Scholars Journal of Applied Medical Sciences
Abbreviated Key Title: Sch J App Med Sci
ISSN 2347-954X (Print) | ISSN 2320-6691 (Online)
Journal homepage: https://saspublishers.com

OPEN ACCESS

Medical Sciences

From Notes to Billing: Large Language Models in Revolutionizing Medical Documentation and Healthcare Administration

Shashank Agarwal[1*], Sumeer Basha Peta[2]

[1]Independent Researcher, Boston, Massachusetts, USA

DOI: https://doi.org/10.36347/sjams.2025.v13i08.005 | **Received:** 05.06.2025 | **Accepted:** 11.08.2025 | **Published:** 14.08.2025

***Corresponding author:** Shashank Agarwal
Independent Researcher, Boston, Massachusetts, USA

Abstract **Review Article**

The growing reliance of medical professionals for quick and accurate documentation, invoicing, and administrative processing has fueled interest in AI-powered automation tools. Large Language Models (LLMs) such as GPT-4, Med-PaLM 2, BioGPT, and ClinicalBERT have the potential to significantly improve medical documentation, clinical workflows, billing accuracy, and administrative load. This paper examines the revolutionary role of LLMs in clinical documentation and administrative automation, including current advances, real-world implementations, and new solutions. The article emphasizes their ability to generate clinical notes, code medical operations, facilitate patient communication, and navigate intricate insurance systems. It also addresses important issues such as model bias, hallucination risks, data privacy, legal accountability, and workforce preparation. This review seeks to offer an in depth yet comprehensible overview for medical practitioners, informaticians, and policymakers intending to securely incorporate LLMs into their existing structures by assessing important deployments and comparing common LLM platforms. While LLMs hold promise in increasing efficiency in medical facilities, their safe implementation will require human supervision, strong ethical standards, and constant system validation.

Keywords: Large Language Models, medical documentation, healthcare administration, text summarization, EHRs, clinical NLP.

I. INTRODUCTION

The global healthcare sector has been facing mounting pressure in the past few decades to manage the expanding amount of health data, rising operating expenses, and a shortage of clinicians while improving quality, safety, and efficiency. Artificial Intelligence (AI) tools, particularly large language models (LLMs), have emerged as a promising innovation in administrative automation and documentation enhancement, especially as a result of the accelerated health system digitization due to the COVID-19 pandemic [1] [2]. Clinical workflows are now actively incorporating LLMs like Microsoft's BioGPT, Google's Med-PaLM 2, and OpenAI's GPT-4, which can assist in creating discharge summaries, deciphering patient-provider conversations, extracting structured data from unstructured notes, and even help with automated billing and coding [3] [4]. These models are appropriate for both front-end clinical documentation and back-end administrative tasks because of their ability to comprehend clinical language, produce patient histories, and generate human-like text in response to real-time data.

LLMs have the potential to improve several facets of healthcare and medicine. Administrative work can be automated, education can be enhanced and customized, decision assistance tools can be enabled, etc [5] [6]. Moreover, integrating Large Language Models into the medical documentation process could save a substantial amount of the time that doctors presently spend on this tedious task. When given a clear and comprehensive prompt (user-provided text input), LLMs are excellent at constructing documentation with an appropriate format and adding pertinent patient information into the documents [7]. Clinicians' administrative workload might be greatly reduced by LLMs automating parts of this procedure, which could increase productivity and lower burnout.

This review paper aims to investigate the current and emerging applications of LLMs in the arenas of administrative automation and medical documentation. In particular, it examines how LLMs accelerate documentation procedures, aid in clinical decision-making, automate medical coding and billing processes, and improve workflows related to insurance, thereby promoting patient communication. Along with examining the top LLM platforms being used in healthcare, it assesses their reliability as well as limitations and discusses the risks and issues related to LLM implementation, such as workforce transformation, safety, and fairness. Although previous research has examined specific uses of LLMs in healthcare, like note summarization or clinical chatbots, there is a lack of comprehensive reviews that concentrate on their application in administrative automation and medical recordkeeping. This review addresses that gap by exploring how LLMs are changing documentation workflows, billing, coding, and insurance procedures—areas that are still underrepresented in the literature today despite having a significant impact on clinician workload and healthcare efficiency.

II. ROLE OF LLMS IN MEDICAL DOCUMENTATION AND ADMINISTRATIVE AUTOMATION

II.1. Customizing patient appointments and Scheduling

People may encounter unforeseen challenges when attempting to manage care for personally or for their loved ones, such as imaging tests or medical procedures that need to be prescreened for safety and appropriateness. But with the assistance of LLMs, patients, along with caregivers, may acquire tailored information on these procedures according to their medical histories. LLMs can analyze a patient's medical information to assist patients in comprehending how to prepare for a diagnostic exam or procedure, answer concerns regarding what to expect, and proactively guide them in scheduling their treatment at a suitable site. LLMs can analyze a patient's medical information [8-10]. LLMs may educate patients about any possible risks and provide alternative treatments. For example, if a patient has a pacemaker or defibrillator and requires a magnetic resonance imaging (MRI) assessment, and has previously experienced an allergic response to a specific intravenous contrast agent. Through the processing of discrete EMR data, interaction notes from specialist appointments, and pertinent scanned documents, LLMs are able to notify primary care physicians when patients are past due for follow-up testing [10-12]. By obtaining the data required for ensuring patient safety and providing patients with pre-appointment counseling, LLMs can also expedite the prescreening procedure.

II.2. Improving clinical documentation

LLMs have the potential to assist with clinical documentation and record-keeping [13,14]. OpenNotes, a national initiative to share clinicians' notes with their patients, has been shown to improve record accuracy by giving patients access and edit capabilities [15]. Similarly, LLMs can help manage medical records by flagging potential contradictions or discrepancies, providing smart and dynamic clinical decision support, alerting clinicians to incomplete follow-up recommendations, or flagging actionable test results. This can help ensure that patient data is accurate and up to date, improving the overall quality of care delivered and reducing the likelihood of documentation errors. Furthermore, LLMs can help clinicians document components necessary for appropriate and accurate billing, which can contribute to better patient care and increase revenue [16]. Nevertheless, allowing LLMs to alter patient charts or files, which are considered legal files, poses difficulties. Despite their sophistication, LLMs are not immune to errors. If permitted to run autonomously, they have the potential to introduce unforeseen and unnoticed inaccuracies into patients' records, leading to erroneous medical choices. The privacy of patients' medical records is also jeopardized, particularly if LLMs have access to or can edit critical information. As a result, careful evaluation and execution of proper measures would be required to make sure that LLMs are used safely and accurately in handling patient medical information.

II.3. Facilitating insurance prior authorization

The initial approval procedure is very burdensome and irritating for professionals in the USA [17]. LLMs may aid physicians by assembling the information from the patient's record required to submit a comprehensive and detailed application for prior approval for a specific therapy or procedure [5]. In a similar vein, insurance organizations might leverage LLMs for automating the review of submitted documents and point out elements that contribute to their acceptance or refusal decisions, reducing the necessity for laborious and error-prone human assessment and enhancing the general veracity of the answer [18]. LLMs may analyze healthcare records, insurance plans, and other pertinent data to decide if an individual's health insurance policy supports the treatment or a service. Large Language Models might also be utilized to redesign the peer-to-peer review procedure of insurance by enabling professionals on both ends of the consultation to draw on previous rulings in relevant instances to make better informed and accurate decisions. It may minimize physician, insurance provider, and administrative workloads while also reducing errors and holdups that directly impact the ability of patients to obtain the medical assistance they require [18].

II.4. Improving Medical Coding and Billing Accuracy Using LLMs

Medical coding and billing serve as vital yet labor-intensive aspects of healthcare administration. These mechanisms make sure that medical practitioners are properly reimbursed and that insurance companies have reliable data to substantiate payments. Historically, qualified medical coders manually reviewed clinical evidence for assigning appropriate codes using the "International Classification of Diseases" (ICD), "Current Procedural Terminology" (CPT), and "Healthcare Common Procedure Coding System". Unfortunately, this manual method is susceptible to human mistakes, discrepancies, and delays, which contribute to increased administrative hassles and revenue losses [19,20]. LLMs, which can interpret and synthesize human-like language, are rapidly being investigated as a way to completely transform this sector by automating and improving the precision of medical coding and billing [21].

LLMs such as GPT-4 and Med-PaLM 2 have demonstrated remarkable outcomes for natural language comprehension, allowing them to interpret intricate clinical descriptions and assign them to relevant billing codes. Finlayson *et al.,*[22] demonstrated that LLMs can outperform rule-based systems in identifying coding-relevant segments in electronic medical records (EMRs) when trained using clinical documentation datasets. According to the study [22], LLMs can generalize to new clinical scenarios with excellent precision when using zero-shot and few-shot prompting, especially when giving CPT procedure codes and ICD-10 diagnostic codes. A large language model, for example, can correctly provide the correct ICD-10 code (E11.329) based on an evaluation report that reads, "The patient was diagnosed with type 2 diabetes with mild nonproliferative diabetic retinopathy without macular edema." Instead of requiring a developer to consult numerous code libraries and documentation avenues, this task may now be completed in almost real-time with more consistency.

Frequently, mismatched coding or documentation problems lead to claim denials, which may have a major impact on hospital earnings. A study by the American Medical Association [23] found that between 15 and 20 percent of all claims are rejected on the initial attempt, most frequently because of incorrect coding. For accuracy and payer policy compliance, medical facilities can use LLMs to pre-validate and verify codes with documentation. When utilized in conjunction with an AI-augmented coding assistance framework, LLMs can cut claim refusal rates by 35% [24]. These algorithms identify documentation errors, offer more correct codes, and even offer justification syntax that may be applied to claims to ensure audit compliance [24]. These features not only improve procedures, but they also secure providers' financial viability.

Integrating LLMs into real-time medical documentation settings represents one of their most promising uses. As physicians enter their notes, LLMs integrated into EMRs may provide real-time code recommendations, enabling them to record important billing-relevant details (e.g., complexities, time spent, and treatments). This method ensures that documentation fulfills billing and health care standards at the same time. An LLM-augmented mechanism improved documentation integrity by 22% and minimized after-hours coding requests by 41% [25]. The combined benefit of clinical assistance and coding efficiency signals a transition toward a more interactive and smarter billing process.

III. Addressing Bias, Safety, and Legal Concerns in LLM-Based Healthcare Automation

Incorporation of Large language models (LLMs) into health care records and medical administration offers many advantages. However, it also raises serious issues with safety, bias, and legal accountability. Although LLMs like GPT-4, Med-PaLM, and BioGPT show great strengths in clinical language comprehension, note summarization, and documentation automation, their implementation in actual medical facilities carries risks that need to be meticulously assessed to ensure secure, fair, and legal use.

The data sets utilized in training are the source of bias in Large Language Models. MLMs or Medical language models frequently undergo training using a variety of text sources, such as clinical notes, research papers, and publicly available health-related data. Nonetheless, such sources may tend to produce limitations in care and documentation in the field of medicine, which can drastically yield biased results that may worsen inequality. As an example, Obermeyer *et al.,*[26] found that medical cost data-trained models underestimated the health needs of Black patients due to structural differences in care access. Similarly, an LLM that is trained on the system of bias records can generate results that disregard or distort the symptoms in patients with rare diseases, minorities, as well as in females [27]. It can lead to practical repercussions in computerized documentations, diagnosis recommendations, and even appeals for prior approval, which can perpetuate systemic bias in medical provision.

One of the major safety concerns with LLMs is “hallucination”, i.e., the tendency of the models to generate compelling yet incorrect or otherwise misleading information. Even minor mistakes in the healthcare documentation may cause inaccurate diagnosis of a disease, misleading treatment, inaccurate billing, etc. Gilbert *et al.,*[28] stated that GPT-4 produced fake information that did not exist in the original record of patients, such as incorrect dosage of prescriptions or diagnoses, in 12% of clinical summary generation events. Despite their increasing factual consistency, LLMs are based on probability and therefore do not "know" facts as humans do. Therefore, when their responses impact clinical or legal records, precautions must be taken [28]. To address this, researchers are looking into hybrid systems that integrate Large Language Models with retrieval-augmented generation (RAG), in which the algorithm compares its output responses to organized databases or EMRs [29,30]. Despite this improvement, human monitoring is still required for crucial documentation responsibilities.

The employment of LLMs in creating or editing medical documentation produces major legal concerns. Clinical records are legal documents that are frequently utilized in court, insurance conflicts, and regulatory inspections. Liability can be difficult to determine if an

LLM-generated note results in an incorrect diagnosis or insurance refusal. Would the fault lie with the company, organization, or model developer? There is currently no agreement on how to account for AI in therapeutic contexts. According to a commentary by Goodman and Flanagin [31] in JAMA, unless unambiguous legal precedents have been established, medical professionals should remain the ultimate judges of documentation while avoiding excessive dependence on LLMs in unattended modalities. The FDA and other regulatory organizations have started to outline guidelines for AI in healthcare equipment, but they have not yet addressed documentation and administrative AI use in depth. As a result, hospitals that employ LLMs should develop explicit procedures for model supervision, audit trails, and clinician verification. Several systems now require that any text written or altered by an LLM be specifically marked, allowing users to verify and validate its veracity before it is included in official documentation.

An additional significant security risk is the privacy of patient data. LLMs used in medical facilities should conform to data privacy standards like the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the GDPR in Europe. However, connecting LLMs with EMRs increases the danger of data leakage, particularly when leveraging APIs or cloud-based frameworks that are not expressly developed for safe healthcare settings. The Department of Health and Human Services (HHS) issued advice recommending against using publicly hosted LLMs (such as ChatGPT) to process protected health information (PHI) unless specific Business Associate Agreements (BAAs) have been put in place [32]. This highlights the importance of building on-premises or HIPAA-compliant LLM implementations, such as Microsoft Azure OpenAI or AWS HealthLake-compatible models [32]. Additionally, there are certain challenges with de-identification and long-term compliance for LLMs that store user inputs or conversational history for model improvement. To ensure secrecy, technical measures such as access limits, logging limitations, and automated redaction are essential.

Ⅳ. SYSTEM ARCHITECTURE OF LLMS IN CLINICAL DATA SUMMARIZATION

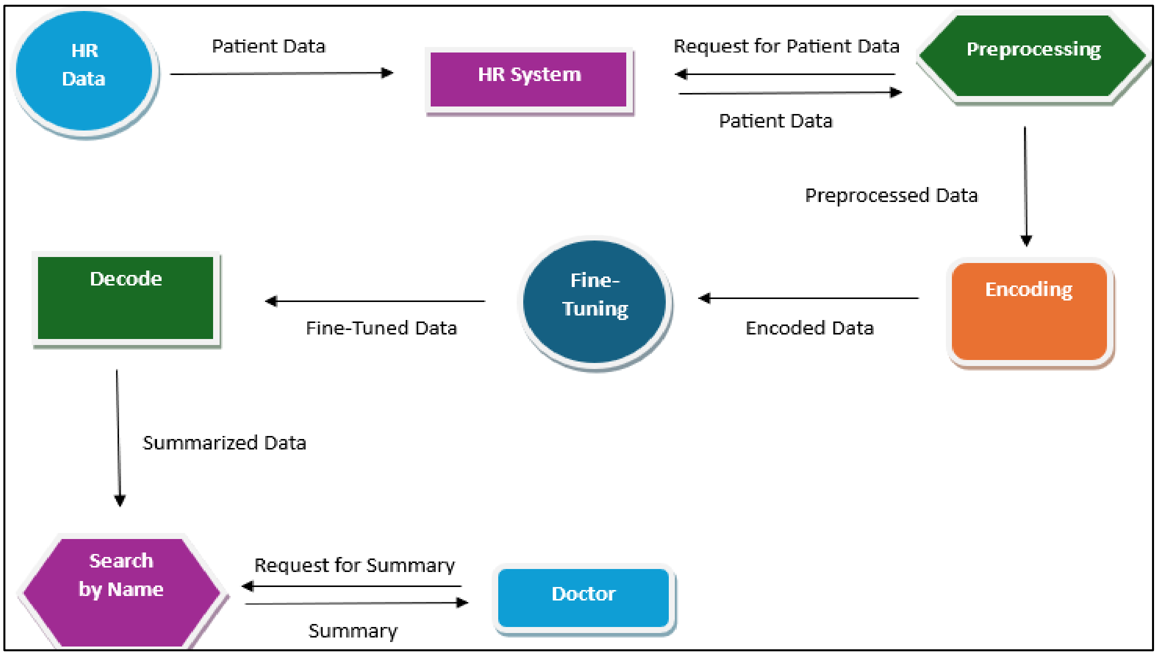

Fig. 1. System Design Process of LLMs in Clinical Data Summarization

With the use of an LLM, the architecture facilitates the effective summarization of medical records, giving medical personnel rapid access to important data. Each component and layer in this system is interrelated and has a specific purpose (Fig.1).

1. User Interaction Layer Doctor:

The end user of this summarization system is the doctor. The physician may inquire of the system for a patient's condensed medical records via a user interface. By eliminating the need to manually go through extensive healthcare data, this layer helps physicians make informed decisions more quickly. Following processing, the algorithm provides the physician a condensed, summarized response [33,34].

2. Core Summarization System:

This key component controls the workflow for summarizing the information. It handles doctor requests, retrieves the required information from the database, and then processes the information via several stages to produce a summary. This technology connects the medical professional and the intricate flow of data processing. The Healthcare System serves as a fundamental data source, containing detailed, raw patient medical records. Once the Summarization System obtains a request, it consults the healthcare records system to obtain the required information, such as medical histories, medical diagnoses, and therapeutic records [33,35].

3. Pipeline for Processing Data:

Preprocessing: The original health record information obtained from the system is first preprocessed. This involves the elimination of noise or any unnecessary data, which is done by cleaning and organizing the data so as to uphold congruence. Some of the preprocessing can include eliminating redundant information, standardizing data formats, fixing missing information, all of which preconditions the data to get efficiently encoded [36].

Encoding: Since data has already undergone preprocessing, it is now coded in an electronic format that can be comprehended by the LLM. This step transforms textual and structural data into vectorized data, or embedded (readable by the artificial intelligence models). The transformation allows the model to understand and interpret complicated medical wording and the patient data [36].

Fine-Tuning: Using the encoded data, the model is fine-tuned following encoding. In the fine-tuning step, the LLM is tailored to manage the nuances of medical information, including common trends in patient records, medical terminology, and abbreviations. Fine-tuning implies that the model's output is accurate and suitable for summarization in the field of healthcare [33,37].

Decoding: After fine-tuning, the algorithm decodes the data into a summary in natural language. The process of decoding converts the encoded health information back into legible and readable text, creating a concise summary that includes the most important details from the patient's medical records [33,38].

4. Summarized Data Store and Response Delivery:

Once the summary of data is generated, it is temporarily kept in the Summarized Data Store for easy access by the physician. It serves as the storage layer, ensuring that the summary data is easily accessible for future use or referencing, avoiding the need for further reprocessing of the raw information. As a response, the doctor is sent the summarized data. This output layer allows the physician to immediately retrieve condensed clinical records without having to read extensive or sophisticated reports, conserving time and allowing for speedy, informed clinical decision-making [33,39].

Ⅴ. WIDELY USED LLMS IN MEDICAL DOCUMENTATION AND ADMINISTRATIVE AUTOMATION

LLMs are rapidly changing the face of clinical documentation and the administration of healthcare. Their propensity to recognize, create, and contextualize medical terms and language has led to broad use in activities that include clinical note summarizing, EHR integration, and patient communication. Various LLMs, both general-purpose and domain-specific, are gaining prominence in real-world clinical settings owing to their superior efficiency, scalability, and adaptability.

GPT-4, created by OpenAI, has become one of the most well-known and extensively used models in healthcare domains [40]. Microsoft's Azure OpenAI platform, offering healthcare organizations HIPAA-compliant solutions, is one of the many applications that employ GPT-4. To help physicians with notetaking and patient message responses, GPT-4 has been integrated into EHR procedures in partnership with Epic Systems. In an ambulatory setting, GPT-4 integration led to major time savings and increased physician satisfaction with documentation activities [40]. The significance of human assessment of AI-generated content to maintain authenticity and security was also underlined in the same study [40].

Med-PaLM 2 is another well-known LLM developed especially for the medical sector by Google Research [3]. By incorporating physician-in-the-loop training, safety enhancements, and medical question-answering benchmarks, Med-PaLM 2 builds upon its predecessor. On the MultiMedQA benchmark, which utilizes datasets such as the USMLE questions and PubMedQA, it achieved expert-level performance. According to Singhal *et al.,*[3], Med-PaLM 2 is a good option for its applications in clinical record keeping and triage assistance since it can answer 85% of medical exam-style questions with expert-level coherence and factuality [3].

BioGPT has evolved as a specialized generative framework for biomedical text mining and generation in the field of biomedical literature processing [4]. BioGPT, designed by Microsoft Research, has demonstrated exceptional performance in biomedical question answering as well as summarization operations after being trained on millions of PubMed abstracts. For certain biological NLP tasks, especially those that focus on domain-specific vocabulary interpretation, BioGPT performed better than conventional BERT-based models [4]. Although not as commonly utilized in direct clinical note summarization, BioGPT is being used extensively in technologies that facilitate the creation of documentation from scientific and clinical trial data.

ClinicalBERT is a refined variant of BERT tailored for clinical narratives from the MIMIC-III dataset, serving as another highly customized model.

Performance in structured operations, including risk classification, adverse event identification, and drug extraction, has been shown by ClinicalBERT. According to Alsentzer *et al.,*[41] ClinicalBERT performed noticeably better than vanilla BERT on named entity recognition and sentence classification in medical texts when tested on a variety of clinical NLP tasks. Although not generative like GPT-4, ClinicalBERT facilitates intelligent documentation workflows by identifying important information in free-text records [41].

GatorTron is among the largest transformer-based clinical language models to date. It was created by NVIDIA and the University of Florida Health. Medical question answering and medical summarization are among the complex healthcare NLP tasks that GatorTron has been tested to do after being trained on more than 90 billion words of clinical content. Yang *et al.,* [42] found that GatorTron had the highest possible scores on five clinical NLP benchmark tasks and demonstrated potential in enhancing clinical concept retrieval and EHR queries. It is perfect for administrative automation and backend documentation assistance because of its size and domain- particular tuning [42].

The commercial adoption of GPT-4 for medical documentation is exemplified by Nuance's Dragon Ambient eXperience (DAX) and its LLM-powered expansion, DAX Copilot [43]. Discussions between clinicians and patients are passively recorded by these devices, which then transform them into structured notes that are integrated into electronic health records. In a multi-site assessment, Bundy *et al.,*[43] confirmed the utility of DAX as a front-end scribing solution driven by LLMs by finding that it retained documentation quality, improved clinician satisfaction, and decreased overtime documentation by upto 50%.

All these models- GPT-4, Med-PaLM 2, BioGPT, ClinicalBERT, GatorTron, and DAX Copilot—fill a distinct role in the medical recordkeeping ecosystem. While GPT-4 and Med-PaLM 2 are generic models with expanding clinical features, ClinicalBERT and GatorTron are tailored to systematic documentation and backend analysis. Their cohabitation illustrates the complexities of medical documentation, which necessitates both generative proficiency and field specificity. Subsequently, the eventual success of administrative automation depends on hybrid approaches that combine these models' characteristics while adhering to stringent ethical, legal, and clinical guidelines. TABLE Ⅰ outlines some of the leading LLMs that are used in clinical documentation and administrative tasks.

Table Ⅰ: Leading LLMS Used in Medical Documentation and Administration

LLM Name	Developer	Primary Applications	Notable Deployments
GPT-4	OpenAI / Microsoft	Clinical note drafting, patient messages, ambient documentation	Epic Systems, Nuance DAX Copilot
ClinicalBERT	MIT, Harvard	Risk stratification, entity extraction, structured tagging	NLP back-end tasks in hospital data systems
Med-PaLM 2	Google Research	Clinical QA, triage support, documentation synthesis	Google Health pilots
GatorTron	University of Florida & NVIDIA	Clinical NLP, question answering	Evaluated across 90B+ tokens in clinical datasets
BioGPT	Microsoft Research	Biomedical summarization, trial data extraction	Research pipelines, publications

TABLE Ⅱ outlines several studies in the literature that emphasize the importance and role of LLMs in healthcare administration.

Table Ⅱ: Studies Highlighting LLMS in Healthcare Administration

Study	Author/Year	Main Outcomes	Study Limitations
[22]	Huang *et al.,* 2024	ChatGPT-3.5 achieved 89–100% accuracy in pathology report extraction	Only tested on two pathology types
[44]	Lee *et al.,* 2025	LLMs applied to 500+ surgical pathology reports demonstrated high accuracy in structured data extraction.	Retrospective single-center study.
[45]	Van Veen *et al.,* 2023	81% of LLM summaries were equal or better than human-written notes	Focused only on 4 clinical summarization types
[46]	Hu *et al.,* 2024	ChatGPT performed competitively in extracting structured information from 847 CT reports via zero-shot prompting.	Lacked comparative evaluation with fine-tuned models.
[47]	Wei *et al.,* 2024	GPT-4 had high specificity and sensitivity for common symptoms	Less effective for rare symptoms

Study	Author/Year	Main Outcomes	Study Limitations
[48]	Pandey, H., & Amod, A. (2024).	A Multi-Agent LLM system achieved 86.2% accuracy for checklist-level and 95.6% overall prior authorization judgments.	Early-stage preprint; real-world deployment not yet validated.
[49]	Liu *et al.,* 2024	LLM outputs rated more readable than baselines	Depended heavily on automated metrics
[50]	Choi *et al.,* 2023	GPT-based prompts effectively extracted clinical variables from breast cancer and ultrasound reports.	Tested only on breast cancer cohort; generalizability unknown.
[51]	Zaretsky *et al.,* 2024	LLMs improved readability and understandability	Limited to 50 samples from one hospital
[52]	Zhang *et al.,* 2023	EHRTutor framework, using LLM-powered Q&A and summarization, outperformed baseline in patient comprehension and engagement.	Prototype tested in controlled setting, not yet in real-world clinical workflows.

Ⅵ. FUTURE RESEARCH DIRECTIONS

Future studies shall focus on the long-term effects of LLMs in real-world clinical situations rather than merely technical performance standards to fully reap the benefits of these models in the medical field. To assess how LLMs impact clinical outcomes, documentation quality, efficiency in workflow, and burnout among clinicians over time, comprehensive research is required. Beyond conventional NLP scores, defined assessment criteria are also necessary to evaluate the factual precision, patient safety, and usefulness of output generated by AI. Frameworks for bias audits should also be created to ensure equal treatment, especially for patients from different socioeconomic backgrounds, races, genders, and languages. As medical facilities shift to being data-driven, researchers should look into safe and adaptable implementation techniques, such as HIPAA-compliant cloud models and on-premises LLM programs that connect with EHRs through interoperable protocols such as HL7 FHIR. Furthermore, studies must inquire into how LLMs might be adapted for low-resource and multilingual settings with an intense administrative workload but restricted access to qualified professionals. Finally, the regulatory and ethical context for LLMs is still largely unknown; future research must give concrete paradigms for accountability, transparency, and informed consent when these models are used in medical and administrative decision-making.

キ. CONCLUSION

Administrative processes and medical documentation are being drastically transformed by Large Language Models. LLMs have an excellent opportunity to lessen clinician workload and enhance operational effectiveness by automating processes, including clinical note writing, coding, billing, and patient communication. In real-world scenarios, models like GPT-4, Med-PaLM 2, and BioGPT have demonstrated excellent performance. Nonetheless, issues with bias, hallucinations, confidentiality, and legal accountability highlight the necessity of cautious application. Secure adoption necessitates regulatory coherence, transparency, and human oversight. This review emphasizes how LLMs can be transformative tools when used meticulously, ethically, and with continual evaluation.

REFERENCES

1. E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” Nat. Med., vol. 25, no. 1, pp. 44–56, 2019. doi: 10.1038/s41591-018-0300-7.
2. F. Jiang *et al.,* “Artificial intelligence in healthcare: past, present and future,” Stroke Vasc. Neurol., vol. 2, no. 4, pp. 230–243, 2017. doi: 10.1136/svn-2017-000101.
3. K. Singhal *et al.,* “Large language models encode clinical knowledge,” Nature, vol. 620, no. 7972, pp. 172–180, 2023. doi: 10.1038/s41586-023-06291-2.
4. R. Luo *et al.,* “BioGPT: generative pre-trained transformer for biomedical text generation and mining,” Brief. Bioinform., vol. 23, no. 6, p. bbac409, 2022. doi: 10.1093/bib/bbac409.
5. R. Bhayana, S. Krishna, and R. R. Bleakney, “Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations,” Radiology, vol. 307, no. 5, p. e230582, 2023. doi: 10.1148/radiol.230582.
6. D. Sykes *et al.,* “Comparison of rule-based and neural network models for negation detection in radiology reports,” Nat. Lang. Eng., vol. 27, no. 2, pp. 203–224, 2021. doi: 10.1017/S1351324920000509.
7. L. B. Pape-Haugaard, C. Lovis, and I. C. Madsen, Digital Personalized Health and Medicine: Proceedings of MIE 2020. IOS Press, 2020. [Online]. Available: https://play.google.com/store/books/details?id=k_jtDwAAQBAJ
8. E. Harris, “Large language models answer medical questions accurately, but can’t match clinicians’ knowledge,” JAMA, vol. 330, no. 9, pp. 792–794, 2023. doi: 10.1001/jama.2023.14311.
9. K. Singhal *et al.,* “Large language models encode clinical knowledge,” Nature, vol. 620, no. 7972, pp. 172–180, 2023. doi: 10.1038/s41586-023-06291-2.

10. S. Jha, “Algorithms at the gate—radiology’s AI adoption dilemma,” JAMA, vol. 330, no. 17, pp. 1615–1616, 2023. doi: 10.1001/jama.2023.16049.
11. C. E. Haupt and M. Marks, “AI-generated medical advice—GPT and beyond,” JAMA, vol. 329, no. 16, pp. 1349–1350, 2023. doi: 10.1001/jama.2023.5321.
12. D. Johnson *et al.,* “Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model,” Res. Sq., 2023. doi: 10.21203/rs.3.rs-2566942/v1.
13. A. J. Moy *et al.,* “Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review,” J. Am. Med. Inform. Assoc., vol. 28, no. 5, pp. 998–1008, 2021. doi: 10.1093/jamia/ocaa325.
14. A. J. Thirunavukarasu *et al.,* “Large language models in medicine,” Nat. Med., vol. 29, no. 8, pp. 1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.
15. A. J. Fossa, S. K. Bell, and C. DesRoches, “OpenNotes and shared decision making: a growing practice in clinical transparency and how it can support patient-centered care,” J. Am. Med. Inform. Assoc., vol. 25, no. 9, pp. 1153–1159, 2018. [Online]. Available: https://academic.oup.com/jamia/article/25/9/1153/5047138
16. M. Castaldi and J. McNelis, “Introducing a clinical documentation specialist to improve coding and collectability on a surgical service,” J. Healthc. Qual., vol. 41, no. 3, pp. e21–e29, 2019. doi: 10.1097/JHQ.0000000000000146.
17. M. A. Psotka *et al.*, “Streamlining and reimagining prior authorization under value-based contracts: a call to action from the value in healthcare initiative’s prior authorization learning collaborative,” Circ. Cardiovasc. Qual. Outcomes, vol. 13, no. 7, p. e006564, 2020. doi: 10.1161/CIRCOUTCOMES.120.006564.
18. S. Tripathi, R. Sukumaran, and T. S. Cook, “Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care,” J. Am. Med. Inform. Assoc., vol. 31, no. 6, pp. 1436–1440, 2024. doi: 10.1093/jamia/ocae087.
19. Motics Team, “AI in medical coding: overcoming billing inefficiencies in healthcare,” Motics Blog, Jan. 7, 2025. [Online]. Available: https://www.motics.ai/blog/ai-in-medical-coding-overcoming-billing-inefficiencies-in-healthcare/
20. S. Vestevich, “Medical coding: solutions for avoiding revenue loss,” MedLearn Publishing, Apr. 17, 2023. [Online]. Available: https://racmonitor.medlearn.com/medical-coding-solutions-for-avoiding-revenue-loss/
21. J. Vrdoljak, Z. Boban, M. Vilović, M. Kumrić, and J. Božić, “A review of large language models in medical education, clinical decision support, and healthcare administration,” Healthcare, vol. 13, no. 6, p. 603, 2025. doi: 10.3390/healthcare13060603.
22. J. Huang *et al.,* “A critical assessment of using ChatGPT for extracting structured data from clinical notes,” NPJ Digit. Med., vol. 7, no. 1, p. 106, 2024. doi: 10.1038/s41746-024-01015-0.
23. American Medical Association, “2023 National Health Insurer Report Card,” 2023. [Online]. Available: https://www.ama-assn.org.
24. T. M. Maddox *et al.,* “Generative AI in medicine—evaluating progress and challenges,” N. Engl. J. Med., 2025. doi: 10.1056/NEJMp2412345.
25. J. Guenther, “Quality assurance informs large-scale use of ambient AI clinical documentation,” Permanente Medicine, Mar. 26, 2025. [Online]. Available: https://permanente.org/quality-assurance-informs-large-scale-use-of-ambient-ai-clinical-documentation/.
26. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” Science, vol. 366, no. 6464, pp. 447–453, 2019. doi: 10.1126/science.aax2342.
27. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, 2021. doi: 10.1145/3457607.
28. S. Gilbert, J. N. Kather, and A. Hogan, “Augmented non-hallucinating large language models as medical information curators,” NPJ Digit. Med., vol. 7, no. 1, p. 100, 2024. doi: 10.1038/s41746-024-00989-z.
29. O. K. Gargari and G. Habibi, “Enhancing medical AI with retrieval-augmented generation: a mini narrative review,” Digit. Health, vol. 11, p. 20552076251337177, 2025. doi: 10.1177/20552076251337177.
30. S. Liu, A. B. McCoy, and A. Wright, “Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines,” J. Am. Med. Inform. Assoc., vol. 32, no. 4, pp. 605–615, 2025. doi: 10.1093/jamia/ocaf016.
31. K. E. Goodman, H. Y. Paul, and D. J. Morgan, “AI-generated clinical summaries require more than accuracy,” JAMA, vol. 331, no. 8, pp. 637–638, 2024. doi: 10.1001/jama.2023.28150.
32. U.S. Department of Health and Human Services, “Use of artificial intelligence in health care: privacy and security considerations,” 2023. [Online]. Available: https://www.hhs.gov.
33. G. Malode, P. Mahajan, P. Shelar, A. Sardar, and N. Dhamale, “Automated summarization of healthcare record using LLM,” unpublished.
34. A. Winter *et al.,* “Technological perspective: architecture, integration, and standards,” in Health Information Systems: Technological and Management Perspectives, Cham, Switzerland: Springer, 2023, pp. 51–152. doi: 10.1007/978-3-031-27765-5_3.
35. D. Moser, M. Bender, and M. Sariyar, “A pipeline for automating emergency medicine documentation

using LLMs with retrieval-augmented text generation," Appl. Artif. Intell., vol. 39, no. 1, p. 2519169, 2025. doi: 10.1080/08839514.2024.2519169.

36. N. Kanwal and G. Rizzo, "Attention-based clinical note summarization," in Proc. 37th ACM/SIGAPP Symp. Appl. Comput., 2022, pp. 813–820. doi: 10.1145/3477314.3507072.
37. M. S. Ansari, M. S. A. Khan, S. Revankar, A. Varma, and A. S. Mokhade, "Lightweight clinical decision support system using QLoRA-fine-tuned LLMs and retrieval-augmented generation," arXiv preprint arXiv:2505.03406, 2025.
38. A. Yalunin, D. Umerenkov, and V. Kokh, "Abstractive summarization of hospitalisation histories with transformer networks," arXiv preprint arXiv:2204.02208, 2022.
39. S. Rhazzafe *et al.,* "Hybrid summarization of medical records for predicting length of stay in the intensive care unit," Appl. Sci., vol. 14, no. 13, p. 5809, 2024. doi: 10.3390/app14135809.
40. J. Zhang *et al.,* "The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant," J. Am. Med. Inform. Assoc., vol. 31, no. 9, pp. 1884–1891, 2024. doi: 10.1093/jamia/ocad208.
41. E. Alsentzer *et al.,* "Publicly available clinical BERT embeddings," arXiv preprint arXiv:1904.03323, 2019.
42. X. Yang *et al.,* "GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records," arXiv preprint arXiv:2203.03540, 2022.
43. H. Bundy *et al.,* "Can the administrative loads of physicians be alleviated by AI-facilitated clinical documentation?," J. Gen. Intern. Med., vol. 39, no. 15, pp. 2995–3000, 2024. doi: 10.1007/s11606-024-08869-2.
44. D. Lee *et al.,* "Using large language models to automate data extraction from surgical pathology reports: retrospective cohort study," JMIR Form. Res., vol. 9, no. 1, p. e64544, 2025. doi: 10.2196/64544.
45. D. Van Veen *et al.,* "Clinical text summarization: adapting large language models can outperform human experts," Res. Sq., 2023. doi: 10.21203/rs.3.rs-3471520/v1.
46. D. Hu, B. Liu, X. Zhu, X. Lu, and N. Wu, "Zero-shot information extraction from radiological reports using ChatGPT," Int. J. Med. Inform., vol. 183, p. 105321, 2024. doi: 10.1016/j.ijmedinf.2024.105321.
47. W. I. Wei *et al.,* "Extracting symptoms from free-text responses using ChatGPT among COVID-19 cases in Hong Kong," Clin. Microbiol. Infect., vol. 30, no. 1, pp. 142.e1–142.e6, 2024. doi: 10.1016/j.cmi.2023.08.014.
48. H. Pandey and A. Amod, "Advancing healthcare automation: multi-agent system for medical necessity justification," arXiv preprint arXiv:2404.17977, 2024.
49. Y. Liu, S. Ju, and J. Wang, "Exploring the potential of ChatGPT in medical dialogue summarization: a study on consistency with human preferences," BMC Med. Inform. Decis. Mak., vol. 24, no. 1, p. 75, 2024. doi: 10.1186/s12911-024-02496-y.
50. H. S. Choi *et al.,* "Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer," Radiat. Oncol. J., vol. 41, no. 3, pp. 209–217, 2023. doi: 10.3857/roj.2023.00228.
51. J. Zaretsky *et al.,* "Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format," JAMA Netw. Open, vol. 7, no. 3, p. e240357, 2024. doi: 10.1001/jamanetworkopen.2024.0357.
52. Z. Zhang, Z. Yao, H. Zhou, and H. Yu, "Ehrtutor: enhancing patient understanding of discharge instructions," arXiv preprint arXiv:2310.19212, 2023.