER using the UMLS metathesaurus [1] to extract clinical terms, followed by transformation into vector embeddings using a biomedical transformer. These embeddings are augmented with structured knowledge graph representations generated through an LLM-driven extraction process and graph embeddings via TransE [2]. To improve the separation of key semantic features, we apply autoencoder-based dimensionality reduction before concatenating term embeddings with their graph-based counterparts. A feedforward neural network with an attention layer classifies extracted embeddings to determine patient subgroup membership. We evaluate the pipeline on multiple datasets, including extracting a subpopulation taken from Dayton Childrens' Hospital, with experiments demonstrating improvements over baseline BERT-only and keyword-based methods in classifying medical reports by specialty and behavioral health relevance. Our results show that incorporating knowledge graphs and dimensionality reduction enhances precision and interpretability while maintaining adaptability for different research queries.

Keywords

Medical NER, Knowledge Graphs, Transformer, LLM

1. Introduction

Electronic medical records (EMRs) contain enormous amounts of unstructured clinical text, making it difficult to extract meaningful patient subsets for research, clinical trial matching, and other forms of medical decision making [3]. Human annotation of these files takes up an increasingly large amount of clinical time for nurses and doctors in healthcare settings, and is prone to error [4]. Standardized medical coding systems, such as ICD-10, fail to capture the full complexity of patient conditions, leading to gaps in automated classification [5]. While natural language processing (NLP) techniques have made progress in structuring free text data, challenges remain in ensuring both accuracy and interpretability, particularly when dealing with ambiguous or underrepresented conditions [6].

Paradigms and Approaches in Medical NLP: Recent advances in domain-specific NLP models and knowledge representation techniques offer promising solutions for patient subset extraction. Large Language Models (LLMs) offer the ability to manipulate and normalize complex texts, though they come with the costs of large processing requirements, patient confidentiality concerns, and hallucinations. [7]

Transformer-based embeddings, generated by models pretrained on biomedical information, such as BioBERTa [8], have demonstrated strong performance in generating embeddings that effectively capture relationships between medical terms [9]. Knowledge graphs (KGs)

LLM-TEXT2KG 2025: 4th International Workshop on LLM-Integrated Knowledge Graph Generation from Text (Text2KG), June 1 - Jun 5, 2025, co-located with Extended Semantic Web Conference (ESWC), Portoroz, Slovenia

holmes.51@wright.edu (B. Holmes); cogan.shimizu@wright.edu (C. Shimizu)

0000-0003-2166-2532 (B. Holmes); 0000-0003-4283-8701 (C. Shimizu)

further enhance this process by integrating structured relationships between extracted terms, helping to disambiguate concepts and provide context [10]. Peng et al.[11] provide a comprehensive overview of knowledge graphs in AI, highlighting both their promise for structured reasoning and the ongoing challenges in scalability, data integration, and semantic consistency—challenges that motivate the hybrid LLM-based approach proposed here. However, effectively combining these techniques in a scalable way while maintaining patient confidentiality and model governance remains an open research question.

A Novel Pipeline: This work introduces a pipeline for automated patient subset extraction that integrates named-entity recognition (NER), transformer-based embeddings, and graph-based knowledge representations. Our approach builds on prior work such as Dessì et al.[12], who demonstrated the feasibility of using NLP and machine learning techniques to automatically construct knowledge graphs from unstructured technical documents. Recent work such as GraphRAG [13] explores how structured graph representations can enhance LLM-driven summarization and retrieval, aligning with our goal of using knowledge graphs to contextualize medical terms extracted from free-text. We employ NER tools that leverage the UMLS metathesaurus to extract clinical entities, which are then embedded using BioBERTa and enriched with graph embeddings generated through TransE. To improve classification performance, we apply autoencoder-based dimensionality reduction before concatenating term embeddings with their graph-based counterparts. A feedforward neural network (FFNN) with an attention mechanism classifies these enriched embeddings, improving accuracy and interpretability.

Overall, we propose a five-stage pipeline that integrates established and emerging tools in clinical NLP:

1. Named Entity Recognition (NER) is used to extract medically relevant terms from free-text notes using tools grounded in the UMLS Metathesaurus.
2. Transformer-based embeddings (via BioBERTa) are generated for each extracted entity to capture semantic relationships.
3. Knowledge graph construction is performed using a large language model to identify medically meaningful relations (e.g., associated_with, is_a), enabling graph embeddings via TransE.
4. Dimensionality reduction is optionally applied using a guided autoencoder that preserves task-relevant structure in embedding space.
5. Classification is performed by a feedforward neural network (FFNN) with an attention layer to determine subgroup membership.

We evaluate our approach using multiple datasets with real-world patient data, including MIMIC-IV [14], and a set of medical records provided for a behavioral health classification task. Our results demonstrate that incorporating knowledge graphs and dimensionality reduction improves precision over baseline BERT-only and keyword-based methods. This work contributes A flexible, scalable pipeline for extracting complex patient subpopulations from medical free text that is:

- Adaptable to diverse research questions
- Able to be customized, without requiring deep machine learning knowledge
- An on-premises solution, not requiring exporting sensitive data to a public model

The rest of the paper is structured as follows: Section 2 discusses key related work, Section 3 covers our approach in detail, Section 4 presents several experiments designed to validate the framework, and Section 5 discusses results and possible extensions to this work.

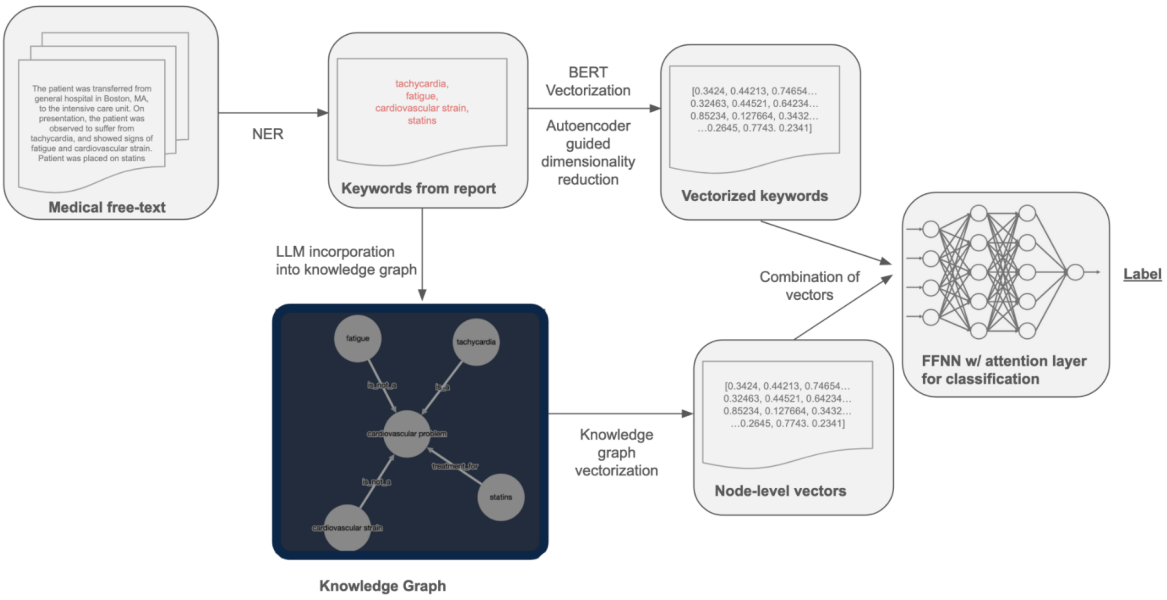

Figure 1: The pipeline for extraction of patient subsets from medical free-text: from the text "The patient was transferred from general hospital in Boston, MA, to the intensive care unit. On presentation, the patient was observed to suffer from tachycardia, and showed signs of fatigue and cardiovascular strain. Patient was placed on statins.", NER extracted "tachycardia", "fatigue", "cardiovascular strain", and "statins".

2. Background

Extracting structured information from unstructured medical texts is an ongoing challenge in clinical natural language processing (NLP). Traditional approaches, such as rule-based systems and dictionary-based methods, struggle with variability in medical language and context-dependent terminology. Recent advances in deep learning and knowledge representation techniques offer more robust solutions, leveraging transformer-based embeddings, knowledge graphs (KGs), and dimensionality reduction to improve classification performance.

Medical Named Entity Recognition (NER)

Early approaches to medical NER relied heavily on rule-based systems and dictionary-based methods, which aimed to map free-text medical descriptions to standardized terminologies. These methods used hand-crafted rules, regular expressions, and lexicons such as the Unified Medical Language System (UMLS) to identify clinical entities like diseases, medications, and procedures [15]. While impressive in scope, complete reliance on predefined vocabularies made these tools sensitive to variations in clinical language. Additionally, rule-based methods struggled with ambiguity, context-dependent meanings, and scalability, limiting their effectiveness in large-scale clinical datasets[16].

NLP Framework Approaches

Traditional machine learning approaches, such as support vector machines (SVMs) and conditional random fields (CRFs), enabled more flexible recognition of medical entities by learning patterns in annotated corpora. However, these models still required extensive feature engineering and were limited in their ability to generalize beyond predefined feature sets [17]. The advent of deep learning, particularly recurrent neural networks (RNNs) and later transformer-based models, significantly enhanced entity extraction by capturing more complex contextual relationships. Models like BioBERT and ClinicalBERT [18], pre-trained on biomedical literature, demonstrated superior performance in medical text processing by generating dense,

contextualized embeddings for clinical terms. [18] Despite these advances, transformer-based models remained limited in their ability to incorporate structured medical knowledge, often relying solely on textual context without explicit relationships between medical entities [19].

Large Language Models
Recent advances in large language models (LLMs) have further transformed medical NLP by enabling automated entity extraction, relationship discovery, and knowledge graph construction from free-text clinical notes [20]. Unlike earlier transformer models that required task-specific fine-tuning, LLMs such as GPT-4 can perform zero-shot and few-shot learning, allowing them to extract structured medical concepts with minimal supervision [21]. These models leverage massive pretraining corpora, capturing a broad range of clinical knowledge and linguistic patterns. However, LLMs face challenges such as hallucinations, lack of domain-specific grounding, and ethical concerns related to patient data privacy [22, 23, 24]. To address these limitations, hybrid approaches that integrate LLMs with structured representations, such as knowledge graphs and graph embeddings, are emerging as a promising direction for medical NLP. These methods combine the contextual understanding of deep learning models with the interpretability and relational structure of expert-curated medical ontologies, improving both accuracy and reliability in clinical text classification [25, 26].

Limitations in Transformer-Based Embeddings Transformer-based models, such as BioBERT and ClinicalBERT, have demonstrated state-of-the-art performance in biomedical text embedding [27, 28]. Pre-trained on large corpora of scientific literature and clinical notes, these models generate dense embeddings that capture semantic relationships between medical terms. However, standalone transformer embeddings may not incorporate structured knowledge, leading to difficulties in disambiguating similar terms or understanding hierarchical relationships between clinical entities [19].

Knowledge Graphs for Clinical NLP Knowledge graphs provide a structured representation of medical knowledge by encoding relationships between entities. In clinical applications, KGs have been used to enhance decision support, improve explainability, and address sparsity in textual data [29, 30, 31]. Graph embedding techniques, such as TransE, allow models to incorporate relational information into vector space representations, improving classification and retrieval tasks [2]. Despite their advantages, automated KG construction remains an open challenge, as large language models (LLMs) can introduce noise, redundancies, or inconsistencies when extracting structured relationships from text [32].

Dimensionality Reduction High-dimensional embeddings from transformer models can capture rich contextual information but may introduce redundancy and computational inefficiency [33, 34]. Autoencoders provide a means of reducing dimensionality while preserving key semantic distinctions between medical concepts [35, 36]. Guided dimensionality reduction techniques have shown promise in improving cluster separation for symptom classification and disease subgroup identification [37]. However, effective dimensionality reduction requires careful tuning to avoid loss of clinically relevant information.

This work builds on these advances by integrating medical NER, transformer embeddings, knowledge graphs, and dimensionality reduction into a unified pipeline for automated patient subset extraction. By combining these techniques, we aim to improve the accuracy, interpretability, and adaptability of clinical NLP systems for diverse research and decision-support applications.

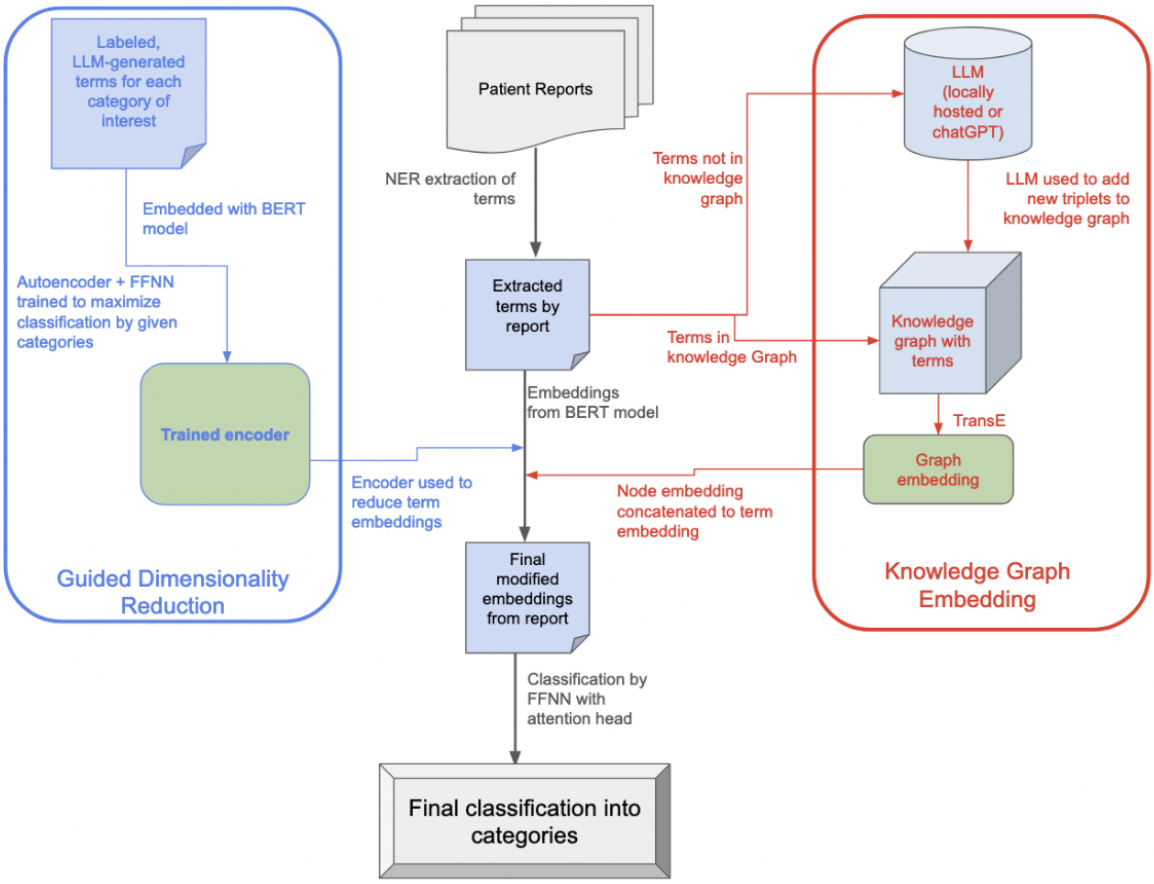

Figure 2: The full proposed pipeline

3. Approach

To extract clinically meaningful patient subpopulations from unstructured medical text, we propose a pipeline that integrates named entity recognition (NER), transformer-based embeddings, knowledge graph (KG) representations, and guided dimensionality reduction (Fig. 2). Our approach consists of five main steps:

1. Entity extraction with medical NER
2. Embedding generation using domain-specific transformers
3. Automated KG construction and graph embeddings
4. Dimensionality reduction with an autoencoder
5. Classification using a feedforward neural network (FFNN) with an attention layer

3.1. Dataset Description

We evaluate our pipeline on two sets of data: one synthetic and exploratory, and the other derived from clinical records.

- Synthetic Evaluation Terms: To test the dimensionality reduction and graph construction modules, we created sets of 40 medical terms using ChatGPT[38]. These were evenly split into two groups (e.g., psychological vs. cardiovascular symptoms), and labeled according to their semantic class. While limited in size, these sets serve to illustrate feasibility and separability under controlled conditions.

Table 1
Semantic types captured by MedCAT in this study. Note that misspellings ('bioplar') are correctly captured.

Semantic Type	Example Phrase	Entities Extracted	Normalized Entities
Mental or Behavioral Dysfunction	"...bipolar, PTSD, presented from OSH ED..."	bipolar PTSD	Bipolar Disorder Post-Traumatic Stress Disorder
Disease or Syndrome	"____ presented with HCV cirrhosis c/b ascites"	HCV cirrhosis ascites	Hepatitis C Liver Cirrhosis Ascites
Sign or Symptom	"... worsening abd distension and discomfort"	discomfort	Discomfort
Injury or Poisoning	"She had a food poisoning a week ago from eating stale cake"	food poisoning	Food Poisoning
Neoplastic Process	"... also receiving ongoing chemotherapy for liver cancer..."	liver cancer	Cancer (liver)

- Dayton Children's Hospital (DCH) Dataset: This dataset contains free-text reports from 100 de-identified patients. Reports were parsed to extract symptoms using MedCAT[39], and were evaluated for relevance to behavioral health counseling. Each report may contain multiple symptoms, and the classification task was binary at the patient level. After post-processing and label refinement, this dataset included 100 labeled examples and over 800 extracted symptom mentions.
- Medical Specialty Dataset (Kaggle): We used a subset of 430 reports across five specialties (gastroenterology, neurology, orthopedics, radiology, urology). Relevant symptom/procedure/body-part terms were extracted and embedded as centroids.

These datasets serve both as realistic and exploratory test beds for analyzing modular improvements in embedding, dimensionality reduction, and classification.

3.2. Named-Entity Recognition and Term Normalization

The pipeline begins with extracting medical terms from free-text clinical notes using an NER tool such as MedCAT, which leverages the UMLS metathesaurus for entity recognition and normalization [39]. The normalization this tool allows ensures consistency in extracted terms by mapping variations (e.g., "heart attack" vs. "myocardial infarction") to standardized concepts. MedCAT and other tools that use the UMLS metathesaurus (such as MetaMap) also allow for finding particular semantic types in terms. These semantic types are broad categories of medically relevant terms, like "Sign or Symptom", "Disease or Syndrome", or "Body Part". These allow for the extraction of sub-types of medical terms. In this study, we use the semantic types found in Fig. 3.

3.3. Transformer-Based Embeddings for Medical Terms

To encode extracted terms into dense vector representations, we employ BioBERTa, a transformer-based model pre-trained on biomedical corpora. This allows the pipeline to

capture semantic similarities between clinical concepts while preserving contextual relationships. These embeddings represent the "main path" of the pipeline. A simple pipeline would simply generate embeddings from NER-extracted terms, combine these embeddings, and give them as input to a classification model, a FFNN or other deep learning architecture. Indeed, this exact pipeline has been succesfully used in medical NLP, [40] but does suffer limitations of understanding complex clinical relationships between embeddings.
Multiple ways of extracting embeddings from symptoms are possible with this pipeline. For projects where an overall 'snapshot' of the patient's visit is desired, terms could be extracted, have embeddings created via transformer model, then the centroid of these embeddings could be taken to be representative of the overall visit embedding.
A known challenge in using LLMs for knowledge graph construction is hallucination—the generation of inaccurate or overly confident statements. To mitigate this, we employed a constrained prompt format with binary yes/no outputs and avoided free-form explanations. Additionally, nodes and relations were verified by comparing against known structured knowledge when available. However, we acknowledge that some errors may persist, and future versions will include ensemble querying, majority voting across multiple prompts, or grounding through biomedical knowledge bases such as UMLS or SNOMED CT.
In the context of Dayton Childrens' Hospital, the presence of any one symptom predisposing a patient for behavioral health counseling was sufficient to categorize a patient, so each symptom for each patient was embedded and considered separately for classification - any positive results would label the patient overall as needing counseling.

3.4. Knowledge Graph Construction and Graph Embeddings

To enrich term representations with structured domain knowledge, we integrate a lightweight knowledge graph (KG) constructed using outputs from a large language model (LLM). Extracted terms from the NER step are evaluated in relation to predefined clinical concepts using ChatGPT (GPT-4), which assesses whether a symptom or finding is meaningfully associated with a broader medical condition.

A consistent prompt was used throughout this process to ensure structured output: *"You are a medical notation expert. Indicate, with a one-word yes or no answer, if [symptom] is associated with [condition]."* Based on this response, edges were created in the knowledge graph. If the response was yes, an is_a edge was added between the term and the condition; if the response was no, an is_a edge was created to a node called [not condition].

The resulting knowledge graph consisted of nodes representing both extracted terms and reference categories (e.g., behavioral health counseling), and a small set of binary relation types. Although this initial work uses only is_a and is_not_a relationships, future versions of the pipeline will incorporate more clinically expressive relations such as treats, is_treated_by, causes, and is_caused_by, guided by structured ontologies or curated schema, such as the UMLS metathesaurus.

After KG construction, we used the TransE algorithm from the PyKEEN library[41] to generate graph embeddings for all nodes. These embeddings capture relational context and are concatenated with the transformer-based embeddings of each term prior to classification, enhancing both accuracy and semantic interpretability.

3.5. Dimensionality Reduction via Autoencoders

To improve classification efficiency and semantic separation, we optionally apply a guided dimensionality reduction step using an autoencoder. The autoencoder used for dimensionality

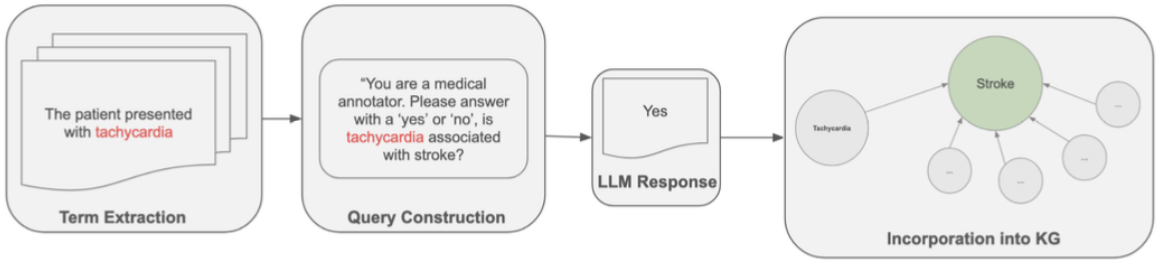

Figure 3: Automated query construction and KG building from NER.

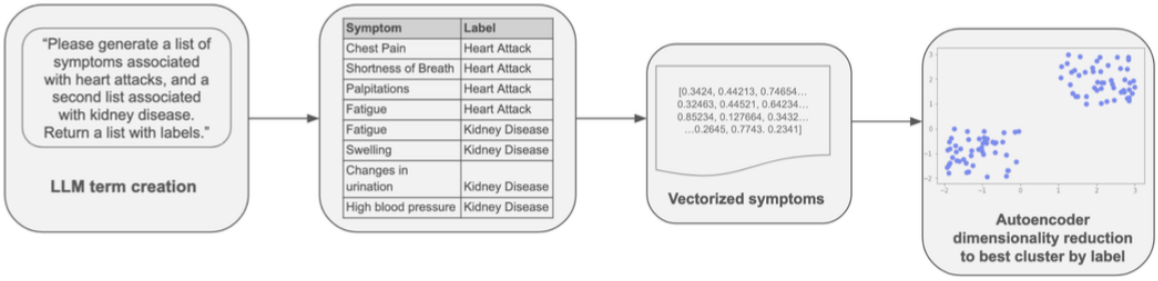

Figure 4: Optional guided dimensionality reduction via LLM-generated terms and labels

reduction consisted of three linear layers in both the encoder and decoder, with ReLU activation. The encoder compressed 768-dimensional BioBERTa embeddings to a latent space of 64 dimensions, using the architecture [768 → 256 → 128 → 64]. The autoencoder was trained with a mean squared error (MSE) loss for reconstruction, and supervised loss was added to encourage separation between semantically distinct classes during training. Training was conducted for 30 epochs with a batch size of 16 using Adam optimizer. All implementation was done in PyTorch.

The autoencoder is trained to optimize term clustering according to predefined categories, ensuring that similar terms are projected into distinct clusters based on clinical relevance. This step mitigates the high-dimensionality challenge of transformer embeddings while preserving key medical distinctions. For instance, in a project to find psychological symptoms, a number of symptoms relating to physical problems, and an equal number of symptoms relating to psychological problems are generated either by an expert or an LLM. These terms are embedded using the transformer model to be used in the NER step. The autoencoder is trained to create encodings that maximize accuracy of classification of these two groups into different clusters. In this way, information relating to the clinically relevant aspects of the terms are preserved in the encoding.

3.6. Classification with Feedforward Neural Network and Attention Layer

The final step involves concatenating the reduced BioBERTa[8] embeddings with their corresponding graph embeddings and passing them through an FFNN with an attention layer. The attention mechanism enhances classification performance by selectively weighting relevant features within the combined embeddings [42]. The FFNN is trained to classify patient records based on extracted symptom embeddings, enabling the identification of clinically relevant patient subgroups.

Our approach is designed to be modular and adaptable, supporting a range of clinical NLP tasks such as patient stratification, specialty classification, and behavioral health assessment. By integrating multiple representation techniques, our pipeline aims to improve accuracy, interpretability, and robustness in extracting complex patient subsets from medical free text.

4. Results

4.1. Sample classification of data using guided dimensionality reduction

Given the exploratory nature of this study, some evaluation datasets—such as those involving guided dimensionality reduction—were partially constructed using synthetic terms generated via ChatGPT, intended as proof-of-concept demonstrations rather than definitive clinical validations. To validate the autoencoder guided dimensionality reduction technique, 40 terms were generated by chatGPT, designed so they could be divided in half either by association with cardiovascular/psychological problems, or, in a different split, by their severity, mild or severe. Using sample terms generated by ChatGPT, 20 referencing cardiovascular problems, and 20 referencing psychological problems, an autoencoder and FFNN architecture was trained. This architecture was used to classify the 40 original terms. All of these terms were classified correctly. 40 new terms, synonyms for either “severe” or “mild” were then used to train a new autoencoder/FFNN architecture, and this correctly classified 37/40 terms. While the number of terms used in our synthetic evaluations is small, the goal of these experiments is not generalization but controlled testing of architectural components such as embedding separability and KG-based enrichment. Future work will expand evaluation to larger curated sets.
[1]

4.2. Automated construction of a knowledge graph, TransE creation of embeddings

To test the approach of automated KG construction, and the ability of TransE to generate spacially separated embeddings, ChatGPT was used to create 40 terms associated exclusively with one of two categories: allergic reactions to environmental allergens (pollen allergies, perfume allergies, gluten sensitivity), and 40 terms associated with medication allergies (hives from sulfa drugs, cough from ACE inhibitors, etc.). These terms were then given to a new session of ChatGPT, and asked to classify them as “environmental allergies” or “drug reactions”, and each term was added as a triplet to the knowledge graph containing nodes “environmental allergies” and “drug reactions”. Using the Pykeen library, the TransE algorithm was used to generate graph embeddings.

4.3. Application of the attention layer in classifying medical specialty

To test the utility of the attention layer, we utilized the Kaggle challenge “Medical Specialty Classification”[43]. In this test, a classifier must be designed that will determine the medical speciality that a report is associated with. This competition was especially useful for displaying the utility of the attention head on the end of the feed-forward neural network, because different aspects of a report (the symptoms, procedures, etc.) are important to consider in relation to each other to determine the specialty. For instance, the symptom "broken bone" might equally apply to orthopedics or radiology, while the procedure "lumbar puncture"

[1] The misclassified terms (“bipolar disorder” was classified as mild, not severe, and "palpitations" and "elevated cholesterol" were classified as severe, not mild) can be seen to reflect the subjectivity around a division such as severe/mild.

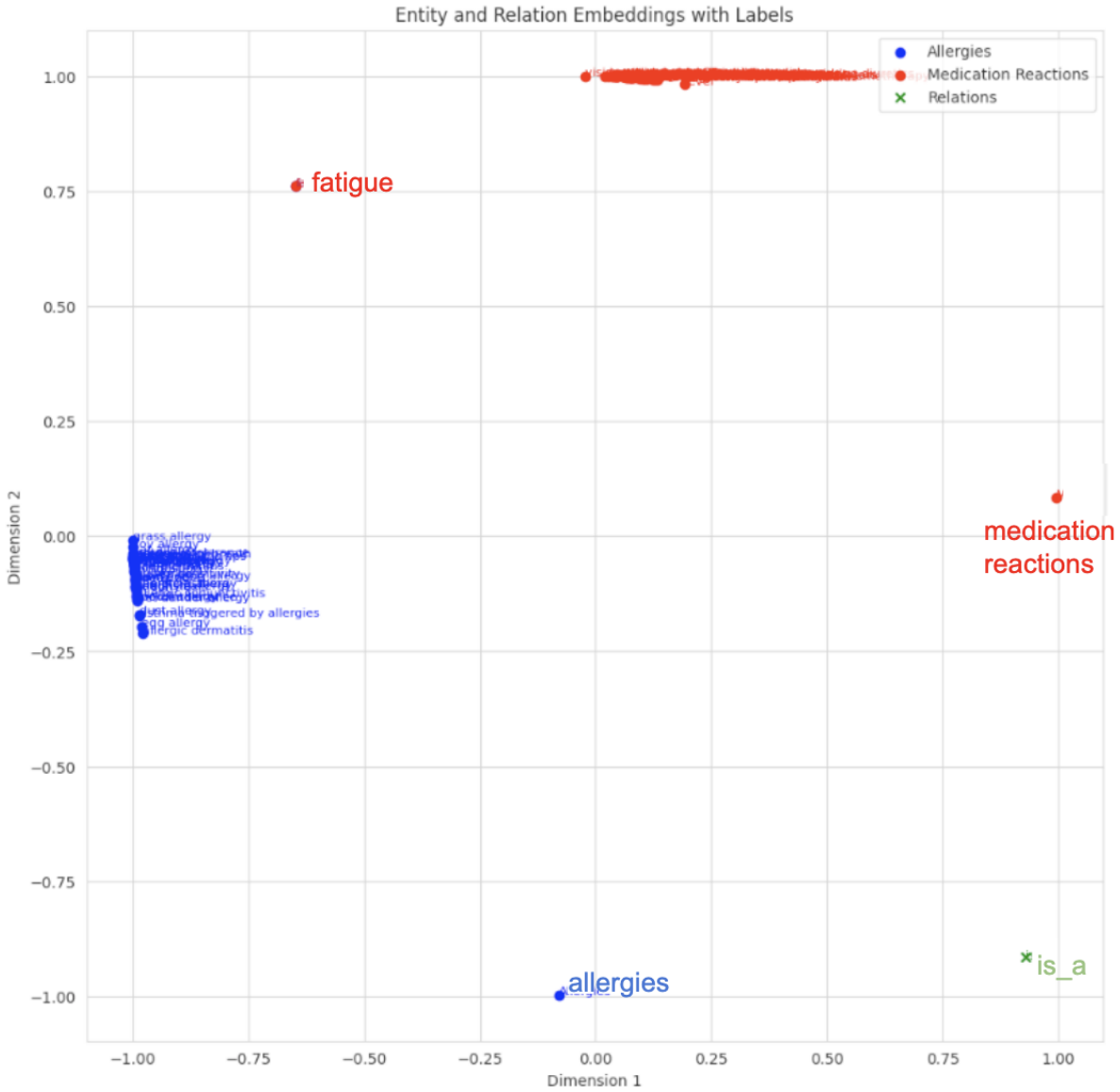

Figure 5: Figure 5: Graph embedding visualization of terms related to allergies (blue) and medication reactions (red). "Fatigue" (red) was added to both categories. Two primary clusters are visible based on TransE embeddings, though visual separation is subtle without cluster overlays. Clustering metrics (e.g., silhouette score: 0.9007) quantitatively confirm distinct grouping.

could easily apply to radiology or neurology. Understanding how multiple embeddings relate to each other can lead to a much clearer picture of the likely specialty. A subset of 430 of these reports, taken from the specialties: gastroenterology, neurology, orthopedics, radiology, and urology were selected. These reports had relevant terms extracted that matched medCAT categories for: symptoms, body parts, and medical procedures. Embeddings were created using the BioBERT sentence encoder. For each report, the centroid of the symptom, body part, and medical procedure terms was taken, and these centroid embeddings were stacked together. A FFNN with one hidden layer and relu activation was used to classify these reports. An attention layer was added before the FFNN, applied to the reduced-dimensionality concatenated centroids before FFNN classification. The addition of this layer improved the performance of the model from an accuracy of 0.8242 to 0.8837.

4.4. Classification of patients from Dayton Children's Hospital with encoder and FFNN

A preliminary test of the system using data taken from Dayton Children's Hospital (DCH), in which free-text patient reports were analyzed to determine if the patient they discussed would benefit from behavioral health counseling.
This is a complex determination, as it includes not only extraction of behavioral and psychological symptoms, but determining if these symptoms are frequently associated with behavioral health.
As an example, while some kinds of delirium would merit behavioral health counseling, as could be symptoms of an underlying psychological condition, "emergence delirium" is a side-effect of waking from anaesthesia, and is not an indicator for behavioral health counseling.
100 deidentified patients taken from Dayton Children's Hospital were put through a partial pipeline: their symptoms were extracted with NER, and embeddings created with the BioBERT-mnli-snli-scinli-scitail-mednli-stsb sentence transformer (BioBERT).
Additionally, guided dimensionality reduction was performed with ChatGPT-generated terms for 15 categories, corresponding to reasons for behavioral health counseling:

- suicide
- self-harm
- homicidal ideation
- assaultive behavior
- command hallucinations
- cognitive impariment
- dementing disorder
- mood disorder
- anxiety
- psychotic symptoms
- attention disorder
- eating disorder
- toxic reactions to psychiatric medication
- primary substance abuse[44]

Terms from patient reports were extracted with MedCAT, which also tracked negations. Embeddings were created using BioBert, were dimensionally reduced with the autoencoder, put through a FFNN, and were classified as needing behavioral health counseling if any symptom from a patient's report was classified as belonging to one of the provided categories (any output node output activation of >0.6).
This resulted in an initial F1 score of 0.9020. With a review of the missing categories, it was determined that insomnia and developmental delay were being tagged as positive signs (for the categories 'mood disorder' and 'cognitive impairment' respectively), despite not being indicators for behavioral health. Modification of the input terms resulted in an improvement of the F1 score to 0.9333. This kind of post-run generalization was implemented because of an updated understanding of the underlying problem, and reflects the benefit of review by subject-matter experts of the categories used for dimensionality reduction in this pipeline.
Terms were additionally provided to ChatGPT 4.0 using the prompt described in section 3. A TransE embedding of this knowledge graph was stacked with the reduced-dimensionality embeddings, and this change resulted in a further improvement of two reports being correctly labeled as positive, and an improvement of the F1 score to 0.95.

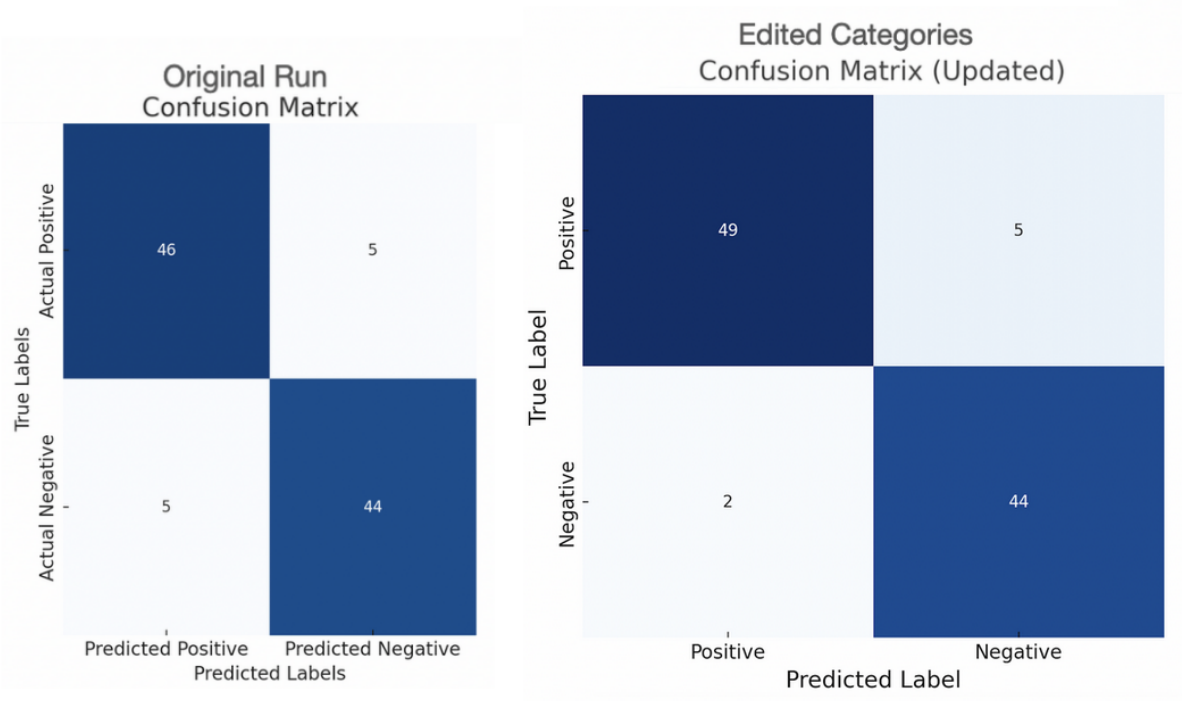

Figure 6: Confusion matrix of results using original training categories (left) and after modification to remove insomnia and developmental delay from categories (right)

5. Discussion

Our results demonstrate that integrating transformer-based embeddings, knowledge graph representations, and guided dimensionality reduction provides an accurate and sensitive method of patient subset extraction from clinical free text. The pipeline effectively captures relevant medical concepts, enhances term relationships through structured knowledge, and refines embeddings for improved classification.

One key advantage of this approach is its adaptability. By leveraging named-entity recognition for term extraction, and knowledge graphs to effectively place those terms within a clinical framework, the system can be applied to a variety of classification tasks, from identifying patients with specific conditions to categorizing reports by medical specialty. The guided dimensionality reduction step further enhances interpretability, allowing embeddings to be optimized for different classification goals. Additionally, the pipeline's modularity enables customization for domain-specific research questions, making it a flexible tool for clinical NLP applications.

Our results highlight the contribution of individual components to overall performance. Removing KG embeddings lead to a decline in F1-score, suggesting that structured knowledge provides valuable context to transformer-based embeddings. Similarly, disabling the attention layer in the feedforward neural network (FFNN) reduced classification accuracy, suggesting that selective weighting of features enhanced decision-making. These findings align with prior work showing that combining structured and unstructured knowledge can improve medical NLP tasks.

While the current pipeline demonstrates strong performance on curated datasets, generalizability to broader clinical settings, like full EHR narratives, must remain a major focus. Importantly, using EHRs as a source of grounding or enrichment (as structured fields, or unstructured, given to LLM-based knowledge extraction) domain drift, documentation variability, and coding inconsistencies may impact performance. Future extensions of this pipeline could integrate

EHR-derived ontologies or structured medication/problem lists to anchor LLM outputs more reliably and reduce hallucination risk. Establishing reliable data provenance and incorporating clinician feedback will be essential for broader deployment.
Of important note: our claim that guided dimensionality reduction improves interpretability is based on qualitative observation and separability metrics such as silhouette scores. Before any live clinical implementation of these systems will require structured expert evaluation, visualization, and possibly formal explanation tools such as SHAP or counterfactual analysis. Future work will incorporate domain expert feedback and formal interpretability benchmarks to validate these claims.
This work involves de-identified patient data and synthetic text generated for methodological validation. Nevertheless, ethical considerations remain central, particularly regarding privacy, consent, and fairness. The use of LLMs must be carefully monitored to prevent leakage of sensitive information or biased decision-making. In clinical deployment scenarios, data should be processed on-premises when possible, and all model outputs should be reviewed within a human-in-the-loop framework. Future research should include fairness audits and bias checks across different patient subpopulations to ensure equitable outcomes.

6. Conclusions and Future Work

We present a novel pipeline for automated patient subset extraction that integrates NER, transformer-based embeddings, knowledge graphs, and guided dimensionality reduction. By combining structured and unstructured representations, this approach improves classification accuracy and interpretability while maintaining adaptability for diverse clinical NLP tasks.
While these early results are encouraging, there remains much work to be done. The utilization of an existing KG framework for placing NER-derived terms in is one active area of work. Utilizing the UMLS API, a knowledge graph centered around NER-derived terms can be automatically created, and specific desired relations (like "treats", between a medication and a symptom" can be pulled out, allowing for a deeper knowledge graph structure.

Refining the way embeddings are handled is another area of work - simply stacking the graph embedding and term embeddings is a simple solution, and more advanced methods of combining embeddings, as well as alternative architectures to the classifier structure, which, currently, is just a FFNN with basic relu activation, is likely to yield gains.
Properly defining the circumstances under which different aspects of the model yield the greatest benefit is another area of ongoing work - the improvements seen after adding the knowledge graph, for instance, showed improvements only by converting false negatives to true positives. This one-sided improvement could be an artifact of the relatively small number of improperly-classified reports, and could benefit from a larger base of reports.
Finally, more integration with large language models is possible. From providing alternative annotations to build consensus for the decisions this model makes, to constructing knowledge graphs based on text documents provided for an individual institution's specific research goals, integration of LLMs is a promising endeavor for this work.
We are excited by the promise of this work, and look forward to further development of this pipeline

7. Acknowledgments

We are grateful for data provided by Dayton Childrens' Hospital, and the involvement and guidance of Dr. Katherine Winner.

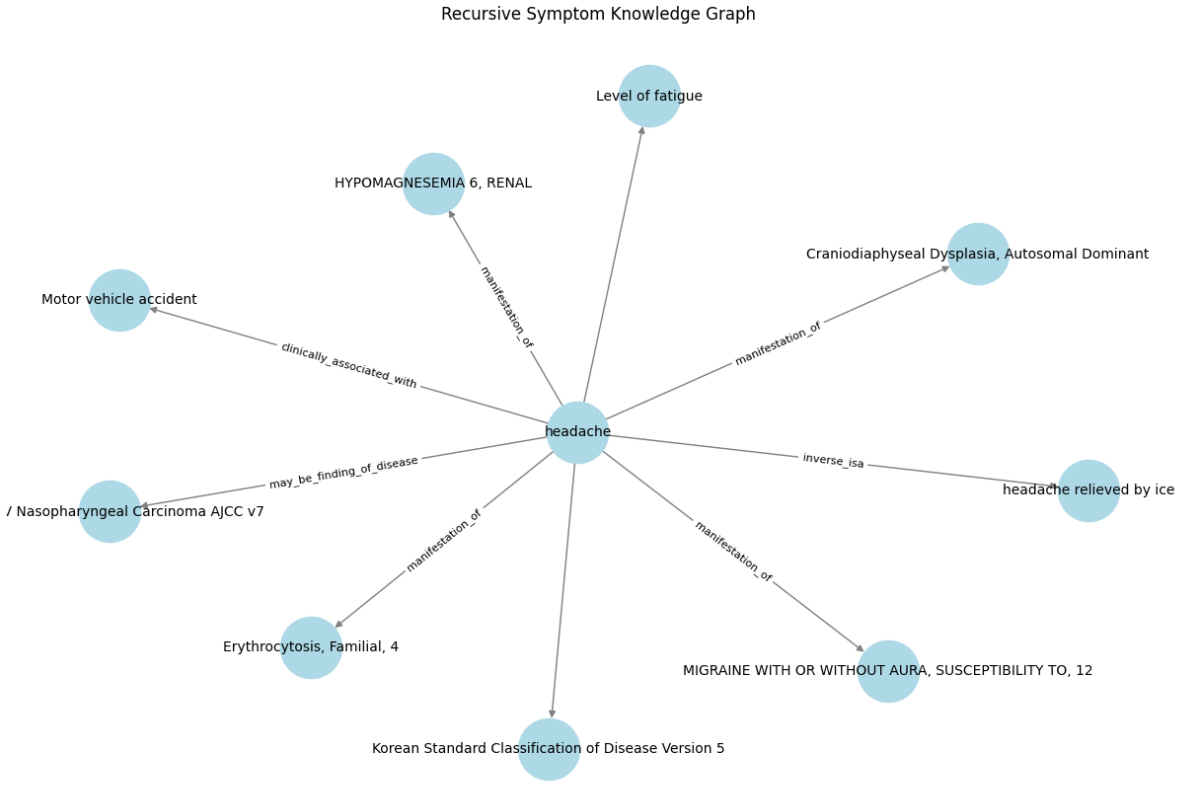

Figure 7: Knowledge graph created using the UMLS API with the symptom "headache"

8. Declaration on Generative AI

The authors did not use Generative AI in writing this paper

9. Bibliography

References

[1] N. I. of Health, Umls metathesaurus, 2016. URL: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html.

[2] H. Yang, J. Liu, Knowledge graph representation learning as groupoid (2021). doi:https://doi.org/10.1145/3459637.3482442.

[3] M. N. Zozus, C. Pieper, C. M. Johnson, T. R. Johnson, A. Franklin, J. Smith, J. Zhang, Factors affecting accuracy of data abstracted from medical records, PLOS ONE 10 (2015) e0138649. URL: https://pdfs.semanticscholar.org/32e4/62ddee2e63e3483fd97059152131a843bdab.pdf. doi:https://doi.org/10.1371/journal.pone.0138649.

[4] K. B. Johnson, W. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, J. L. Snowdon, Precision medicine, ai, and the future of personalized health care, Clinical and Translational Science 14 (2020). doi:https://doi.org/10.1111/cts.12884.

[5] A. Goel, A. Gueta, O. Gilon, C. Liu, S. Erell, L. H. Nguyen, X. Hao, B. Jaber, S. Reddy, R. Kartha, Llms accelerate annotation for medical information extraction, in: Proceedings of the 40th International Conference on Machine Learning, 2023. URL: https://proceedings.mlr.press/v225/goel23a.html.

[6] D. Newman-Griffis, G. Divita, B. Desmet, A. Zirikly, C. P. Rosé, E. Fosler-Lussier, Ambi-

guity in medical concept normalization: An analysis of types and coverage in electronic health record datasets, Journal of the American Medical Informatics Association 28 (2020) 516–532. doi:https://doi.org/10.1093/jamia/ocaa269.

[7] T. Minssen, E. Vayena, I. G. Cohen, The challenges for regulating medical use of chatgpt and other large language models, AI in Medicine (2023). URL: https://jamanetwork.com/journals/jama/fullarticle/2807167. doi:https://doi.org/10.1001/jama.2023.9651.

[8] K. Huang, J. Altosaar, R. Ranganathm, Biobert: a pre-trained biomedical language representation model for biomedical text mining, ArXiv (2019). doi:https://doi.org/10.48550/arXiv.1904.05342.

[9] P. Deka, A. Jurek-Loughrey, N. Deepak, Unsupervised keyword combination query generation from online health related content for evidence-based fact checking, Proceedings of iiWAS2021 (2021). doi:https://doi.org/10.1145/3487664.3487701.

[10] C. Fang, G. Alonso, I. Kagiampakis, M. H. Khalid, E. Jacob, K. C. Bulusu, N. Markuzon, Integrating knowledge graphs into machine learning models for survival prediction and biomarker discovery in patients with non–small-cell lung cancer, Journal of Translational Medicine 22 (2024). doi:https://doi.org/10.1186/s12967-024-05509-9.

[11] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, Artificial Intelligence Review 56 (2023). doi:https://doi.org/10.1007/s10462-023-10465-9.

[12] D. Dessì, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain, Future Generation Computer Systems 116 (2021). doi:https://doi.org/10.1016/j.future.2020.10.026.

[13] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. N. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, Computation and Language (2025). doi:https://doi.org/10.48550/arXiv.2404.16130.

[14] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, R. Mark, Mimic-iv, 2024. URL: https://physionet.org/content/mimiciv/3.1/.

[15] G. Karystianis, K. Thayer, M. Wolfe, G. Tsafnat, Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews, Journal of Biomedical Informatics 70 (2017) 27–34. doi:https://doi.org/10.1016/j.jbi.2017.04.004.

[16] W. K. Tan, S. Hassanpour, P. J. Heagerty, S. D. Rundell, P. Suri, H. T. Huhdanpaa, K. James, D. S. Carrell, C. P. Langlotz, N. L. Organ, E. N. Meier, K. J. Sherman, D. F. Kallmes, P. H. Luetmer, B. Griffith, D. R. Nerenz, J. G. Jarvik, Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain, Academic Radiology 25 (2018) 1422–1432. URL: https://www.sciencedirect.com/science/article/pii/S1076633218301211. doi:https://doi.org/10.1016/j.acra.2018.03.008.

[17] F. Zhu, B. Shen, Combined svm-crfs for biological named entity recognition with maximal bidirectional squeezing, PLoS ONE 7 (2012) e39230. doi:https://doi.org/10.1371/journal.pone.0039230.

[18] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, arXiv Preprint (2019). doi:https://doi.org/10.48550/arxiv.1904.05342.

[19] S. Gao, M. Alawad, M. T. Young, J. Gounley, N. Schaefferkoetter, H.-J. Yoon, X.-C. Wu, E. B. Durbin, J. Doherty, A. Stroup, L. Coyle, G. D. Tourassi, Limitations of transformers on clinical text classification, IEEE Journal of Biomedical and Health Informatics (2021) 1–1.

URL: https://ieeexplore.ieee.org/abstract/document/9364676. doi:https://doi.org/10.1109/JBHI.2021.3062322.

[20] Z. A. Nazi, W. Peng, Large language models in healthcare and medical domain: A review, Informatics 11 (2024) 57. doi:https://doi.org/10.3390/informatics11030057.

[21] M. Wornow, A. Lozano, D. Dash, J. Jindal, K. W. Mahaffey, N. H. Shah, Zero-shot clinical trial patient matching with llms, NEJM AI 2 (2025). doi:https://doi.org/10.1056/aics2400360.

[22] S. V. Shah, Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records, JAMA Network Open 7 (2024) e2425953. doi:https://doi.org/10.1001/jamanetworkopen.2024.25953.

[23] J. Chiat, S. Y.-H. Chang, W. William, A. J. Butte, N. H. Shah, L. Sui, N. Liu, F. Doshi-Velez, W. Lu, J. Savulescu, D. Shu, Medical ethics of large language models in medicine, NEJM AI (2024). doi:https://doi.org/10.1056/aira2400038.

[24] D. P. Panagoulias, M. Virvou, G. A. Tsihrintzis, Augmenting large language models with rules for enhanced domain-specific interactions: The case of medical diagnosis, Electronics 13 (2024) 320–320. doi:https://doi.org/10.3390/electronics13020320.

[25] M. Cheng, L. Li, Y. Ren, Y. Lou, J. Gao, A hybrid method to extract clinical information from chinese electronic medical records, IEEE Access 7 (2019) 70624–70633. doi:https://doi.org/10.1109/access.2019.2919121.

[26] D. Wu, L. Nie, R. A. Mumtaz, K. Agarwal, A llm-based hybrid-transformer diagnosis system in healthcare, IEEE Journal of Biomedical and Health Informatics (2024) 1–12. doi:https://doi.org/10.1109/jbhi.2024.3481412.

[27] R. Zhu, X. Tu, J. X. Huang, Utilizing bert for biomedical and clinical text mining, Elsevier eBooks (2020) 73–103. doi:https://doi.org/10.1016/b978-0-12-819314-3.00005-7.

[28] A. Turchin, S. Masharsky, M. Zitnik, Comparison of bert implementations for natural language processing of narrative medical documents, Informatics in Medicine Unlocked 36 (2023) 101139. URL: https://www.sciencedirect.com/science/article/pii/S2352914822002763. doi:https://doi.org/10.1016/j.imu.2022.101139.

[29] Y. Lan, S. He, K. Liu, X. Zeng, S. Liu, J. Zhao, Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion, BMC Medical Informatics and Decision Making 21 (2021). doi:https://doi.org/10.1186/s12911-021-01622-7.

[30] G. Zhou, H. E, Z. Kuang, L. Tan, X. Xie, J. Li, H. Luo, Clinical decision support system for hypertension medication based on knowledge graph, Computer Methods and Programs in Biomedicine 227 (2022) 107220–107220. doi:https://doi.org/10.1016/j.cmpb.2022.107220.

[31] F. Teng, W. Yang, L. Chen, L. Huang, Q. Xu, Explainable prediction of medical codes with knowledge graphs, Frontiers in Bioengineering and Biotechnology 8 (2020). doi:https://doi.org/10.3389/fbioe.2020.00867.

[32] W. Li, H. Zhou, J. Dong, Q. Zhang, Q. Li, G. Baciu, J. Cao, X. Huang, Constructing low-redundant and high-accuracy knowledge graphs for education, Lecture Notes in Computer Science (2023) 148–160. doi:https://doi.org/10.1007/978-3-031-33023-0_13.

[33] F. Dalvi, H. Sajjad, N. Durrani, Y. Belinkov, Analyzing redundancy in pretrained transformer models, arXiv (Cornell University) (2020). doi:https://doi.org/10.48550/arxiv.2004.04010.

[34] Y. Bao, Breaking the bottleneck advances in efficient transformer design (2025). doi:https://doi.org/10.20944/preprints202502.2271.v1.

[35] V. Bellini, T. Di Noia, E. D. Sciascio, A. Schiavone, Semantics-aware autoencoder, IEEE Access 7 (2019) 166122–166137. doi:https://doi.org/10.1109/access.2019.

2953308.
[36] T.-D. Le, R. Noumeir, J. Rambaud, G. Sans, P. Jouvet, Adaptation of autoencoder for sparsity reduction from clinical notes representation learning, IEEE Journal of Translational Engineering in Health and Medicine 11 (2023) 469–478. doi:https://doi.org/10.1109/jtehm.2023.3241635.
[37] N. Sadati, M. Z. Nezhad, R. B. Chinnam, D. Zhu, Representation learning with autoencoders for electronic health records: A comparative study, arXiv (Cornell University) (2018). doi:https://doi.org/10.48550/arxiv.1801.02961.
[38] OpenAI, Chatgpt (oct. 2024 version), 2024. URL: https://chat.openai.com/chat.
[39] Z. Kraljevic, D. Bean, A. Mascio, L. Roguski, A. Folarin, A. Roberts, R. Bendayan, Richard, Medcat – medical concept annotation tool, arXiv (Cornell University) (2019). doi:https://doi.org/10.48550/arxiv.1912.10166.
[40] Z. Yu, X. Yang, G. L. Sweeting, Y. Ma, S. E. Stolte, R. Fang, Y. Wu, Identify diabetic retinopathy-related clinical concepts and their attributes using transformer-based natural language processing methods, BMC Medical Informatics and Decision Making 22 (2022). doi:https://doi.org/10.1186/s12911-022-01996-2.
[41] M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, S. Sharifzadeh, V. Tresp, J. Lehmann, Pykeen 1.0: A python library for training and evaluating knowledge graph embeddings, Machine Learning (2000). doi:https://doi.org/10.48550/arXiv.2007.14175.
[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. URL: https://arxiv.org/abs/1706.03762. doi:https://doi.org/10.48550/arXiv.1706.03762.
[43] T. K. Comp, Kaggle medical specialty competition, 2025. URL: https://www.kaggle.com/competitions/medical-specialty-classification.
[44] U. of Toldeo Nursing, U of toledy behavioral health criteria, 2025. URL: https://www.utoledo.edu/policies/utmc/nursing/unit/senior-behavioral-health/pdfs/3364-120-6-admission-criteria.pdf.