Writing assignment

Applications of Large Language Models (LLMs) in Healthcare

Famke Schulting,
MSc. Bioinformatics & Biocomplexity, Utrecht University
Supervisors: Tuur Leeuwenberg[1], Anne de Hond[1]
Second Reviewer: Anneke Damen[1]
[1]University Medical Center Utrecht, Utrecht
26 January 2024

Table of Contents

Abstract

In recent years, large language models (LLMs) have revolutionized language-related applications across various fields. These powerful models, trained on massive datasets, can understand and generate natural language for tasks like summarization, question-answering, and information extraction. One notable application of LLMs is their integration into the healthcare sector. Although numerous clinical applications have been proposed, from extracting medication information to providing patient support through chatbots, widespread implementation is still in progress.

This review aimed to identify clinical tasks that make use of LLMs and can be applied within the healthcare sector, from relevant literature. From the 1008 founded publications, a random subset was included in this review. After thorough screening, 129 clinical tasks were described within the resulting 89 publications. These clinical tasks were categorized into the overarching tasks: 'Clinical workflow', 'Patient education and communication' or 'Healthcare management'. The categorization of these clinical tasks, as well as the identification of the underlying classical NLP tasks, aimed to provide a comprehensive understanding on the described clinical tasks and the potential capabilities of utilizing LLMs in healthcare.

Although many utilities of LLMs in healthcare were described, the majority was not yet implemented within clinical settings. This indicates that the future holds promise for the widespread implementation of these clinical tasks, but further development and validation are essential for realizing their full potential in transforming healthcare services.

Layman Summary

In recent years, large language models (LLMs) have changed how we use language in different areas. These models, which are really powerful and trained on huge amounts of data, can understand and generate natural language and can be used for tasks like summarizing text, answering questions, and extracting information. These kinds of tasks can also be utilized within healthcare settings. LLMs can for instance summarize medical documents for patients, help doctors in making clinical decisions or automating tasks like patient appointment scheduling. However, there is currently no clear overview on what kind of tasks can be used in healthcare and how these can be categorized.
Therefore, within this review, literature was investigated to identify different clinical tasks that make use of LLMs. In total, 129 different clinical tasks were found, with most of them being related to the clinical workflow, followed by patient education and communication, and healthcare management. By categorizing these founded clinical tasks, an overview was made, focusing on the language tasks and their impact on healthcare. While a lot of clinical tasks were described in the literature, the majority of these was currently not being used in healthcare. This shows that even though there is a lot of promise in using LLMs for clinical tasks, future work is still needed to implement these tasks in healthcare.

Introduction

The emergence of large language models (LLMs) has facilitated the development of many language-orientated applications in multiple domains. LLMs, which are Transformer models that are trained on massive datasets, possess the capability to understand and generate natural language for a variety of natural language processing (NLP) tasks, such as text summarization, question-answering and information extraction. Especially since the launch of ChatGPT by OpenAI in November 2022, chatbots have gained immense popularity in a short time. Currently, ChatGPT boasts over 100 million weekly active users who employ the interacting chatbot for a wide array of purposes, including informational queries, educational or creative writing assistance, language translation as well as coding and programming help (1).

Another purpose of utilizing chatbots or LLMs is their integration within the healthcare sector. Despite numerous proposed clinical LLM applications, which span from extracting medication information to identifying potential adverse drug effects, generating clinical reports, and providing patient support through chatbot systems, the overall implementation is not yet widespread. However, this integration is becoming increasingly prominent as it holds the potential to profoundly reshape healthcare delivery.

While previous research outlines NLP tasks based on publicly available electronic health record data from patients (2), a clear description of the exact clinical NLP tasks that can be used in healthcare settings is lacking. Having this description is crucial, as it facilitates a comprehensive understanding of how the described clinical tasks can be categorized and what the potential capabilities of utilizing LLMs in healthcare are.

Therefore, this review aims to create a comprehensive overview of clinical tasks that make use of LLMs, that have been described in literature. These described clinical tasks only involve LLM applications that are applicable within a clinical setting, thereby excluding those intended for medical education and research. Moreover, these existing publications on LLMs in healthcare will be categorized, focusing on the overarching clinical application, but also the underlying classical NLP task, as well as the end-user.

Background of LLMs

Short history of natural language processing (NLP)

The origins of NLP can be traced back to 1950, when Alan Turing proposed the 'Turing test', in order to determine a machine's ability to achieve human intelligence (3). From the 1950s to 1980s, NLP was mainly focused on applying rule-based approaches, that were designed based on linguistic rules and patterns. However, due to complexity and flexibility of human language, this was a difficult task. Since the 1980s, a shift was seen, as statistical NLP systems were designed that were able to extract features from the large amount of digitally available texts. The utilization of statistical and machine learning algorithms gradually replaced the rule-based NLP systems, as machine learning models were able to learn linguistic patterns in a text without following a set of fixed rules (4).

In 2003, Bengio et al. published a paper describing a neural probabilistic language model. This model departed from the use of probabilities in a statistical language model. Instead, the model employed the concatenation of word embeddings through a one-hidden layer feed-forward neural network to predict the next word in a sequence (3).

The evolution of NLP took a significant leap in 2017, when Vaswani et al. introduced the Transformer model architecture, thereby replacing previously used recurrent neural networks (5), long short-term memory networks (6) and Word2Vec (7). Transformers utilize self-attention mechanisms, which enable the model to consider different parts of the input sequence when processing each element. This enables Transformers to process input in parallel and learn long-term dependencies, thereby enhancing the model's capability to better comprehend context and relationships within sentences.

Large language models (LLMs) are Transformer models, which are extensively trained on large datasets. A popular model is BERT (Bidirectional Encoder Representations from Transformers). In contrast to directional models, BERT reads text input sequentially, meaning that during training the entire sequence of words from both the left and right direction are taken into account (8). This bidirectional approach enables the BERT model to comprehensively understand the context and relationships within a given text.
As mentioned before, another well-known model is GPT (Generative Pre-trained Transformer), which is developed by OpenAI. GPT takes a generative approach, as it is trained to predict the next word or token in a sequence based on the context provided by the preceding words. By learning statistical patterns and relationships in the training data, GPT can generate text that is both coherent and contextually relevant.

Description of NLP tasks

The classical NLP tasks cover a diverse set of objectives, that are all aimed at enabling machines to understand, interpret, and generate human language. Within the clinical tasks that are described within this review, multiple classical NLP tasks, as described by Gao et al. (2) could be identified (Table 1).

Table 1. Overview of classical NLP tasks, description and examples.

Classical NLP task	***Task description***	***Example***
Text Generation (TG)	Generate text based on given inputs, aiming to achieve the appearance of being indistinguishable from human-written text	Generation of a patient report based on information from the electronic health care record
Summarization (Summ)	Summarize the main ideas of given a piece of text.	Summary of patient letter in layman style
Question answering (QA)	Providing answers to questions raised by humans in a natural language.	Use of a chat prompt to ask medical question
Named Entity Recognition (NER)	Identify and classify named entities from texts.	Example of identified entities: *Disease: chronic kidney disease* *Medication: epinephrine* *Surgery: tonsillectomy*
Information Extraction (IE)	Extraction of entities and relations from a text	"Patient Jane Smith was diagnosed with diabetes and prescribed insulin" could be processed by IE to extract the information that "Jane Smith" has a diagnosis of "diabetes" and has been prescribed "insulin."
Speech Recognition (SR)	Convert human speech into text information.	Generation of clinical interview transcripts based on a patient encounter

Methodology

Inclusion criteria

For this scoping review, relevant academic literature published from 1 January 2010 up to 14 December 2023 was searched on Pubmed (https://pubmed.ncbi.nlm.nih.gov/). These data were selected to ensure a comprehensive coverage of relevant publications, with 2010 chosen as the starting point, given the prominence that LLMs have gained over the last decade. The search string consisted of a combination of terms that were related to 'large language models' and 'healthcare' (Supplementary Table 1).
Due to time limitations, a randomized sampling procedure was performed, in which 100 publications from the total found publications, were included in the screening procedure. This screening procedure entailed screening both the title and abstract of each paper, to make sure that the publication matched our inclusion criteria. As the aim of this review was to identify LLM applications in healthcare that can applied within a clinic, all publications related to medical education and research were excluded.
Additional literature was included through manual inspection of the reference lists of identified documents to ensure a level of saturation on the various clinical tasks reported.

Data synthesis and summarization

Among the papers included in this review, characteristics of the clinical LLM application were summarized into tables. Following Jianning Li et al.'s categorization (8), the identified applications were categorized into overarching clinical tasks, including the 'Clinical workflow,' 'Patient communication and education,' and 'Healthcare management.' In short, 'Clinical workflow' refers to all the processes involved in patient care, such as diagnosis, treatment, and follow-up. 'Patient communication and education' involve both interactions between healthcare providers and patients, as well as providing medical information and guidance to patients. 'Healthcare management' pertains to the organization and administration of healthcare services.

These described overarching clinical task could be linked to different end-users, including ‘patient/relatives’, ‘healthcare professionals’, ‘clinical centers’ and ‘lawyers and regulators’. Besides documenting the clinical task description and the clinical relevance of this task, also the classical NLP tasks (e.g. ‘Information extraction’ or ‘Text generation’) that could be identified, were documented. Furthermore, the type of input data and output data was described. Similair to Jianning Li et al. (9), a tag (level 1 – 3) was assigned to the selected papers, which indicated the depth and particularity in which the clinical task was described. Where level 1, indicated that the task was only described by using generic comments (e.g. "ChatGPT can aid patients before radiologic-guided procedures." (10)), level 2 papers provided some more depth about the clinical task as a specific medical specialty and/or scenario for the task was described (e.g. "Our system makes the simplifying assumption that patient-level smoking status determination can be achieved by accurately classifying individual sentences from a patient's record." (11)). The papers that had a level 3 tag, described besides the clinical task itself, also the purpose and/or relevance of the clinical task, (e.g. "This study developed and validated a rule-based classification algorithm for prediabetes risk detection using natural language processing from home care nursing notes." (12)). As a result, level 3 papers offer a clearer depiction of actual capability of LLMs in healthcare. To assess the current extent of LLMs integration in healthcare, it was also documented whether the specified tasks had already been implemented in clinical settings. A full overview of the extracted form can be found in Supplementary Table 2.

Results

Using the search string, a total of 1008 publications were found for review. From these, a random subset of 100 publications was included in the data analysis. This number represented a saturation point at which no new clinical LLM applications could be identified. After the first review phase, 27 papers were excluded based on the exclusion criteria. A total of 16 papers was added after manually inspecting the reference lists of the included articles. This led to a total of 89 publications included in the review (Figure 1).

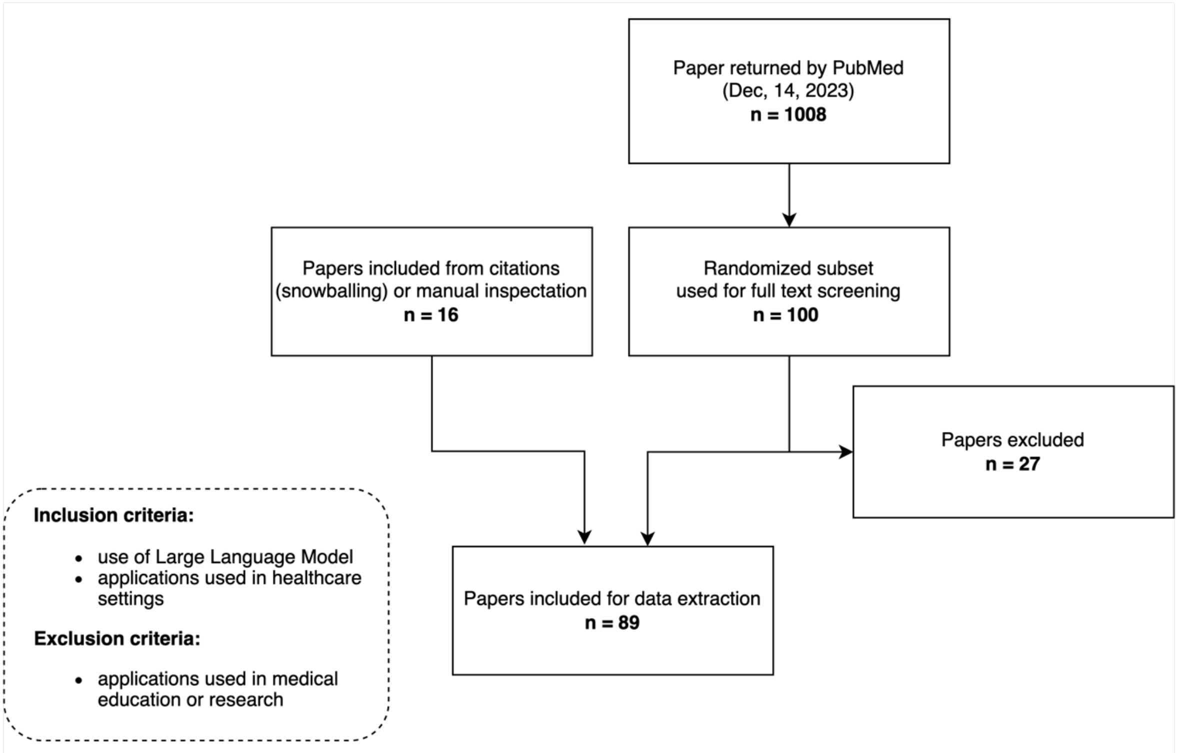

Figure 1: Flow diagram of screening strategy

Within the found publications, 129 clinical tasks were identified. Out of these described tasks, the majority (97 tasks) could be applied within the clinical workflow (Figure 3). The second group of 23 reported tasks was utilized for patient education or communication (Figure 2). Only 4 of the described tasks had usage in healthcare management (Figure 3).

Patient education and communication

The clinical tasks that were described within the patient education and communication field often relied on chatbots, utilizing a classic NLP task known as 'Question Answering'. The usage of chatbots allows patients to receive support whenever healthcare professionals are not available. For instance, a chatbot service can provide patients with intensive behavioral counseling recommended for weight management programs for obesity and support behavior change by applying various mental health approaches (13). Chatbots can also be

used to help patients to maintain a healthy life style by supporting physical activity, healthy eating or assisting with tobacco cessation (14,15). Besides providing support, chatbots were also used to provide patients with personalized education materials, on diabetes self-management education (16), otolaryngology-related information (17), or about ophthalmic plastic and reconstructive surgery topics (18). Additionally, the usage of chatbots to answer medical questions on various topics were often reported (19–27).
Moreover, chatbots can improve health literacy by answering patient questions about insurance coverage of specific services or clinical research participation (28).
Such applications are valuable as they alleviate the burden on healthcare professionals, thereby offering financial benefits as well.

The NLP task 'Text Generation' and 'Summarization' was also often described. LLMs can convert medical information or letters in layman style in order to make them accessible and understandable for patients (28–34).

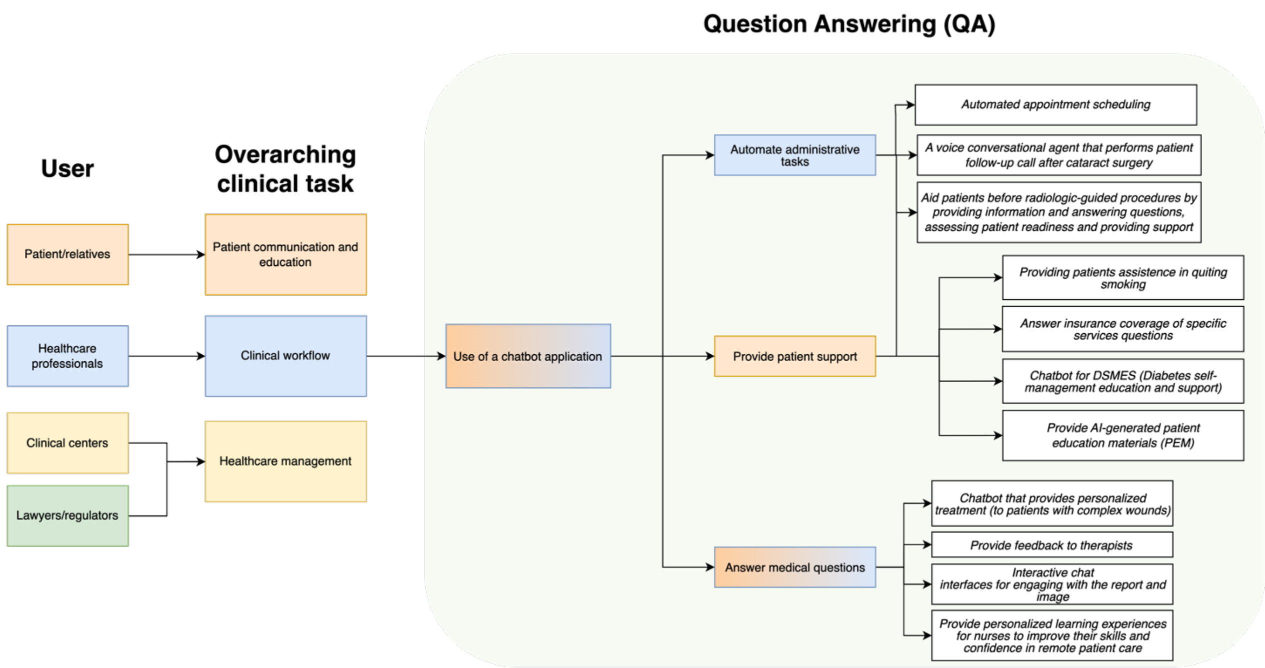

Figure 2: Overview of LLM applications in healthcare that utilize the classical NLP task of 'Question Answering' (QA)

Clinical workflow

Within the clinal workflow many different clinical tasks were described. For instance, just as chatbots can provide patients with personalized education materials, healthcare professionals can also access personalized learning material and recommendations for remote patient monitoring (31,35), as well as receiving feedback based on free-text

transcripts of patient encounters (36). Another 'Question Answering' NLP task within the clinical workflow is the use of interactive chat interfaces that allow healthcare professionals to engage with the report and image (30) or providing answers to medical questions (37,38). Gupta et al. describe how a chatbot can provide personalized treatment and lifestyle recommendations for patients with complex wounds based on the provided patient information, wound characteristics and medical history (39).
Furthermore, Wang J. et al. developed 'PhenoPad', which is an intelligent clinical note-taking interface that can capture free-form notes and phenotypic information from various modalities, such as speech and NLP techniques, but also from handwriting recognition (40).

The utilization of free-text electronic health records (EHRs) to extract information was often described in literature. The classical NLP tasks 'Information Extraction' and 'Named Entity Recognition' can be applied for multiple purposes. For instance, the extraction of medication information from EHRs can be useful to detect substance abuse (41,42), identify potential adverse drug effects (27,33) or pharmacovigilance in general (33,43,44).
The task of identifying smoking status from EHRs has also been reported, which is important for both informing and improving the assessment of smoking behavior, as it impacts not only diagnosis but also treatment options (11,42).

Moreover, by extracting clinical outcomes from EHRs, predictive analysis can be performed (45–47), such as the prediction of seizure recurrence (48), overall survival of rectal cancer patients (49) or readmission following an acute myocardial infarction (50). Furthermore, extracting temporal information from EHRs enables the generation of a complete timeline of a patient's medical events, which is crucial for many medical reasoning tasks (51).
Another common use of 'Information Extraction' involves its application in clinical decision support (20,25,27,29,30,33,40,52–60). One of the described functionalities involves incorporating LLMs in the process of radiologic decision-making. This is achieved by assessing the appropriate imaging modalities for various clinical presentations related to breast cancer screening and breast pain (54).

The extraction of predefined or ontology-based entities, such as specific medical terms, categories, or concepts, from patient-generated data or EHRs can be used for multiple purposes (33,61–65). Firstly, the extracted ontologies can be used to characterize clinical characteristics of patients to aid disease surveillance (31,34,66–69), such as in cystic fibrosis (70) or surveilling heart failure symptoms (71,72).
These observations are useful for disease classification (33,34,47,56,73), which is for instance useful for diseases like glaucoma which is difficult to identify in early stages, due to existence of variation on physiologic characteristics (74).
Moreover, extracted ontologies can aid in identifying risk groups for diseases like prediabetes (12) or can be used to identify suspicious findings in breast cancer (75).

Lastly, ontology-based entities are used to extract phenotypical information that is described in pathological reports (76), for detailed patient-specific data for biosurveillance (77) or to extract prognostic and predictive biomarker status of breast cancer patients (78).

Extracting clinical information extends beyond medical conditions, as it is also used to capture information from goals-of-care discussions between patients and healthcare professionals (79,80). This involves documentation of patients' values, goals and treatment preferences or end-of-life discussions, which are very important in palliative care (81,82). In palliative care, manual chart reviews are frequently used, even though the majority of crucial information regarding end-of-life care is found in free-text notes. Utilizing NLP tools can facilitate the extraction of this information, thereby enhancing palliative care consultations.

Another clinical LLM application involves patient cohort selection for trials. This can be achieved by extraction information from EHRs, but also by extracting the clinical trial eligibility criteria from clinical trial protocols (20,33,34,78,83–85). Patient-trial matching is generally conducted manually, which typically requires a significant amount of time from clinicians or administrative staff with specialized knowledge. By automating this process, both time and resources can be saved.

Besides extracting information EHRs, the classical NLP task of 'Named Entity Extraction' is also used in the (re)structuring of clinical notes. This can include rewriting clinical notes to include a specific metric in a report (30) or integrating clinical information from various sources (45,86). Another application that has been described by Jiu J et al. is the removal of duplicated information from EHRs, as this has been shown to have a positive impact on clinical NLP models (87). Furthermore, the conversion from free-text notes into structured reports is also often described, as it enhances communication, promotes collaboration among health care professionals and standardizes reporting language across institutions (19,30,32,88,89). Lastly, de-identification of clinical records is another application, that is critical to facilitate the use of unstructured clinical records while protecting patient privacy and confidentiality (61,90–92).

The classical NLP task 'Text generation' and 'Summarization' have also been described within the clinical workflow. Healthcare professionals can for instance use LLMs to generate (summarized) patients reports (25,30,93,94) or write patient letters or discharge summaries (45,95). The generation of these medical documents results in time savings for healthcare professionals.

Finally, LLMs can also contribute the clinical workflow by automating administrative tasks of healthcare professionals (28). A chatbot can for instance be used to automate appointment scheduling (19,25). But it can also prepare patients before certain procedures. Ismail A et al.

described how LLMs can be used to aid patient before radiologic-procedures by providing the patients with information and support, while also assessing their readiness for the procedure (10). Also after a certain procedure, such a cataract surgery, voice conversational agents can perform follow-up calls (96).
Another clinical NLP application, as described by Were MC et al., involves extracting provider follow-up information from discharge summaries. This task is crucial, as medical errors frequently occur during the transition of care from the inpatient to outpatient setting. This is particularly true in cases where there is a requirement to contact the follow-up provider long after the patient has been discharged from the hospital (97).
The automation these clinical repetitive tasks, enables healthcare professionals to be redeployed to higher-value activities, where their skills can best be used.

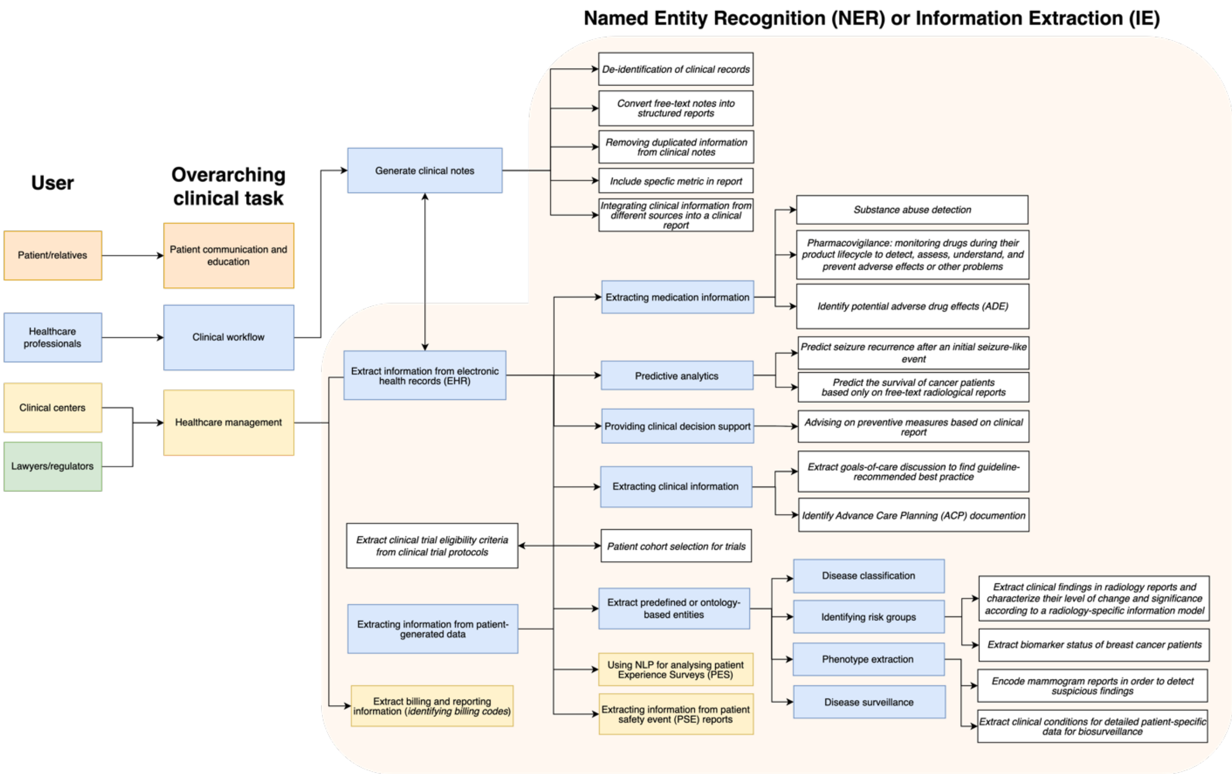

Figure 3: Overview of LLM applications in healthcare that utilize the classical NLP task of 'Named Entity Recognition' (NER) and 'Information Extraction' (IE)

Healthcare management

In the field of healthcare management, the NLP task 'Information Extraction' can also be utilized. One application is the extraction of information from patient experience surveys. These reports are leveraged to evaluate accessibility, communication and overall satisfaction, aiming to improve patient experience (98). Furthermore, patient safety event reports, in which adverse events and errors in healthcare are documented, can be utilized to extract contributing factors that led to the event. This will help safety officers to address safety issues that occurred within the organization (99).

Another application involves the extraction of billing codes from EHRs. The identification of these billing codes is crucial for healthcare organizations, as it enables them to receive payment for the provided healthcare services (33).

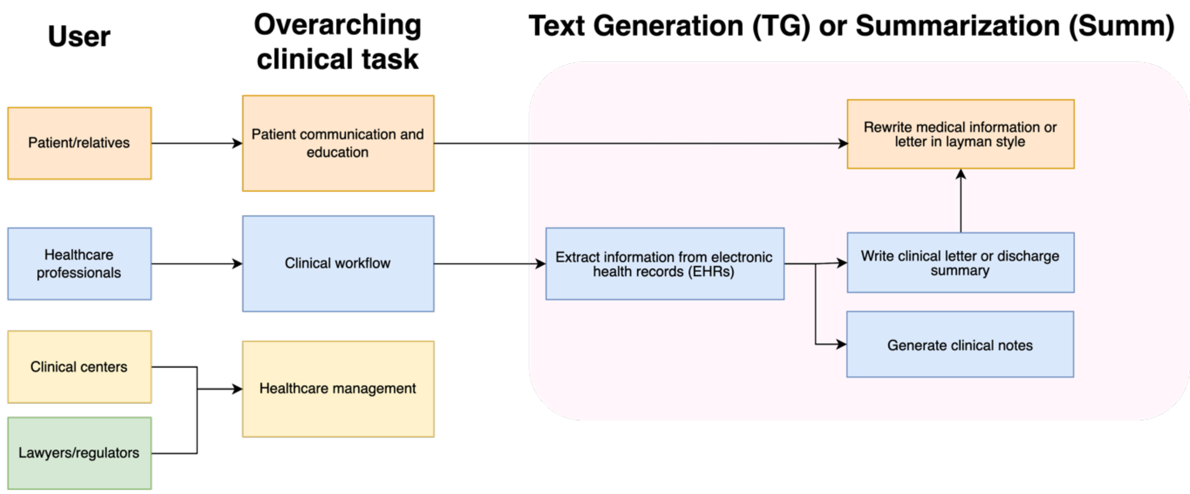

Figure 4: Overview of LLM applications in healthcare that utilize the classical NLP task of 'Text generation' (TG) and 'Summarization' (Summ)

Current usage and outlook

Out of all the publications, only five reported to have conducted a trial within in a clinic, meaning the majority of identified clinical tasks have not yet been implemented within clinical settings. Furthermore, the level in which the clinical task was described was often found to be either level 1 or 2, meaning the exact relevance or expected usage of the clinical task was not described. For instance, Wu H et al. describe the clinical task of disease surveillance by extracting disease information from EHRs (34). However, the specific details of the extracted information and its practical implications for healthcare professionals are not mentioned. Another example is the publication by Guergana K. Savova et al., which addresses the identification of the patient's smoking status from EHRs (11). However, also here the publication fails to elaborate on how this information can be utilized in healthcare.

Discussion and Conclusion

This scoping review provides a comprehensive overview of the applications of LLMs in healthcare, that can be utilized in clinical settings. The found clinical applications could be assigned to different overarching clinical tasks, including ‘Patient education and communication’, ‘Clinical workflow’ and ‘Healthcare management’, which each focused on different end-users and tasks. Additionally, within each sector, the various classic NLP tasks underlying the clinical workflows were identified.

The 'Patient education and communication' sector primarily involves utilizing the classical NLP task 'Question Answering.' Chatbots play a key role in enabling patients to ask medical questions, seek guidance, and receive support related to specific diseases or conditions. Additionally, the 'Text generation' NLP task is employed to create personalized education materials for patients and translate medical documents into layman's terms.
Within the ‘Clinical workflow’, many different clinical tasks were identified. The classical NLP tasks 'Question Answering' and 'Text generation' are again utilized, but with a focus on providing healthcare professionals with medical information and assisting them in routine tasks such as generating patient reports or handling administrative tasks like appointment scheduling. Within the clinical workflow, the NLP task of ‘Information Extraction’ was most often described, as the clinical information from EHRs proves to be valuable for diverse clinical tasks. The extracted ontologies could subsequently be utilized for diagnosis, disease surveillance, predictive analytics, clinical decision support or patient cohort selection.
In the 'Healthcare management' sector, the 'Information Extraction' task could also be employed. For instance, extracted data from patient surveys can be utilized to measure overall satisfaction. Similarly, identifying billing codes from EHRs is crucial for the accurate processing of payments.

While existing literature on LLM applications in healthcare outlines numerous potential clinical tasks, many have not been implemented yet. Additionally, the specific purposes of these tasks are often not clearly defined, as generic statements like 'AI models may assist in clinical decision support, clinical trial recruitment, clinical data management, research support, patient education, and other fields' (100) were frequently encountered. This shows that the interest in the application of LLMs in healthcare is growing, but the actual implementation of the described clinical tasks is still in development.

One limitation of our scoping review methodology is the random sampling procedure, where only 100 publications out of the 1008 identified by the search string were included in the screening process. However, upon thorough inspection of this subset and the additional publications identified through manual inspection of reference lists, we believe that a level of saturation was achieved. This allowed us to identify, if not all, at least the most prevalent applications of LLMs in healthcare.

Another limitation is that this review did not specifically address or validate the accuracy of the output generated by the described clinical tasks. However, one could argue that this was beyond the scope of the review, as its primary purpose was to offer an overview of the described clinical tasks that have or may hold promise for future implementation.

One thing to appreciate from this review, is the identification of both the classic NLP task and its clinical purpose within the described tasks. Providing this information is crucial for understanding how the described clinical tasks can be categorized and sets our review apart from previous work.

In conclusion, this review was able to provide an overview on the applications of LLMs in healthcare settings, focusing both on the underlying classical NLP tasks, as well as their clinical implications. Although current literature shows the potential of utilizing LLMs in healthcare, their widespread implementation is still on the horizon, representing a focus for future developments.

References

1. Shewale R. ChatGPT Statistics — User Demographics (January 2024) [Internet]. 2024 [cited 2024 Jan 26]. Available from: https://www.demandsage.com/chatgpt-statistics/

2. Gao Y, Dligach D, Christensen L, Tesch S, Laffin R, Xu D, et al. A scoping review of publicly available language tasks in clinical natural language processing. Journal of the American Medical Informatics Association. 2022 Sep 12;29(10):1797–806.

3. Turing AM. Computing Machinery and Intelligence. Mind. 1950 Oct 1;LIX(236):433–60.

4. Zhou B, Yang G, Shi Z, Ma S. Natural Language Processing for Smart Healthcare. IEEE Rev Biomed Eng. 2022;1–17.

5. Medsker L, Jain LC, editors. Recurrent Neural Networks: Design and Applications. In: 0 ed. CRC Press; 1999 [cited 2024 Jan 18]. Available from: https://www.taylorfrancis.com/books/9781420049176

6. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997 Nov 15;9(8):1735–80.

7. Church KW. Word2Vec. Nat Lang Eng. 2017 Jan;23(1):155–62.

8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv; 2019 [cited 2024 Jan 18]. Available from: http://arxiv.org/abs/1810.04805

9. Li J, Dada A, Kleesiek J, Egger J. ChatGPT in Healthcare: A Taxonomy and Systematic Review [Internet]. Health Informatics; 2023 Mar [cited 2023 Dec 11]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.03.30.23287899

10. Ismail A, Ghorashi NS, Javan R. New Horizons: The Potential Role of OpenAI's ChatGPT in Clinical Radiology. Journal of the American College of Radiology. 2023 Jul 1;20(7):696–8.

11. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo Clinic NLP System for Patient Smoking Status Identification. J Am Med Inform Assoc. 2008;15(1):25–8.

12. Jeon E, Kim A, Lee J, Heo H, Lee H, Woo K. Developing a Classification Algorithm for Prediabetes Risk Detection From Home Care Nursing Notes: Using Natural Language Processing. Comput Inform Nurs. 2023 Jul 1;41(7):539–47.

13. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med. 2019 May 16;9(3):440–7.

14. Sadasivan C, Cruz C, Dolgoy N, Hyde A, Campbell S, McNeely M, et al. Examining Patient Engagement in Chatbot Development Approaches for Healthy Lifestyle and Mental Wellness Interventions: Scoping Review. J Particip Med. 2023 May 22;15:e45772.

15. Olano-Espinosa E, Avila-Tomas JF, Minue-Lorenzo C, Matilla-Pardo B, Serrano Serrano ME, Martinez-Suberviola FJ, et al. Effectiveness of a Conversational Chatbot (Dejal@bot) for the Adult Population to Quit Smoking: Pragmatic, Multicenter, Controlled, Randomized Clinical Trial in Primary Care. JMIR Mhealth Uhealth. 2022 Jun 27;10(6):e34273.

16. Nassar CM, Dunlea R, Montero A, Tweedt A, Magee MF. Feasibility and Preliminary Behavioral and Clinical Efficacy of a Diabetes Education Chatbot Pilot Among Adults With Type 2 Diabetes. J Diabetes Sci Technol. 2023 Jun 6;19322968231178020.

17. Nielsen JPS, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Oto-Laryngologica. 2023 Nov 1;143(9):779–82.

18. Eid K, Eid A, Wang D, Raiker RS, Chen S, Nguyen J. Optimizing Ophthalmology Patient Education via ChatBot-Generated Materials: Readability Analysis of AI-Generated Patient Education Materials and The American Society of Ophthalmic Plastic and Reconstructive Surgery Patient Brochures. Ophthalmic Plastic & Reconstructive Surgery. :10.1097/IOP.0000000000002549.

19. Kantor J. ChatGPT, large language models, and artificial intelligence in medicine and health care: A primer for clinicians and researchers. JAAD Int. 2023 Dec;13:168–9.

20. The potential impact of ChatGPT in clinical and translational medicine - Xue - 2023 - Clinical and Translational Medicine - Wiley Online Library [Internet]. [cited 2023 Dec 18]. Available from: https://onlinelibrary-wiley-com.proxy.library.uu.nl/doi/10.1002/ctm2.1216

21. Gabriel J, Shafik L, Alanbuki A, Larner T. The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. Int Urol Nephrol. 2023 Nov 1;55(11):2717–32.

22. Yin R, Neyens DM. Examining how information presentation methods and a chatbot impact the use and effectiveness of electronic health record patient portals: An exploratory study. Patient Education and Counseling. 2024 Feb 1;119:108055.

23. Campbell DJ, Estephan LE. ChatGPT for patient education: an evolving investigation. Journal of Clinical Sleep Medicine. 19(12):2135–6.

24. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. JAMA Netw Open. 2023 Aug 1;6(8):e2330320.

25. Parsa A, Ebrahimzadeh MH. ChatGPT in Medicine; a Disruptive Innovation or Just One Step Forward? Arch Bone Jt Surg. 2023;11(4):225–6.

26. Babayiğit O, Tastan Eroglu Z, Ozkan Sen D, Ucan Yarkac F. Potential Use of ChatGPT for Patient Information in Periodontology: A Descriptive Pilot Study. Cureus. 15(11):e48518.

27. Botelho F, Tshimula JM, Poenaru D. Leveraging ChatGPT to Democratize and Decolonize Global Surgery: Large Language Models for Small Healthcare Budgets. World J Surg. 2023 Nov 1;47(11):2626–7.

28. Jin JQ, Dobry AS. ChatGPT for healthcare providers and patients: Practical implications within dermatology. Journal of the American Academy of Dermatology. 2023 Oct 1;89(4):870–1.

29. Alessandri-Bonetti M, Liu HY, Giorgino R, Nguyen VT, Egro FM. The First Months of Life of ChatGPT and Its Impact in Healthcare: A Bibliometric Analysis of the Current Literature. Ann Biomed Eng [Internet]. 2023 Jul 23 [cited 2023 Dec 18]; Available from: https://doi.org/10.1007/s10439-023-03325-8

30. Lungren MP, Fishman EK, Chu LC, Rizk RC, Rowe SP. More is Different: Large Language Models in Healthcare. Journal of the American College of Radiology [Internet]. 2023 Dec 1 [cited 2023 Dec 18]; Available from: https://www.sciencedirect.com/science/article/pii/S1546144023009626

31. Nashwan AJ, Abujaber AA. Harnessing Large Language Models in Nursing Care Planning: Opportunities, Challenges, and Ethical Considerations. Cureus. 2023 Jun;15(6):e40542.

32. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2023 Apr 1;17(4):102744.

33. Demner-Fushman D, Elhadad N. Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. Yearb Med Inform. 2016 Nov 10;(1):224–33.

34. Wu H, Wang M, Wu J, Francis F, Chang YH, Shavick A, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. npj Digit Med. 2022 Dec 21;5(1):1–15.

35. Sharma M, Sharma S. A holistic approach to remote patient monitoring, fueled by ChatGPT and Metaverse technology: The future of nursing education. Nurse Educ Today. 2023 Dec;131:105972.

36. Schueller SM, Morris RR. Clinical science and practice in the age of large language models and generative artificial intelligence. J Consult Clin Psychol. 2023 Oct;91(10):559–61.

37. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq. 2023 Feb 28;rs.3.rs-2566942.

38. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr. 2023 Mar 1;7(2):pkad015.

39. Gupta S, Gupta S, McMath K, Sugandh S. Enhancing Complex Wound Care by Leveraging Artificial Intelligence: An Artificial Intelligence Chatbot Software Study. Wounds. 2023;35(8):e265–7.

40. Wang J, Yang J, Zhang H, Lu H, Skreta M, Husić M, et al. PhenoPad: Building AI enabled note-taking interfaces for patient encounters. npj Digit Med. 2022 Jan 27;5(1):1–9.

41. Yetisgen M, Vanderwende L. Automatic Identification of Substance Abuse from Social History in Clinical Text. In: Ten Teije A, Popow C, Holmes JH, Sacchi L, editors. Cham: Springer International Publishing; 2017 [cited 2024 Jan 15]. p. 171–81. (Lecture Notes in Computer Science; vol. 10259). Available from: http://link.springer.com/10.1007/978-3-319-59758-4_18

42. Ni Y, Bachtel A, Nause K, Beal S. Automated detection of substance use information from electronic health records for a pediatric population. J Am Med Inform Assoc. 2021 Aug 1;28(10):2116–27.

43. Matsuda S, Ohtomo T, Tomizawa S, Miyano Y, Mogi M, Kuriki H, et al. Incorporating Unstructured Patient Narratives and Health Insurance Claims Data in Pharmacovigilance: Natural Language Processing Analysis of Patient-Generated Texts About Systemic Lupus Erythematosus. JMIR Public Health Surveill. 2021 Jun 29;7(6):e29238.

44. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010;17(1):19–24.

45. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930–40.

46. Ahmed MS, Kornblum D, Oliver D, Fusar-Poli P, Patel R. Associations of remote mental healthcare with clinical outcomes: a natural language processing enriched electronic health record data study protocol. BMJ Open. 2023 Feb 10;13(2):e067254.

47. Au Yeung J, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, et al. AI chatbots not yet ready for clinical use. Front Digit Health. 2023 Apr 12;5:1161098.

48. Beaulieu-Jones BK, Villamar MF, Scordis P, Bartmann AP, Ali W, Wissel BD, et al. Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study. The Lancet Digital Health. 2023 Dec 1;5(12):e882–94.

49. Kim S, Lee CK, Choi Y, Baek ES, Choi JE, Lim JS, et al. Deep-Learning-Based Natural Language Processing of Serial Free-Text Radiological Reports for Predicting Rectal Cancer Patient Survival. Front Oncol. 2021;11:747250.

50. Brown JR, Ricket IM, Reeves RM, Shah RU, Goodrich CA, Gobbel G, et al. Information Extraction From Electronic Health Records to Predict Readmission Following Acute Myocardial Infarction: Does Natural Language Processing Using Clinical Notes Improve Prediction of Readmission? J Am Heart Assoc. 2022 Apr 5;11(7):e024198.

51. Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. Journal of the American Medical Informatics Association. 2013 Sep 1;20(5):828–35.

52. Zhang Q, Liang Y. Comments on "ChatGPT and its Role in the Decision-Making for the Diagnosis and Treatment of Lumbar Spinal Stenosis: A Comparative Analysis and Narrative Review." Global Spine Journal. 2023 Dec 13;21925682231222268.

53. Hassanpour S, Bay G, Langlotz CP. Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing. J Digit Imaging. 2017 Jun;30(3):314–22.

54. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making [Internet]. medRxiv; 2023 [cited 2023 Dec 12]. p. 2023.02.02.23285399. Available from: https://www.medrxiv.org/content/10.1101/2023.02.02.23285399v1

55. Sarioglu E, Choi HA, Yadav K. Clinical Report Classification Using Natural Language Processing and Topic Modeling. In: 2012 11th International Conference on Machine Learning and Applications [Internet]. 2012 [cited 2023 Dec 19]. p. 204–9. Available from: https://ieeexplore.ieee.org/document/6406751

56. Aramaki E, Yano K, Wakamiya S. MedEx/J: A One-Scan Simple and Fast NLP Tool for Japanese Clinical Texts. Stud Health Technol Inform. 2017;245:285–8.

57. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. Journal of Medical Internet Research. 2023 Aug 22;25(1):e48659.

58. Salihu A, Gadiri MA, Skalidis I, Meier D, Auberson D, Fournier A, et al. Towards AI-assisted cardiology: a reflection on the performance and limitations of using large language models in clinical decision-making. EuroIntervention. 2023 Dec 4;19(10):e798–801.

59. Seth I, Xie Y, Hunter-Smith DJ, Seifman MA, Rozen WM. Re: Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance. ANZ J Surg. 2023 Oct 9;

60. Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: Harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol. 2023 Sep 20;13(4):170–8.

61. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearb Med Inform. 2015 Aug 13;10(1):183–93.

62. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc. 2018 Mar 1;25(3):331–6.

63. Trivedi G, Pham P, Chapman WW, Hwa R, Wiebe J, Hochheiser H. NLPReViz: an interactive tool for natural language processing on clinical text. J Am Med Inform Assoc. 2017 Jul 22;25(1):81–7.

64. Kersloot MG, van Putten FJP, Abu-Hanna A, Cornet R, Arts DL. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. Journal of Biomedical Semantics. 2020 Nov 16;11(1):14.

65. Hasikin K, Lai KW, Satapathy SC, Sabanci K, Aslan MF. Editorial: Emerging applications of text analytics and natural language processing in healthcare. Frontiers in Digital Health [Internet]. 2023 [cited 2024 Jan 7];5. Available from: https://www.frontiersin.org/articles/10.3389/fdgth.2023.1227948

66. Loscertales J, Abrisqueta-Costa P, Gutierrez A, Hernández-Rivas JÁ, Andreu-Lapiedra R, Mora A, et al. Real-World Evidence on the Clinical Characteristics and Management of Patients with Chronic Lymphocytic Leukemia in Spain Using Natural Language Processing: The SRealCLL Study. Cancers. 2023 Jan;15(16):4047.

67. Ortiz A, Portoles J, Pino-Pino MD, Barea J, López M, de Sequera P, et al. Clinical Characteristics and Management of Patients with Secondary Hyperparathyroidism Undergoing Hemodialysis: A Feasibility Analysis of Electronic Health Records Using Natural Language Processing. Kidney Diseases. 2023 Jan 23;9(3):187–96.

68. Pethani F, Dunn AG. Natural language processing for clinical notes in dentistry: A systematic review. Journal of Biomedical Informatics. 2023 Feb 1;138:104282.

69. Naseri H, Kafi K, Skamene S, Tolba M, Faye MD, Ramia P, et al. Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. J Biomed Inform. 2021 Aug;120:103864.

70. Hussain SA, Sezgin E, Krivchenia K, Luna J, Rust S, Huang Y. A natural language processing pipeline to synthesize patient-generated notes toward improving remote care and chronic disease management: a cystic fibrosis case study. JAMIA Open. 2021 Sep 29;4(3):ooab084.

71. Chae S, Song J, Ojo M, Topaz M. Identifying Heart Failure Symptoms and Poor Self-Management in Home Healthcare: A Natural Language Processing Study. Stud Health Technol Inform. 2021 Dec 15;284:15–9.

72. Le TD, Noumeir R, Rambaud J, Sans G, Jouvet P. Detecting of a Patient's Condition From Clinical Narratives Using Natural Language Representation. IEEE Open J Eng Med Biol. 2022;3:142–9.

73. Sivarajkumar S, Wang Y. HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing [Internet]. arXiv; 2022 [cited 2023 Dec 18]. Available from: http://arxiv.org/abs/2203.05061

74. Delsoz M, Raja H, Madadi Y, Tang AA, Wirostko BM, Kahook MY, et al. The Use of ChatGPT to Assist in Diagnosing Glaucoma Based on Clinical Case Reports. Ophthalmol Ther. 2023 Dec 1;12(6):3121–32.

75. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp. 1997;829–33.

76. Patel TA, Puppala M, Ogunti RO, Ensor JE, He T, Shewale JB, et al. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. Cancer. 2017 Jan 1;123(1):114–21.

77. Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Stud Health Technol Inform. 2004;107(Pt 1):487–91.

78. Zhao B. Clinical Data Extraction and Normalization of Cyrillic Electronic Health Records Via Deep-Learning Natural Language Processing. JCO Clin Cancer Inform. 2019 Sep;3:1–9.

79. Lee RY, Kross EK, Torrence J, Li KS, Sibley J, Cohen T, et al. Assessment of Natural Language Processing of Electronic Health Records to Measure Goals-of-Care Discussions as a Clinical Trial Outcome. JAMA Netw Open. 2023 Mar 1;6(3):e231204.

80. Lindvall C, Deng CY, Moseley E, Agaronnik N, El-Jawahri A, Paasche-Orlow MK, et al. Natural Language Processing to Identify Advance Care Planning Documentation in a Multisite Pragmatic Clinical Trial. J Pain Symptom Manage. 2022 Jan;63(1):e29–36.

81. Ross L, Danforth CM, Eppstein MJ, Clarfeld LA, Durieux BN, Gramling CJ, et al. Story Arcs in Serious Illness: Natural Language Processing features of Palliative Care Conversations. Patient Educ Couns. 2020 Apr;103(4):826–32.

82. Brizzi K, Zupanc SN, Udelsman BV, Tulsky JA, Wright AA, Poort H, et al. Natural Language Processing to Assess Palliative Care and End-of-Life Process Measures in Patients With Breast Cancer With Leptomeningeal Disease. Am J Hosp Palliat Care. 2020 May;37(5):371–6.

83. Kaskovich S, Wyatt KD, Oliwa T, Graglia L, Furner B, Lee J, et al. Automated Matching of Patients to Clinical Trials: A Patient-Centric Natural Language Processing Approach for Pediatric Leukemia. JCO Clin Cancer Inform. 2023 Sep;(7):e2300009.

84. Li X, Liu H, Kury F, Yuan C, Butler A, Sun Y, et al. A Comparison between Human and NLP-based Annotation of Clinical Trial Eligibility Criteria Text Using The OMOP Common Data Model. AMIA Jt Summits Transl Sci Proc. 2021;2021:394–403.

85. Li J, Wei Q, Ghiasvand O, Chen M, Lobanov V, Weng C, et al. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. BMC Medical Informatics and Decision Making. 2022 Sep 6;22(3):235.

86. Alekhya B, Sasikumar R. An ensemble approach for healthcare application and diagnosis using natural language processing. Cogn Neurodyn. 2022 Oct 1;16(5):1203–20.

87. Liu J, Capurro D, Nguyen A, Verspoor K. “Note Bloat” impacts deep learning-based NLP models for clinical prediction tasks. J Biomed Inform. 2022 Sep;133:104149.

88. Villena F, Dunstan J. [Automatic keyword retrieval from clinical texts: an application of natural language processing to massive data of Chilean suspected diagnosis]. Rev Med Chil. 2019 Oct;147(10):1229–38.

89. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study [Internet]. [cited 2023 Nov 30]. Available from: https://pubs.rsna.org/doi/epdf/10.1148/radiol.230725

90. Evaluating the State-of-the-Art in Automatic De-identification | Journal of the American Medical Informatics Association | Oxford Academic [Internet]. [cited 2024 Jan 6]. Available from: https://academic-oup-com.proxy.library.uu.nl/jamia/article/14/5/550/720189

91. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of Biomedical Informatics. 2015 Dec 1;58:S20–9.

92. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of Patient Notes with Recurrent Neural Networks [Internet]. arXiv; 2016 [cited 2024 Jan 15]. Available from: http://arxiv.org/abs/1606.03475

93. Guirguis CA, Crossley JR, Malekzadeh S. Bilateral Vocal Fold Paralysis in a Patient With Neurosarcoidosis: A ChatGPT-Driven Case Report Describing an Unusual Presentation. Cureus. 2023 Apr;15(4):e37368.

94. Milan-Ortiz V, Damughatla AR, Qazi AM, Kamatham S, Oli S, Koleti P, et al. Neutropenic Enterocolitis Following Autologous Stem Cell Transplantation: A Compelling Clinical Case Report Written With the Assistance of ChatGPT. Cureus. 15(3):e36390.

95. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health. 2023 Apr;5(4):e179–81.

96. de Pennington N, Mole G, Lim E, Milne-Ives M, Normando E, Xue K, et al. Safety and Acceptability of a Natural Language Artificial Intelligence Assistant to Deliver Clinical

Follow-up to Cataract Surgery Patients: Proposal. JMIR Res Protoc. 2021 Jul 28;10(7):e27227.

97. Were MC, Gorbachev S, Cadwallader J, Kesterson J, Li X, Overhage JM, et al. Natural language processing to extract follow-up provider information from hospital discharge summaries. AMIA Annu Symp Proc. 2010;2010:872–6.

98. Haynes D, Lin SK, Deitermann AM, Etzkorn JR, Miller CJ, Higgins HW. Natural Language Processing of Mohs Micrographic Surgery Patient Experience Surveys: A Single Center Review. J Cutan Med Surg. 2023 Nov;27(6):660–2.

99. Tabaie A, Sengupta S, Pruitt ZM, Fong A. A natural language processing approach to categorise contributing factors from patient safety event reports. BMJ Health Care Inform. 2023 May 1;30(1):e100731.

100. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. Clinical and Translational Medicine. 2023;13(3):e1216.

Supplementary Materials

Supplementary Table 1. The search query

Database	***Date***	***Search string***
Pubmed	01-01-2010 to 14-12-2023	("large language model"[ti] OR "LLM"[ti] OR "Natural-Language"[ti] OR "language models"[ti] OR "AI language models"[ti] OR "chatbot"[ti] OR "natural language inference"[ti] OR "natural language processing"[ti] OR "NLP"[ti] OR "ChatGPT"[ti] OR "BERT"[ti] OR "LLaMA"[ti] OR ("AI"[ti] AND "language model"[tiab])) AND ("healthcare"[ti] OR "health care"[ti] OR "clinical"[ti] OR "clinical workflow"[ti] OR "clinic"[ti] OR "medicine"[ti] OR "hospital"[ti] OR "clinics"[ti] OR "care"[ti] OR "patient"[ti] OR "treatment"[ti])

Supplementary Table 2. Data extraction sheet format

Domain	**Extracted item**	**Additional explanation**
Paper information	First author	*Surname is enough*
	Title	
	Extracted by	
	Year of publication	
	Type of publication	
	Depth/particularity level of clinical application	*Level 1: Generic comments about the LLM task* *Level 2: LLM task described in a specific medical specialty and/or scenario* *Level 3: LLM task described in a specific medical specialty and/or scenario. Also the purpose and/or relevance of the clinical task is described.*
Application information	Description of application	*As described in paper*
	User orientation	*- Patients/relatives* *- Healthcare professionals* *- Clinical centers* *- Lawyers/regulators*
	User-specialization	*E.g. radiologists*
	Overarching clinical task	*- Patient communication* *- Patient education* *- Clinical workflow* *- Healthcare management*
	Task description	*Description in own words*

	Type of application	
	Purpose/relevance of application	
	Type of classical NLP task	- *Text generation (TG)* - *Summarization (Summ)* - *Named Entity Recogntion (NER)* - *Information Extraction (IE)* - *Question Answering (QA)* - *Speech Recognition (SR)*
	Type of input data	
	Example of input data	
	Type of output data	
Usage of application	Implemented in clinic?	
	Number of scenarios tested	