Title: Evaluating Large Language Models for Medical Text Generation and Summarization: Performance, Challenges, and Clinical Relevance

Abstract

The growing reliance on Electronic Health Records (EHRs) has led to serious increases in the documentation burden for care providers. The emergence of Large Language Models (LLMs) as a means to automate the creation and summarization of health-related texts or documents intends to help mitigate clinician burden and improve efficiency of medical documentation. However, domain language, factual inaccuracy, and data privacy are all concerns regarding the trustworthiness of LLMs in clinical care. This study examines the performance of leading LLMs including Flan-T5, BART, and PEGASUS to generate or summarize doctor-patient consultations. In these studies, we compare the summaries generated by LLMs and calculate their accuracy, coherence, and clinical relevance to traditional summaries. After a thorough analysis of our data, we consider the issue of bias and error patterns as well as presenting some of the risks of AI-generated medical documentation to healthcare. We propose the use of the ACI-BENCH data set, containing structured doctor-patient dialogues and summaries, and we clean the data through de-identification and tokenizing for both data privacy and modeling efficiencies, we then input the cleaned and structured dialogues into the LLM to generate abstractsMoreover, the system is able to create new input dialogues, and these are summarized (again, in real-time and with interface engagement), and these are assessed with automated (quantitative metrics, ROUGE, BLEU) and manual (qualitative assessment, clinician) measures for both clinical accuracy and usability. In conclusion, although LLMs can help with the ease and efficacy of documentation, to make any clinically reliable, they would need to be validated via high quality clinical data. Furthermore, due to the absence of standardized frameworks to evaluate precision of LLMs, it would be difficult to properly gauge the clinical reliability amongst the clinicians. No matter the order, this research has demonstrated the need for hybrid models that integrate LLMs with clinician supervision for enhanced usability in medical contexts, and to help promote trust.

Keywords

Medical Text Summarization, Large Language Models, Clinical Documentation, AI in Healthcare, NLP in Medicine

1. Introduction

Medical documentation is an important component of health care that includes electronic health records (EHRs), clinical notes, discharge summaries, and research articles. While these documents are important for patient care, medical research, and legibility, they also contribute to an increasing documentation burden on health professionals. The importance of effective summarization, accurate text generation, and accessing medical information has never been greater.Recent progress in Large Language Models (LLMs), such as GPT-4, BioBERT, and MedPaLM, have shown considerable potential in natural language understanding and generation. These AI-based models could assist in automating medical documentation, summarize lengthy clinical notes, and improve patient-provider communications. However, incorporating LLMs into medical practice has challenges, such as, accuracy of information, understanding domain-specific language, data privacy, and ethical considerations.This study evaluates the role of LLMs in medical text generation and summarization, with opportunities, limitations, and real-world applicability in healthcare.

1.1 Motivation

The digitization of healthcare has caused a huge increase in the amount of documentation being generated, including electronic health records (EHRs), clinical notes, discharge summaries, and diagnostic reports. Documentation is necessary to continue healthcare treatment, support medical research, and comply with guidelines. These documents come at a significant cost and workload burden to health provision, taking time away from clinical time. Physicians and clinicians are often reported to spend much of their work hours writing and managing documentation, which leads to inefficiencies in workflow, increased administrative workload, and burnout among physicians.

Recent advancements in Large Language Models (LLMs) such as GPT-4, BioBERT, and MedPaLM can help alleviate some of the challenges associated with medical documentation through automated summarization, structured reporting, and clinical text analysis. LLMs can help consolidate lengthy patient records, re-interpret complex medical terminology into lay terms, and summarize research articles to determine their applicability to a decision-making process. Regardless of the promise of LLMs, there are still challenges such as accuracy, privacy of data, bias, and ethics which impede accomplishment in healthcare. This study reviews the feasibility and associated risks of LLM-driven medical text generation.

1.2 Challenges in Medical NLP

While LLMs have demonstrated remarkable capabilities in natural language understanding and generation, their application in medical NLP presents unique challenges that must be addressed for safe and effective deployment in healthcare.

1.2.1 Domain-Specific Language and Context Understanding

Medical language is a highly specialized form of communication, with terminologies, abbreviations, and framed information that general NLP models may struggle to understand. For example, “MI” could translate to myocardial infarction (heart attack) or mitral insufficiency (a valve disorder) depending on the context in which it is being used. Furthermore, "SOB" may mean shortness of breath in most medical contexts, but outside of medical culture it would likely be misunderstood. General LLMs often struggle with these types of distinctions, and possible miscommunication to a clinician could have consequences for clinical decision-making. Therefore, creating a domain-specific model trained on high-quality medical datasets is essential to have improved accuracy and context when considering clinical terminology.