HealAI: A Healthcare LLM for Effective Medical Documentation

Sagar Goyal
sagar@deepscribe.tech
DeepScribe Inc.
San Francisco, California, USA

Eti Rastogi
eti@deepscribe.tech
DeepScribe Inc.
San Francisco, California, USA

Sree Prasanna Rajagopal
sree@deepscribe.tech
DeepScribe Inc.
San Francisco, California, USA

Dong Yuan
dong@deepscribe.tech
DeepScribe Inc.
San Francisco, California, USA

Fen Zhao
fen@deepscribe.tech
DeepScribe Inc.
San Francisco, California, USA

Jai Chintagunta
jai@deepscribe.tech
DeepScribe Inc.
San Francisco, California, USA

Gautam Naik
gautam@deepscribe.tech
DeepScribe Inc.
San Francisco, California, USA

Jeff Ward
jeffw@deepscribe.tech
DeepScribe Inc.
San Francisco, California, USA

ABSTRACT

Since the advent of LLM's like GPT4 everyone in various industries has been trying to harness their power. Healthcare is an industry where this is a specifically challenging problem due to the high accuracy requirements. Prompt Engineering is a common technique used to design instructions for model responses, however, its challenges lie in the fact that the generic models may not be trained to accurately execute these specific tasks. We will present our journey of developing a cost-effective medical LLM, surpassing GPT4 in medical note-writing tasks. We'll touch upon our trials with medical prompt engineering, GPT4's limitations, and training an optimized LLM for specific medical tasks. We'll showcase multiple comparisons on model sizes, training data, and pipeline designs that enabled us to outperform GPT4 with smaller models, maintaining precision, reducing biases, preventing hallucinations, and enhancing note-writing style.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

Large language models, medical note writing, EHR, healthcare, domain-specific LLM, prompt engineering, medical domain, fine-tuning, retrieval, pretraining, long context LLM

ACM Reference Format:
Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. 2024. HealAI: A Healthcare LLM for Effective Medical Documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3616855.3635739

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WSDM '24, March 4–8, 2024, Merida, Mexico
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0371-3/24/03.
https://doi.org/10.1145/3616855.3635739

1 OUTLINE OF THE PRESENTATION

In this presentation, we will talk about our innovative AI solution which streamlines the creation of SOAP (Subjective, Objective, Assessment, and Plan) notes [4] from doctor-patient conversations. SOAP notes are widely used structured documents that help healthcare professionals track patients' health records and effectively communicate with other providers. Our presentation will focus on the end-to-end development cycle of designing and training our own domain-specific LLM. We will be talking about (a) SOAP Notes and challenges in healthcare industry, (b) Designing the solution with GPT4 and its challenges, (c) Designing our LLM for note creation, and finally, (d) Quantization and Model Deployment.

1.1 SOAP Notes and challenges in healthcare industry

SOAP (Subjective, Objective, Assessment, and Plan) notes are an essential healthcare communication tool used by providers to document essential patient information. Using AI for medical note-writing comes with a variety of challenges. These include ensuring accuracy to prevent potential misdiagnoses, correctly interpreting complex medical language, and addressing hallucinations and biases that could influence patient care and results. Both of these could severely affect the precision of note writing, potentially leading to unsuitable therapeutic decisions or clinical procedures.

In this section, we will thoroughly explore these challenges, supplemented with real-life examples.

1.2 Designing the solution with GPT4 and its challenges

In this segment of the presentation, we will talk about how we leveraged GPT-4 in solving the task of medical documentation, including the safeguards we built to address hallucinations and biases [2]. We will also explore different prompting strategies such as CoT [6], ReAct [8] delving into noteworthy examples of prompts and how they improved the quality of the talk. We will then explore the challenges and the limitations of using a general LLM such as GPT-4 which struggles in understanding sophisticated medical

All authors contributed equally to this work terminologies necessitating the need for a custom LLM for medical documentation purposes.

1.3 Designing our custom LLM for note creation

Given the challenges and restrictions we faced with prompt engineering and current models for domain-specific tasks, it was logical for us to progress towards developing our own custom LLM. In this section, our discussion will revolve around the end-to-end life cycle of designing this model, aimed to efficiently handle diverse doctor-patient interactions.

We will begin by going over our Instruction Fine-Tuning [9] technique done over an already existing open-source LLM [3] which imparts enhanced capabilities to the model to follow diverse instructions, including those necessary for performing medical documentation tasks. We also aim to elaborate on the special tasks and datasets curated by us exclusively for the purpose of training this model.

The second part of this section will include details about the Supervised Fine-Tuning of actual medical note-creation data generated through GPT4 prompting and the use of human scribes. We will talk about metrics that we use to rate and compare different notes which also serve as guardrails to guide our training design. We will talk about the challenges stemming from longer patient-doctor interactions and how even the largest context-size model is not able to capture and generate the most accurate note. We will also talk about doctor-preferred medical writing styles and getting our models to emit more acceptable outputs. We will also talk about how this LLM can be trained to make edits on a medical note as the human (doctor) instructs - saving a lot of time for the doctor.

In the final part of this section we will talk about designing the note-creation pipeline using Retrieval Augmented Generation (RAG) [1] [5] - solving our large context problem and which has shown to perform better than using raw LLM directly [7].

1.4 Quantization and Model Deployment

In this final section, we will talk about the need for quantization, the options available, the impact on performance, and the path to production. We will discuss various options available in the industry and things specific to the healthcare industry such as HIPAA compliance which need to be taken into consideration.

2 CONCLUSIONS

Medical Documentation is complex due to critical healthcare aspects, hard-to-understand terminologies, and diversity in interactions that can lead to system failures. Existing SoTA LLM's are insufficient and can be expensive over time. Developing our own LLM allows for cost-effective, customizable solutions with better transparency and continual improvement opportunities. We believe the real power of technology is harnessed when we efficiently use it in the industry where it has the opportunity to directly impact lives.

3 COMPANY PORTRAIT

DeepScribe is healthcare's first enterprise-grade fully automated AI medical scribe. Unlike other AI scribe companies, DeepScribe is 100% customizable, allowing clinicians to create complete and accurate documentation that best suits their needs - all while saving up to 4 hours a day and removing administrative bottlenecks. Not only does DeepScribe improve the well-being of clinicians, it has also been proven to improve patient care, as well as increase revenue and reimbursement.

4 PRESENTERS' BIO

Sagar Goyal is a senior NLP Engineer at Deepscribe, leading initiatives for instruction fine-tuning in the design of custom domain-specific LLMs. In this role, he has led the team to have the first in-house trained deployable LLM. Previously, he has worked at Snap, Microsoft, and MPI-Inf in various AI/ML roles. He completed his Bachelors in Technology from IIT Delhi in CSE and has published papers in WWW, NeurIPS, and PAKDD.

Eti Rastogi is a senior NLP Engineer at Deepscribe. Her work at DeepScribe has been mainly focused on developing user-friendly AI-powered medical applications by collaborating closely with medical professionals. She spearheaded the initiative to develop a customized Medical LLM designed to follow instructions for medical documentation tasks. Prior to this, she served as an ML Engineer at C3.ai, where she developed distributed ML pipelines. Eti earned her M.S. degree in ECE from Carnegie Mellon University and has specialized in the intersection of Deep Learning and NLP.

5 ACKNOWLEDGMENTS

We thank the DeepScribe AI team and leadership for playing a key role in executing this project.

REFERENCES

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

[2] Eti Rastogi. 2023. Overcoming Hallucinations and Biases in LLM: A Step Towards Reliable Medical Application. (2023). https://www.deepscribe.ai/resources/overcoming-hallucinations-and-biases-in-llm-a-step-towards-reliable-medical-applications?utm_content=260768543&utm_medium=social&utm_source=linkedin&hss_channel=lcp-19018424

[3] Azizi S. Tu T. et al. Singhal, K. 2023. Large language models encode clinical knowledge. (2023). https://doi.org/10.1038/s41586-023-06291-2

[4] Sassan Ghassemzadeh Vivek Podder, Valerie Lew. 2022. *SOAP Notes*. StatPearls Publishing, Treasure Island (FL).

[5] Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. 2023. InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining. *arXiv preprint arXiv:2310.07713* (2023). https://doi.org/10.48550/arXiv.2310.07713

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

[7] Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets Long Context Large Language Models. *arXiv preprint arXiv:2310.03025* (2023). https://doi.org/10.48550/arXiv.2310.03025

[8] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022). https://doi.org/10.48550/arXiv.2210.03629

[9] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206* (2023). https://doi.org/10.48550/arXiv.2305.11206