A Benchmark of Domain-Adapted Large Language Models for Generating Brief Hospital Course Summaries

Asad Aali[1,*], Dave Van Veen[2, 6], Yamin Ishraq Arefeen[1], Jason Hom[5], Christian Bluethgen[5, 9], Eduardo Pontes Reis[6, 8], Sergios Gatidis[3, 6], Namuun Clifford[7], Joseph Daws[10], Arash S. Tehrani[10], Jangwon Kim[11], and Akshay S. Chaudhari[3, 4, 6]

[1]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA
[2]Department of Electrical Engineering, Stanford University, Stanford, CA, USA
[3]Department of Radiology, Stanford University, Stanford, CA, USA
[4]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[5]Department of Medicine, Stanford, CA, USA
[6]Center for Artificial Intelligence in Medicine and Imaging, Palo Alto, CA, USA
[7]School of Nursing, The University of Texas at Austin, Austin, TX, USA
[8]Albert Einstein Israelite Hospital, São Paulo, Brazil
[9]University Hospital Zurich, Zurich, Switzerland
[10]One Medical, San Francisco, CA, USA
[11]Amazon, Seattle, WA, USA
[*]Corresponding author: asad.aali@utexas.edu

ABSTRACT

Background: Brief hospital course (BHC) summaries are common clinical documents generated by summarizing clinical notes. While large language models (LLMs) depict remarkable capabilities in automating real-world tasks, their capabilities for healthcare applications such as BHC synthesis have not been shown. To enable the adaptation of LLMs for BHC synthesis, we introduce a novel benchmark consisting of a pre-processed dataset extracted from MIMIC-IV notes, encapsulating clinical note, and brief hospital course (BHC) pairs. As part of the benchmark, we assess the performance of two general-purpose LLMs and three healthcare-adapted LLMs to improve BHC synthesis from clinical notes.

Methodology: Using clinical notes as input for generating BHCs, we apply prompting-based (using in-context learning) and fine-tuning-based adaptation strategies to three open-source LLMs (Clinical-T5-Large, Llama2-13B, FLAN-UL2) and two proprietary LLMs (GPT-3.5, GPT-4). We quantitatively evaluate the performance of these LLMs across varying context-length inputs using conventional natural language similarity metrics. We further perform a qualitative study where five diverse clinicians blindly compare clinician-written BHCs and two LLM-generated BHCs for 30 samples across metrics of comprehensiveness, conciseness, factual correctness, and fluency. We compare reader preferences for the original and LLM-generated summary using Wilcoxon Signed-Rank tests.

Results: The Llama2-13B fine-tuned LLM outperformed other domain-adapted models given quantitative evaluation metrics of BLEU and BERT-Score. GPT-4 with prompting adaptation showed more robustness to increasing context lengths of clinical note inputs than fine-tuned Llama2-13B. The reader study depicted a significant preference for summaries generated by GPT-4 with in-context learning compared to both Llama2-13B fine-tuned summaries and the original summaries ($p < 0.001$).

Conclusion: We present a new benchmark and pre-processed dataset for using LLMs in BHC synthesis from clinical notes. We observe high-quality summarization performance for both in-context proprietary and fine-tuned open-source LLMs using both quantitative metrics and a qualitative clinical reader study. We propose our work as a benchmark to motivate future works to adapt and assess the performance of LLMs in BHC synthesis.

Introduction

Clinicians spend significant time with clinical documentation activities, which are time-consuming and reduce the time that can be spent on patient-facing activities[1–3]. Brief hospital course (BHC) summaries are encapsulated within clinical notes, which

Table 1. **a)** A sample of our novel pre-processed clinical notes dataset, extracted from raw MIMIC-IV notes.

Input	Example
SEX	F
SERVICE	SURGERY
ALLERGIES	No Known Allergies
CHIEF COMPLAINT	Splenic laceration
MAJOR PROCEDURE	NONE
HISTORY OF PRESENT ILLNESS	s/p routine colonoscopy this morning with polypectomy (report not available) ...
PAST MEDICAL HISTORY	Mild asthma, hypothyroid
FAMILY HISTORY	Non-contributory
PHYSICAL EXAM	Gen: Awake and alert CV: RRR Lungs: CTAB Abd: Soft, nontender, nondistended
PERTINENT RESULTS	03:45 PM BLOOD WBC-5.5 RBC-3.95 Hgb-14.1 ...
MEDICATIONS ON ADMISSION	1. Levothyroxine Sodium 100 mcg PO DAILY 2. Flovent HFA (fluticasone) ...
DISCHARGE DISPOSITION	Home
DISCHARGE DIAGNOSIS	Splenic laceration
DISCHARGE CONDITION	Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive ...
DISCHARGE INSTRUCTIONS	You were admitted to ... in the intensive care unit for monitoring after a ...

Output	Example
BRIEF HOSPITAL COURSE	Ms. ... was admitted to ... on .... After getting a colonoscopy and polypectomy, she ...

b) Relevant statistics for the pre-processed dataset split across multiple context length ranges for adaptation tasks.

Context Range	Samples	Input Tokens	BHC Tokens
0 - 1,024	2,000	711 ± 199	104 ± 43
1,024 - 2,048	2,000	$1,471 \pm 275$	148 ± 36
2,048 - 4,096	2,000	$2,496 \pm 388$	225 ± 55

are commonly written for hospitalization cases. Synthesizing BHCs from multiple clinical notes requires substantial clinician time and expertise, with errors possibly causing patient harm[4]. Automating or accelerating BHC note generation is important since discharge summaries can either lack essential information or contain incorrect information, and may be incomplete before follow-up appointments[5]. The recent success of generative large language models (LLMs) has provided a promising avenue for this clinically important task of BHC synthesis from clinical notes.

In this study, we present a novel benchmark[1] and dataset[2] curated from MIMIC-IV notes[6] for the task of BHC summarization. We benchmark the performance of domain-adapted LLMs in the BHC summarization task to motivate future benchmarking studies. We compare several state-of-the-art (SOTA) LLMs: proprietary general-purpose LLMs, such as GPT-3.5[7] and GPT-4[8,9], open-source LLMs, such as Llama2[10], FLAN- UL2[11], and a T-5 model pre-trained on medical text[12]. We explore prompting-based (using in-context learning [ICL]) and fine-tuning-based adaptation strategies for the proprietary and open-source LLMs used in our benchmarking study to identify the best model and adaptation strategy. We further benchmark and provide a comprehensive quantitative analysis for BHC synthesis using well-established metrics measuring the syntactic and semantic similarity between LLM-generated and clinician-written BHCs. With such metrics, we further explore the robustness of domain-adapted LLMs across clinical notes with higher context lengths. Finally, we conduct a rigorous clinical reader study involving five expert clinicians reviewing clinical notes and BHC pairs to ensure adequate qualitative clinical evaluation of the LLM-generated summaries across dimensions of comprehensiveness, conciseness, factual correctness, and fluency. Our novel benchmark which includes the preprocessed dataset, LLM adaptation, quantitative and clinical evaluation, seeks to improve the possibilities of using LLMs for clinical text summarization tasks.

[1]We will be releasing the code for data preprocessing, model adaptation, and evaluation upon publication.
[2]Our preprocessed clinical notes dataset will be made publicly available at the time of publication.

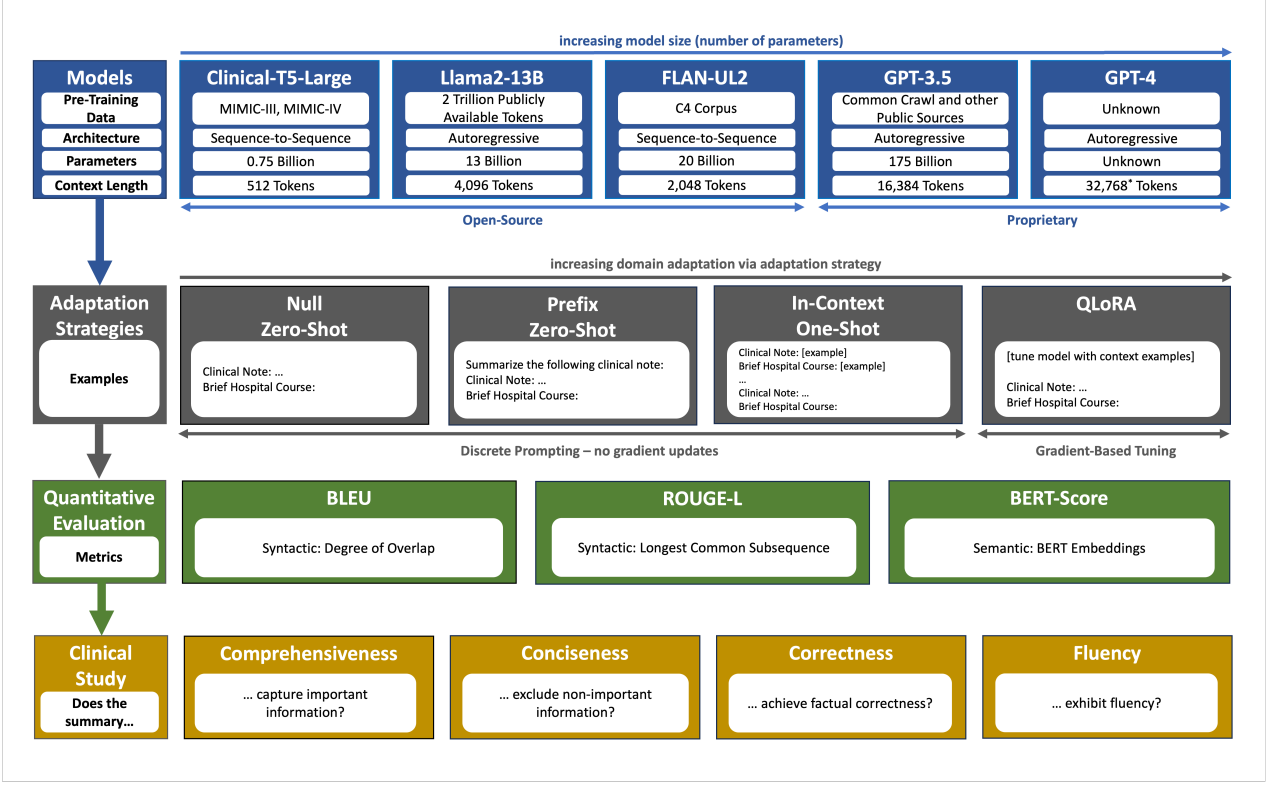

Figure 1. Overall schematic of our study. We evaluate a variety of **models**, including open-source models containing up to 20 billion parameters, and larger-scale proprietary models. Each model is adapted to the summarization task using the **adaptation strategies** displayed (except QLoRA is not applied to GPT-3.5 and GPT-4 since these proprietary models do not provide access to model weights). We evaluate each model's performance by comparing its outputs with expert clinician summaries. Each model paired with the adaptation strategy is evaluated using **quantitative similarity metrics**. Finally, we perform a **clinical study** where five clinicians rate three summaries (randomized order) for every summarization task: best-performing open-source model, best-performing proprietary model, and clinician-written. The * indicates GPT-4's maximum context length at the time of experimentation. Later, the maximum context length was increased to 128,000.

Related Works

Recent advancements in natural language processing (NLP) have been characterized by the adoption of transformer-based language models[13]. Transformer models such as Bidirectional Encoder Representations from Transformers (BERT)[14] and Unidirectional Generative Pre-trained Transformer 2 (GPT-2)[15] paved the way for training large-scale autoregressive models. The evolution of transformer models involved pretraining on large corpora of text data and subsequent fine-tuning on domain-specific data, as demonstrated by GPT-3[16], PaLM[17], and T5[18]. After extensive pre-training, these LLMs are often optimized for alignment with human preferences, typically using instruction tuning[19]. Prior works have demonstrated successful adaptation of LLMs in the medical domain by pretraining or fine-tuning LLMs on clinical text[20].

Methods

Dataset

For our task of synthesizing BHCs from clinical notes, we utilized the raw MIMIC-IV-Note dataset[6], a compilation of 331,794 de-identified discharge summaries from 145,915 patients admitted to the Beth Israel Deaconess Medical Center for a diverse range of medical scenarios. We pass MIMIC-IV[6] notes through our pre-processing pipeline (described below) to extract a novel dataset containing 270,033 clinical notes and BHC pairs. This dataset is tailored for the task of BHC summarization with a focus on the relationship between clinical notes and corresponding BHCs. Creating this benchmark dataset is crucial for others to replicate findings and to compare future works.

Data Processing

Our pipeline uses regular expressions to remove special characters, and unnecessary text, and to separate each section of the clinical note using line breaks to create appropriate section headings. We utilize pre-trained tokenizers from HuggingFace to further process the dataset for LLM training/inference. We extract the BHCs from each clinical note to create a labeled dataset with relevant input-output, clinical notes–BHC pairs. Our pre-processing results generate 270,033 samples with an average input token length of 2,267 $\pm$ 914 and an average output token length of 564 $\pm$ 410. We further create three independent train and test datasets for our adaptation tasks through random sampling as described in Table 1. To allow for an unbiased analysis, we also randomly sample 100 independent test examples from the pre-processed dataset for each context length range. We use the smallest context length dataset (0 - 1,024 tokens) for the initial model evaluation to make extensive quantitative analysis feasible.

Large Language Models

We categorize the LLMs used in this study as open-source LLMs and proprietary LLMs.

Open-Source LLMs

The models in this category contain up to 20 billion parameters and can be fine-tuned on consumer-grade GPUs. These models were trained either with a sequence-to-sequence (seq2seq) or autoregressive objective. The seq2seq architecture maps input sequences directly to output sequences. In contrast, autoregressive architectures predict the next token given the preceding context. The following models were considered:

- Clinical-T5-Large is a "text-to-text transfer transformer"[12] pre-trained on medical text, that employs transfer learning with the Sequence-to-Sequence (seq2seq) architecture to clinical tasks. Clinical-T5-Large supports only a limited context window of 512 tokens.
- FLAN-UL2 hails from the T5[18] family, utilizing the seq2seq architecture. It employs a modified pre-training procedure to incorporate an expanded context length of 2,048 tokens instead of only 512 tokens context length of the original FLAN-T5. FLAN-UL2 training also involved additional instruction tuning[21] to enhance the model's capability to understand complex narratives.
- Llama2-13B stems from the Llama family of LLMs[10] and is an open-source autoregressive model tailored for expanded pretraining on 2 trillion tokens. Llama2-13B allows up to 4,096 tokens as input.

Proprietary LLMs

This category includes large-scale proprietary autoregressive models that utilize reinforcement learning with human feedback (RLHF) to further improve performance over instruction tuning using feedback from expert human evaluators.

- ChatGPT (GPT-3.5)[7] contains 175 billion parameters and has been extensively fine-tuned for general tasks using human feedback, enhancing its ability to capture intricate details within any summarization task. GPT-3.5 can support input context lengths of up to 16,384 tokens.
- GPT-4[9] achieves state-of-the-art overall NLP performance and supports an input context window of 128,000 tokens, which is the highest compared to other models listed in Figure 1, making it a suitable choice in multi-document summarization scenarios.

Clinical-T5-Large[12], and FLAN-UL2[11] have a maximum input context length limit of 512 and 2,048 tokens, respectively. Hence, for the context length analysis in our results section, we only select models that allow at least 4,096 context length inputs, allowing exploration of model behavior for varying extents of the input length. We used the GPT-3.5 and GPT-4 API for all our summarization tasks.

Adaptation Strategies

For each pre-trained model outlined in the previous section, we apply a series of lightweight domain adaptation methods as shown in Figure 1. We only explore adaptation through discrete prompting (no fine-tuning) for large-scale proprietary LLMs since the model weights are not accessible. The adaptation methods mentioned below gradually increase the level of adaptation to a downstream task.

- Null Prompting: This is a discrete zero-shot adaptation strategy[22] where the clinical note is supplied with a basic prompt, such as "brief hospital course". This approach is a baseline technique to evaluate how models respond to minimal task-specific guidance.

- Prefix Prompting: Building upon the null prompting approach, we now provide a more detailed instructional prompt[21,22], like "summarize the following clinical note," which serves to provide the model with additional context for the BHC synthesis task.
- In-Context Prompting: We employ in-context learning (ICL)[21] using a discrete few-shot prompt. ICL involves providing the LLM with examples and instructions within the prompt itself, demonstrating a desired behavior[21]. We enhance our null prompting strategy by prepending one example clinical note and BHC pair, selected randomly from the training dataset.
- Quantized Low-Rank Adaptation (QLoRA)[23,24]: This technique utilizes 4-bit quantization to enable parameter efficient fine-tuning[25] by injecting rank-decomposition matrices into each layer of the model. We use QLoRA to fine-tune LLMs on the supervised task of BHC synthesis using our curated dataset. QLoRA fine-tuning was performed on a mix of NVIDIA A10G and V100 GPU machines.

Our adaptation methods are carefully chosen to strike a balance between computational efficiency and task-specific adaptability. Null prompting and prefixed prompting serve as foundational approaches, allowing us to gauge model responsiveness with minimal and augmented task-specific cues. In-context learning further refines adaptation by introducing contextually relevant examples. QLoRA provides an efficient means of fine-tuning[25] larger models, which is crucial for handling the complexity of the BHC summarization task.

Experimental Setup

Quantitative Evaluation

To identify the best-performing open-source and proprietary models, we conducted a performance analysis, where we selected a subset of 70 independent samples from the 100 test samples in the 0-1,024 context range. The remaining 30 independent samples from this range were separated for a reader study described below, to avoid data leakage. To accurately assess the performance of our BHC summarization models, we employ quantitative metrics commonly used in summarization tasks to evaluate the syntactic similarity and semantic relevance of the generated BHC summaries. Bilingual Evaluation Understudy (BLEU)[26] evaluates the overlap between the generated and reference summaries using a weighted average of 1 to 4-gram precision. ROUGE-L[27] assesses the precision and recall of the longest common subsequence. BLEU and ROUGE-L depict lexical overlap between generated and reference summaries. In contrast, we also use semantic similarity metrics like BERT-Score[28] which leverages contextual BERT embeddings to evaluate the similarity between the generated and reference summaries. Finally, we perform an independent context-length quantitative analysis using fine-tuned Llama2-13B and GPT-4 with ICL, the best-performing open-source and proprietary LLMs allowing context lengths greater than 4,000 tokens as input using all 100 test samples from each of the three context-length test sets. To assess potential health equity implications, we perform a stratified subgroup analysis across patient-reported sex when assessing the quantitative performance of BHC summarization techniques.

Qualitative Evaluation

We conducted a reader study with five board-certified clinicians to compare BHCs written by expert clinicians during clinical practice from the MIMIC dataset to BHCs generated by our best-performing open-source and proprietary models. We utilized 30 independent clinical notes and BHC pairs from the 0 - 1,024 context-length dataset. Each of the 30 independent clinical notes were paired with three BHCs: (1) original BHC written by an expert clinician, (2) BHC generated by the best-performing open-source LLM, and (3) BHC generated by the best-performing proprietary LLM. These were then presented to a group of five diverse clinicians from multiple institutions and different backgrounds, each with a clinical focus in internal medicine, nursing, thoracic radiology, pediatric radiology, and neuroradiology (with 11, 7, 5, 6, and 7 years of experience, respectively). We chose diverse clinical expertise and institutions to reflect the diversity in clinical practice. The BHCs were presented in randomized and blinded order to the clinicians. The five clinicians ranked the 30 note-summary pairs on a Likert scale of 1 (poor) to 5 (excellent) based on four evaluation criteria:

- Comprehensiveness: How well does the summary capture important information? This assesses the recall of clinically significant details from the input text.
- Conciseness: How well does the summary exclude non-important information? This compares how well the summary is condensed, considering the value of a summary decreases with superfluous information.
- Factual Correctness: How well does the summary agree with the facts outlined in the clinical note? This evaluates the precision of the information provided.
- Fluency: How well does the summary exhibit fluency? This assesses the readability and natural flow of the content.

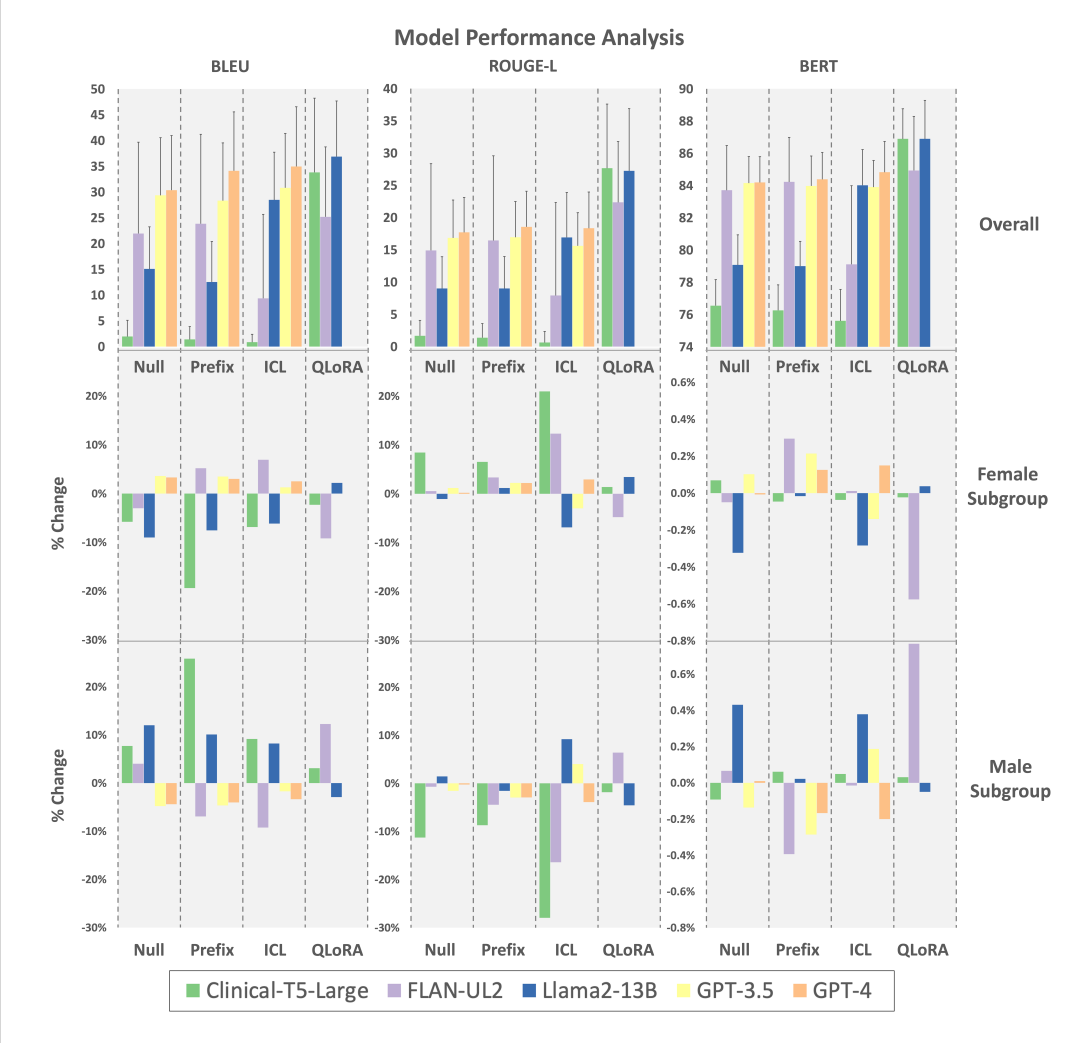

Figure 2. Quantitative metric results for each choice of model, across increasing domain-adaptation strategies. In summary, QLoRA as an adaptation strategy outperforms other adaptation methods. Specifically, QLoRA Llama2-13B outperforms other models in BLEU score, while achieving comparable performance to Clinical-T5-Large in BERT-Score and ROUGE-L.

Statistical Analysis

We explored the following two hypotheses: 1) Does our pool of five diverse clinicians exhibit a preference for LLM-generated summaries over clinician-written ones?, and 2) Does our pool of five diverse clinicians exhibit a preference for summaries generated by adapted proprietary LLMs over summaries generated by adapted open-source LLMs?. We perform these tests across each of the four reader study evaluation criteria, comparing the best-performing proprietary and open-source LLMs with clinician summaries. We performed significance testing using non-parametric Wilcoxon signed-rank tests, with a significance level $\alpha = 0.05$, and with Bonferroni corrections using XLSTAT.

Results

Model Performance Analysis

Despite being the smallest model, Clinical-T5-Large (750M parameters), exhibited competitive performance after QLoRA fine-tuning and achieved the largest performance improvement with increasing adaptation (Figure 2). FLAN-UL2 displayed strong performance across all adaptation strategies, except ICL. Its improvement in performance with fine-tuning was limited. Llama2-13B performance improved substantially following QLoRA fine-tuning, achieving the highest BLEU and BERT scores across all adaptation strategies and LLMs. GPT-3.5 showed good overall performance but exhibited limited benefits with increasing domain adaptation. GPT-4 with ICL outperformed GPT-3.5 and became the top-performing proprietary LLM. When

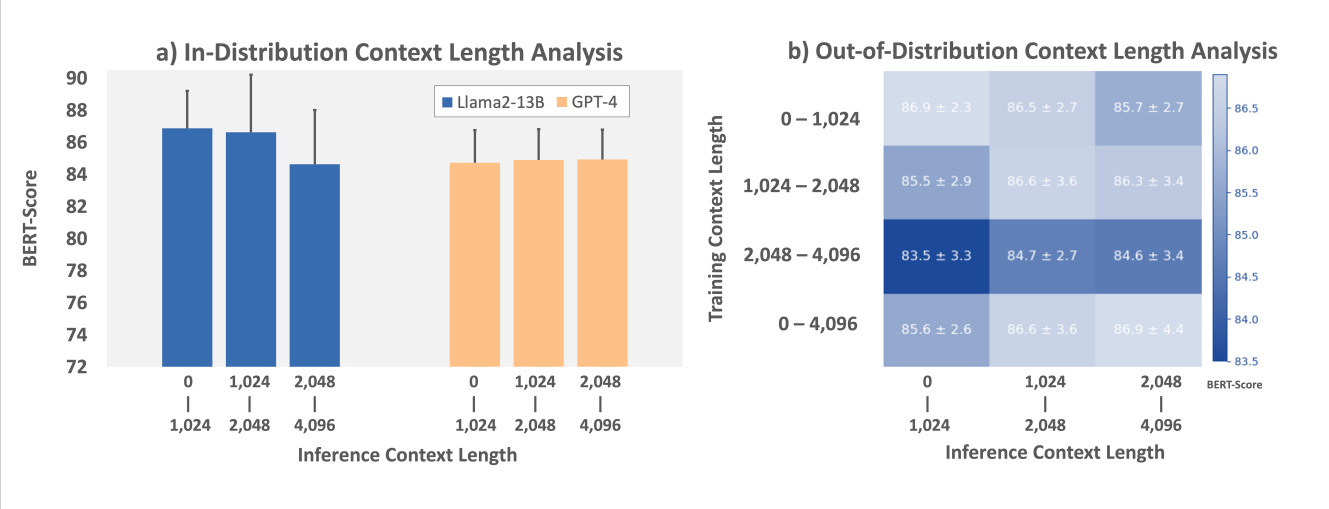

Figure 3. a) Quantitative evaluation metrics across increasing input context lengths. GPT-4 shows consistency in performance whereas Llama2-13B shows a drop in summarization with increasing context length inputs. **b)** Context size analysis for QLoRA Llama2-13B (in/out-of-distribution), where each item on the y-axis displays an independent model fine-tuned on samples from a specific context length range. The summarization performance of the combined model trained on 0 - 4,096 context length inputs outperforms other models with longer input clinical notes at inference (more than 1,024 tokens).

evaluating the deviation of each model's overall summarization performance across the subgroups of patient-reported sex, we noticed large variations with the Clinical-T5-Large model (Figure 2). With increasing adaptation from null prompting to QLoRA, GPT models showed higher variance from their original ROUGE-L and BERT scores. LLMs after QLoRA adaptation generally showed lower variations among subgroups than other adaptation strategies.

Context Length Analysis

We compared the performance of our LLMs when the adaptation (training) and inference (testing) context lengths were identical lengths (in-distribution) (Figure 3 a)). QLoRA Llama2-13B sees a drop in in-distribution performance. In contrast, in-context GPT-4 displays consistent performance in-distribution, across all metrics as the context length of the train and test set increases. Overall, in-context GPT-4 exhibits more robustness than QLoRA Llama2-13b as context lengths increase. We further explored LLM performance with differing training and testing context lengths (Figure 3 b)). We observed that adapted models at testing perform best on samples with context lengths similar to what they were trained on (in-distribution), independent of the length of the samples. Subsequently, the performance of models trained on smaller context-length samples deteriorates when testing on longer-context-length samples.

Clinical Reader Study

We find that our five clinical readers strongly prefer in-context GPT-4 summaries for attributes of comprehensiveness, conciseness, factual correctness, and fluency, compared to the original clinician-written summaries ($p < 0.001$) and QLoRA Llama2-13B summaries ($p < 0.001$) (Figure 4). However, we find no significant differences between reader preferences comparing Llama2 summaries and the original clinician-written summaries ($p = 0.05$). Overall, GPT-4 scores have a narrower range, showcasing lower variation. We also report the proportion of scores from readers between the range of 3 to 5 for GPT-4, Llama2-13B, and clinician summaries: 100%, 47%, and 40%, respectively. This study underscores the strength of LLMs in generating summaries that are both statistically superior and preferred by our panel of clinicians in the majority of cases.

Discussion

In this study, we present a new open-source benchmark for generating BHC summaries from multiple clinical notes using domain-adapted open-source and closed-source LLMs, coupled with quantitative and qualitative evaluation. Our benchmark shows that open-source LLMs can produce high-quality summaries and that Llama2-13B comes out as a superior choice for further analysis of open-source LLMs based on its performance and an expanded context length window. The subgroup analysis further emphasizes the importance of a multi-faceted analysis when evaluating models. Furthermore, our context-length analysis showcases the benefit of fine-tuning LLMs on datasets with a wide range of context lengths, especially when summarizing longer clinical notes.

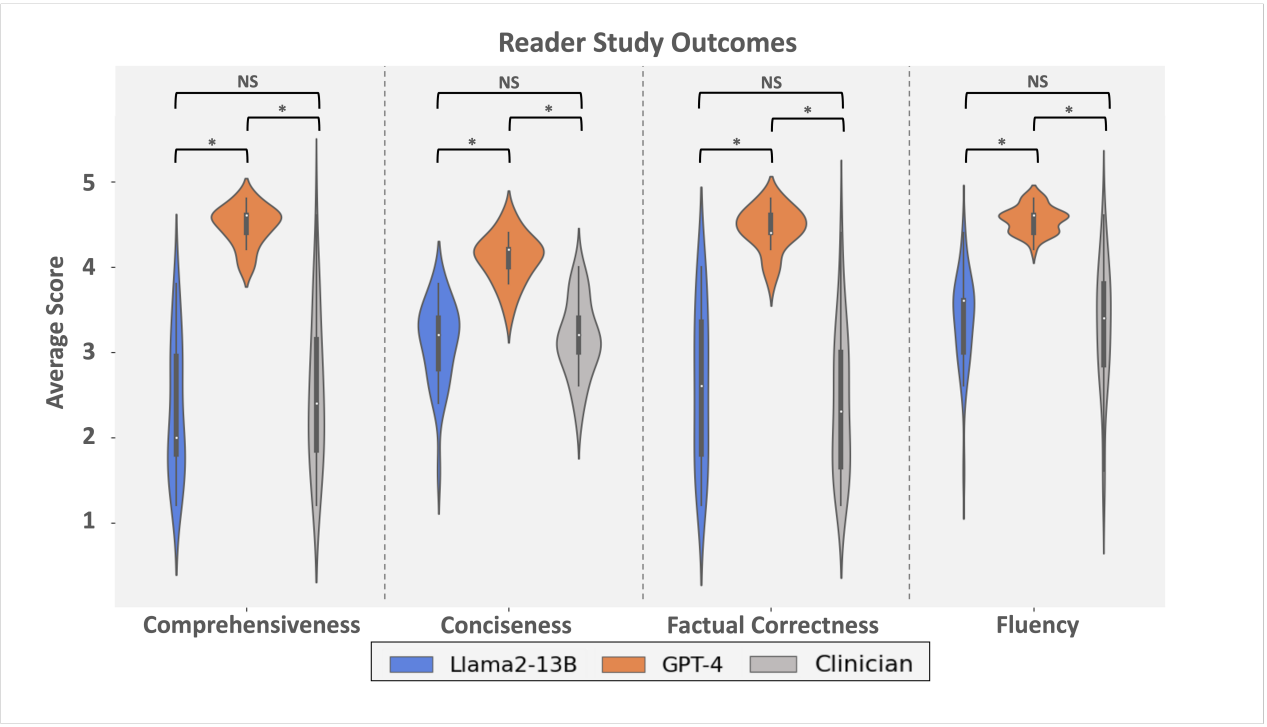

Figure 4. Violin plot showing results from the reader study with five clinicians. Clinicians exhibit a clear and strong preference for in-context GPT-4 (adapted large-scale proprietary LLM) summaries over QLoRA Llama2-13B (adapted open-source LLM) and clinician-written summaries with statistical significance by the Wilcoxon signed-rank test, achieving $^*p < 0.001$ across each attribute. (NS: Not Significant).

Our evaluation of LLM summaries suggests that adapted open-source LLMs like Llama2-13B can produce summaries similar in quality to clinician summaries as depicted in reader preferences (Figure 4). Open-source models also provide the benefit of ease in implementation, where sending data via an API is not possible. These models can be fine-tuned and implemented on-premise, as their weights are publicly available. Our benchmark also reports strong summarization performance from Clinical-T5-Large, while it is the smallest parameter model. A possible explanation of this could be extensive pre-training on medical text. However, Clinical-T5-Large also presented a large variation in performance across subgroups, making it a less suitable choice for deployment.

In our study, we combine evaluation via quantitative metrics and qualitative assessments to capture a comprehensive overview of the performance of different summarization methods. It is crucial to recognize that quantitative measures, such as NLP metrics, and qualitative assessments by clinical readers gauge different aspects of summarization quality. NLP metrics, for instance, quantify the level of (textual, semantic, contextual) similarity between two summaries, whereas clinical readers inform perceptual (subjective) quality differences. We observed that fine-tuned Llama2-13B consistently outperformed in-context GPT-4 in quantitative similarity metrics, suggesting a closer resemblance to clinician-written summaries. However, the reader study results present a nuanced narrative where clinicians express a preference for GPT-4-generated summaries. This finding eludes a critical point: high quantitative similarity scores do not necessarily translate to a "better summary" from a clinician's perspective, as well as across subgroups. Hence, we advocate for the critical role of human evaluation. Enhancing the alignment between quantitative metrics and qualitative assessments will enable a more holistic evaluation for clinical text summarization[31].

Recent adaptations of LLMs in the clinical domain emphasize their potential in understanding medical language. Noteworthy recent works[29–31] have used open-source datasets and displayed the performance of LLMs in multiple summarization tasks in medicine: radiology reports, patient health questions, progress notes, and doctor-patient conversations. Some prior works have further explored the use of proprietary models like ChatGPT[7] in radiology report summarization[32], as well as discharge summarization[33,34]. A recent work presents a large-scale multi-document dataset for brief hospital courses, promoting the advancement of hospital visit summarization[35]. Another work[36] further explores domain adaptation of LLMs through fine-tuning for improved summarization of clinical notes. Models like Clinical-T5[12] and Med-PaLM[37] have shown promise in the clinical domain for tasks such as named entity recognition, natural language inference, and question-answering. Our study

takes inspiration from related works to develop a novel benchmark using our preprocessed dataset, presenting a comprehensive evaluation of adapted LLMs. Through rigorous quantitative, qualitative, and subgroup evaluation, we ultimately aim to bridge the gap between machine learning research and clinical adoption.

While noting the strengths of our study, we also discuss its limitations. Our selection of LLMs for the BHC adaptation task is not comprehensive. As new adaptation strategies are introduced, and new LLMs are trained and released, a promising direction for future works would be to evaluate their performance on our preprocessed dataset and compare it with our proposed benchmark. Another limitation of our work is the limited availability of publicly available datasets for clinical note summarization, making it difficult to gauge the performance of our adapted LLMs in diverse BHC synthesis datasets across hospitals and clinical practices.

Conclusion

In this investigation, we developed a benchmarking dataset and performed a comprehensive quantitative and qualitative evaluation of using LLMs for synthesizing BHCs from clinical notes using adaptation strategies for both open-sourced and closed-sourced models. Via a clinical reader study, we depicted that adapted open-source models can match the quality of clinician-generated summaries, while adapted proprietary models can outperform the quality of clinician-generated summaries across dimensions of comprehensiveness, conciseness, factual correctness, and fluency. Our study can help define a framework for a technical and clinical evaluation of LLMs and our findings have the potential to alleviate the documentation burden on clinicians.

Acknowledgments

Computing resources were partially provided by One Medical and Stanford University. Microsoft provided Azure OpenAI credits for this project via the Accelerate Foundation Models Academic Research (AFMAR) program. A.C. receives research support from NIH grants R01 HL167974, R01 AR077604, R01 EB002524, R01 AR079431, and P41 EB027060; from NIH contracts 75N92020C00008 and 75N92020C00021; from Stanford Center for Artificial Intelligence and Medicine, Stanford Institute for Human Centered AI, from Stanford Center for Digital Health, from Stanford Cardiovascular Institute, from Stanford Center for Precision Health and Integrated Diagnostics; from GE Healthcare, Philips and Amazon.

References

1. Moy, A., Schwartz, J., Chen, R., et al. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *Journal Of The American Medical Informatics Association*. **28**, 998-1008 (2021)
2. Chaiyachati, K., Shea, J., Asch, D., et al. Assessment of inpatient time allocation among first-year internal medicine residents using time-motion observations. *JAMA Internal Medicine*. **179**, 760-767 (2019)
3. Mamykina, L., Vawdrey, D. & Hripcsak, G. How do residents spend their shift time? A time and motion study with a particular focus on the use of computers. *Academic Medicine: Journal Of The Association Of American Medical Colleges*. **91**, 827 (2016)
4. Clough, R., Sparkes, W., Clough, O., Sykes, J., Steventon, A. & King, K. Transforming healthcare documentation: Harnessing the potential of AI to generate discharge summaries. *BJGP Open*. (2023)
5. Kripalani, S., LeFevre, F., Phillips, C., Williams, M., Basaviah, P. & Baker, D. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *Jama*. **297**, 831-841 (2007)
6. Johnson, A., Pollard, T., Horng, S., Celi, L. & Mark, R. MIMIC-IV-Note: Deidentified free-text clinical notes. (PhysioNet,2023)
7. OpenAI OpenAI. ChatGPT. Accessed September 10, 2023. [Online]. (2022), https://openai.com/blog/chatgpt
8. Bubeck, S., Chandrasekaran, V., Eldan, R., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv Preprint ArXiv:2303.12712*. (2023)
9. OpenAI GPT-4 Technical Report. (2023)
10. Touvron, H., Lavril, T., Izacard, G., et al. Llama: Open and efficient foundation language models. *ArXiv Preprint ArXiv:2302.13971*. (2023)
11. Tay, Y., Dehghani, M., Tran, V., et al. Unifying language learning paradigms. *ArXiv Preprint ArXiv:2205.05131*. (2022)

12. Lehman, E. & Johnson, A. Clinical-t5: Large language models built using mimic clinical text. (PhysioNet,2023)

13. Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is All you Need. *Advances In Neural Information Processing Systems*. **30** (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

14. Kenton, J. & Toutanova, L. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings Of NaacL-HLT*. **1** pp. 2 (2019)

15. Radford, A., Wu, J., Child, R., et al. Language models are unsupervised multitask learners. *OpenAI Blog*. **1**, 9 (2019)

16. Brown, T., Mann, B., Ryder, N., et al. Language Models are Few-Shot Learners. *Advances In Neural Information Processing Systems*. **33** pp. 1877-1901 (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

17. Chowdhery, A., Narang, S., Devlin, J., et al. Palm: Scaling language modeling with pathways. *Journal Of Machine Learning Research*. **24**, 1-113 (2023)

18. Raffel, C., Shazeer, N., Roberts, A., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal Of Machine Learning Research*. **21**, 5485-5551 (2020)

19. Zhang, S., Dong, L., Li, X., et al. Instruction Tuning for Large Language Models: A Survey. (2023)

20. Wang, B., Xie, Q., Pei, J., et al. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*. **56**, 1-52 (2023)

21. Lampinen, A., Dasgupta, I., Chan, S., et al. Can language models learn from explanations in context?. *ArXiv Preprint ArXiv:2204.02329*. (2022)

22. Wei, J., Wang, X., Schuurmans, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances In Neural Information Processing Systems*. **35** pp. 24824-24837 (2022)

23. Hu, E., Shen, Y., Wallis, P., et al. Lora: Low-rank adaptation of large language models. *ArXiv Preprint ArXiv:2106.09685*. (2021)

24. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *ArXiv Preprint ArXiv:2305.14314*. (2023)

25. Ding, N., Qin, Y., Yang, G., et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*. **5**, 220-235 (2023)

26. Papineni, K., Roukos, S., Ward, T. & Zhu, W. Bleu: a method for automatic evaluation of machine translation. *Proceedings Of The 40th Annual Meeting Of The Association For Computational Linguistics*. pp. 311-318 (2002)

27. Lin, C. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*. pp. 74-81 (2004)

28. Zhang, T., Kishore, V., Wu, F., Weinberger, K. & Artzi, Y. Bertscore: Evaluating text generation with bert. *ArXiv Preprint ArXiv:1904.09675*. (2019)

29. Chen, Z., Varma, M., Wan, X., Langlotz, C. & Delbrouck, J. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. *ArXiv Preprint ArXiv:2211.08584*. (2022)

30. Van Veen, D., Van Uden, C., Attias, M., et al. RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models. *ArXiv Preprint ArXiv:2305.01146*. (2023)

31. Van Veen, D., Van Uden, C., Blankemeier, L., et al. Clinical text summarization: adapting large language models can outperform human experts. *ArXiv Preprint ArXiv:2309.07430*. (2023)

32. Lyu, Q., Tan, J., Zapadka, M., et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Visual Computing For Industry, Biomedicine, And Art*. **6**, 9 (2023)

33. Patel, S. & Lam, K. ChatGPT: the future of discharge summaries?. *The Lancet Digital Health*. **5**, e107-e108 (2023)

34. Singh, S., Djalilian, A. & Ali, M. ChatGPT and ophthalmology: exploring its potential with discharge summaries and operative notes. *Seminars In Ophthalmology*. pp. 1-5 (2023)

35. Adams, G., Alsentzer, E., Ketenci, M.,et al. What's in a summary? laying the groundwork for advances in hospital-course summarization. *Proceedings Of The Conference. Association For Computational Linguistics. North American Chapter. Meeting*. 2021 pp. 4794 (2021)

36. Wang, H., Gao, C., Dantona, C., Hull, B. & Sun, J. DRG-LLaMA: Tuning LLaMA Model to Predict Diagnosis-related Group for Hospitalized Patients. *ArXiv Preprint ArXiv:2309.12625*. (2023)

37. Singhal, K., Azizi, S., Tu, T., et al. Large language models encode clinical knowledge. *Nature*. 620, 172-180 (2023)