Clinical Text Summarization with LLM-Based Evaluation

Stanford CS224N Custom Project

Daphne Barretto
daphnegb@stanford.edu

Matthew Jin
mjin73@stanford.edu

Bora Oztekin
boztekin@stanford.edu

Abstract

Clinical text summarization is important for transfer of care, record-keeping, and patient access, but can be time consuming and error-prone. In this work, we investigate the use of GPT-4o and Llama3 for three clinical text summarization tasks. We evaluate the model outputs using four established heuristic methods– `BLEU`, `ROUGE-L`, `BERTScore`, and `MEDCON`–and a novel LLM-based evaluation metric, `LLM-EVAL`, where GPT-4o acts as a "human" evaluator. We find on average, fine-tuned Llama3 often outperforms GPT-4o without fine-tuning in the heuristic methods. However, GPT-4o often outperforms Llama3 in `LLM-EVAL`, which appears to be able to analyze semantic meaning beyond BERT embeddings and pre-specified medical concepts. These results show promise for using both GPT-4o and fine-tuned Llama3 for clinical text summarization over prior older models, as well as shows promise for `LLM-EVAL` and LLM-based evaluation in general for potential use for evaluation clinical text summarization performance. We additionally explore using `LLM-EVAL` for Direct Preference Optimization and provide preliminary findings.

1 Key Information to include

TA mentor: Archit Sharma **External collaborators, External mentor, Sharing project**: No
Team Contributions: Daphne set up the baseline using Azure OpenAI to inference GPT-4o for clinical text summarization. Matthew and Bora fine-tuned Llama3-8B on each of our three datasets using QLoRA. Bora inferenced our fine-tuned Llama3-8B models for each dataset's task. Daphne quantitatively evaluated the inferenced responses using GPT-4o to create `LLM-EVAL`. Matthew used the GPT-4o evaluations as preferences for DPO fine-tuning. Daphne calculated our five evaluation methods for GPT-4o (baseline) and our fine-tuned Llama3-8B models on each dataset's task. All team members supported other techical aspects of the project and prepared the project deliverables.

2 Introduction

Electronic health records have drastically increased the workload of documentation on clinicians, with reports of up to two hours for each patient interaction and 60% of nurses' working time spent on documentation alone. In addition to diverting attention from direct patient care, this clinical documentation workload has shown to directly contribute to stress and burnout in clinicians, leading to worse patient outcomes and decreased job satisfaction. A key part of this workload is clinical text summarization, or summarizing key information from electronic health records, which is immensely important for transfer of care, record-keeping, and patient access, but can be time-consuming and error-prone. Van Veen et al. (2024) claims that prior large language model (LLM) benchmarks for general natural language processing (NLP) tasks do not evaluate performance on relevant clinical tasks. However, prior work has already shown potential for LLM use in clinical NLP tasks through methods including training a new model, fine-tuning an existing model, and supplying task-specific examples in the model prompt. Van Veen et al. (2024) aims to demonstrate adapted LLMs that

perform at least as well as experts in clinical text summarization and to analyze potential medical harm from expert and LLM-generated summaries, alleviating documentation burden and improving patient outcomes.

In this project, we build upon Van Veen et al. (2024) and explore current state-of-the-art large language models, GPT-4o and Llama 3 8B, for three clinical text summarization tasks. We use GPT-4o as a baseline, and we fine-tune Llama 3 to summarize the prompt at the level of a medical expert. In addition to established heuristic metrics–BLEU, ROUGE-L, BERTScore, and MEDCON, we create a novel LLM-based evaluation metric, dubbed `LLM-EVAL`, where GPT-4o acts as a "human" evaluator. We then explore `LLM-EVAL`'s potential for use in direct preference optimization (DPO). Our novel contributions include include clinical text summarization with GPT-4o and fine-tuned Llama 3, an LLM-based evaluation metric, and experimental DPO training with an LLM as the evaluator.

3 Related Work

Our reference paper which our project draws most of its ideas and evaluation methods from is the clinical summarization project Van Veen et al. (2024). It establishes a benchmark of several models, datasets, and metrics to evaluate performance on the task of clinical text summarization. Notably, since Llama 3 is considerably new at the time of writing, it is not included in Van Veen et al. (2024). That said, Llama 3 is a significant improvement over Llama 2, especially with regard to the size of the tokenizer.

One limitation of Van Veen et al. (2024) is that it uses likelihood-based objectives; however, there are new techniques such as Direct Preference Optimization (DPO) Rafailov et al. (2023) that can directly optimize for the performance on metrics like BLEU or ROUGE. Given that Van Veen et al. (2024) uses these metrics, it felt sensible to directly optimize for them. For this, we used GPT-4o to evaluate responses that can be benchmarked with those from the fine-tuned Llama 3. Comparison to an instruction-tuned LLM was an ambitious goal, but felt necessary in the landscape of superior instruct models. While we used Llama 3 pre-trained, we found in experimentation that the instruct model was considerably better at summarization than the base model, which was expected.

DPO Rafailov et al. (2023) represents a significant paradigm shift from using reward functions to measure the favorability of an inference result. It is oftentimes computationally and mathematically challenging to set up the complicated reward function found in traditional reinforcement learning. By optimizing directly for the user's preferences in a mechanism that is based on tuples of different responses, there is a possibility to more directly learn preference, circumventing a construct that is in some ways a proxy.

On the topic of optimizing for preferences, the evaluation techniques of LLMs has also become more robust. While metrics like BLEU Papineni et al. (2002) have been cornerstones of natural language processing for decades, there are possible shortcomings in these metrics that serve as a rigid heuristic for evaluating quality. Namely, G-Eval Liu et al. (2023), given a metric, criteria, and evaluation steps, can evaluate inference quality using LLMs. Large foundation models have a significant and breakthrough ability to generalize, so the trend of using them to evaluate the quality of more task-specific LLMs is likely to be perpetuated.

Fine-tuning LLMs that have become exponentially larger over the years is infeasible for the full set of parameters, which consume copious amounts of memory. When considering the memory also used for gradients, this task becomes even more daunting. An approach to parameter-efficient fine tuning (PEFT) is low-rank adaptation (LoRA) Hu et al. (2021). This consists of using a set of low-rank tensors that are representative of latent factors of the original weights. Only this small representation of the entire set of weights is tuned, so memory for gradients is drastically reduced. QLoRA Dettmers et al. (2023a) further reduces the finetuning memory footprint by quantizing the base model and only storing gradients for the low-rank adapters which are trained in full fp16. The library, bitsandbytes Dettmers is used to facilitate the quantization to 4-bit.

4 Approach

In this project, we investigate the use of LLMs for clinical text summarization tasks using LLM-based evaluation. Specifically, we conduct clinical text summarization using state-of-the-art LLMs, GPT-4o

and fine-tuned Llama3; we introduce `LLM-EVAL`, an LLM-based quantitative metric where GPT-4o is used to evaluate generated summaries; and we experiment with LLM-based DPO, a novel DPO method where an LLM-based evaluation metric, in this case `LLM-EVAL`, is used as a proxy for human feedback for further model fine-tuning.

Baseline: GPT-4o Prior work that predated GPT-4o and Llama3 found that GPT-4 with in-context learning was the best performing model tested on clinical text summarization tasks at the time (Van Veen et al. (2024)). Building upon this work, we use GPT-4o, a newly released version of GPT-4, as our baseline. We prompt GPT-4o with medical professional expertise and task-specific instructions without fine-tuning to generate clinical text summaries. We inference GPT-4o using the Azure OpenAI Service for our experiments.[1] We used a low temperature of 0.1 to reduce hallucinations.

Fine-tuning Llama3-8B with QLoRA Unlike typical causal language modeling, where an autoregressive model uses the prompt as a prior and finds the tokens with the greatest probability of occurring next, summarization tasks are instead meant to condense the prior. Since this is not the default behavior of Llama3 model without instruction-tuning, fine-tuning on datasets that map a long set of text to a corresponding summary was used to train the desired summarization behavior. Details on our preliminary experiment using Llama3-8B without fine-tuning are provided in Section A.4.

To fine-tune Llama3-8B on a commercial 24GB GPU (RTX 4090), we reduced its fine-tuning memory footprint using low rank approximation techniques and quantization by adapting prior work on QLoRA Dettmers et al. (2023b,a). Details on our preliminary experiment for fine-tuning Llama3-8B on the Alpaca Dataset to develop our fine-tuning system are provided in Section A.5.

QLoRA mitigates out-of-memory issues by loading in a quantized version of the model and training a separate matrix called an adapter head on top of the model's existing weight matrices following

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

where W_0 represents the model's original parameters and W represents the weights of the adapter head. In order to save space, the QLoRA approach decomposes W (say, $\mathbb{R}^{d \times d}$) into B ($\mathbb{R}^{d \times r}$) and A ($\mathbb{R}^{r \times d}$) to reduce the space complexity from $O(d^2)$ to $O(d)$. In addition, it ignores the gradient calculation for W_0 as the original model weights remain frozen during fine-tuning. In addition, using 4-bit and 8-bit quantization to load the model further decreased the fine-tuning memory requirements to about 12GB and 18GB respectively.

The tokenization used for fine-tuning is comprised of an input with the instruction, prompt, and summary concatenated. Additionally, a label contains only the summary. The tokenization and label enable summarization rather than generation of new text. In this summarization tasks, generation of new text would be considered a hallucination. We additionally utilized left-padding using the format `<pad><pad>...<begin><instruction>...<prompt>...<begin><label>...<end>`.

The model was quantized to 4-bit with a LoRA alpha and rank of 32 using the bitsandbytes library.[2] In general, LoRA alpha and rank are set to be equivalent, except for tasks where the model's response format is being adapted (e.g., extrapolation to summarization) where a higher rank is recommended. These quantization and low-rank adaptation hyperparameter settings consumed a nontrivial amount of VRAM, but this was somewhat offset by the quantization to 4-bits. Although the traditional recommendation for learning rate is $2 * 10^{-5}$, we were able to obtain a low loss with a learning rate of $2 * 10^{-4}$, although the training loss had high variance as a result. We trained for 50 epochs, allowed the training loss to plateau.

When inferencing fine-tuned Llama3-8B, we used a temperature of 0.1 to avoid hallucinations in sensitive healthcare settings with a low tolerance for errors. However, this often results in duplicate responses when inferencing with the same reference text. The figures in Section A.3 show the fine-tuning training loss and samples per second.

LLM-Based Evaluation Our LLM-based evaluation metric uses GPT-4o prompted for custom task evaluation as described in Table 1. The full prompt text is provided in Table 4. This general structure is adapted from Liu et al. (2023) to be used for clinical text summarization. We used a low

[1] `https://azure.microsoft.com/en-us/products/ai-services/openai-service/`

[2] `https://pypi.org/project/bitsandbytes/`

temperature of 0.1 to reduce hallucinations and help retain the proper reference text and inferences provided in the prompt during evaluation.

The result of our prompting is three sub-metrics each scaled from $[-10, +10]$ for completeness, correctness, and conciseness. Larger values indicate a better score for each sub-metric. These values mirror previous clinical reader studies used to evaluate LLM-generated clinical text summarization tasks (Van Veen et al. (2024)). To calculate `LLM-EVAL`, we sum the three sub-metrics into a single value ranging from $[-30, +30]$ and scale the value to range from $[0, 100]$ to align with the other evaluation metrics used.

Evaluation Introduction	Describes the evaluation task for the LLM to conduct.
Task-Specific Instructions	Describes the summarization task for the associated dataset.
Evaluation Criteria	Describes the completeness, correctness, and conciseness metrics and how they are used for evaluation.
Evaluation Steps	Describes the steps to evaluate the inferences, generate quantitative scores based on the evaluation criteria, and format the evaluation.
Reference Text	The original reference text to be summarized.
Inferences	A list of generated summaries as inferenced from the model.

Table 1: Prompt section descriptions used by an LLM to produce `LLM-EVAL` metrics.

LLM-Based Direct Preference Optimization We additionally experimented with Direct Preference Optimization using LLM-based evaluation instead of requiring resource-intensive human evaluators, and investigated whether LLM-based DPO could improve clinical text summarization results. DPO requires a “chosen" and a “rejected" sample to learn what it should and should not output for a given prompt respectively.

To generate the “chosen" and “rejected" samples for the DPO dataset using LLMs, we prompted GPT-4o to evaluate four responses generated from the fine-tuned Llama3 model responding to the same prompt. We used the process outlined in the previous section to calculate `LLM-EVAL` scores. The highest scored response is labeled as the “chosen" sample and the lowest scored response is labeled as the “rejected" sample for the corresponding prompt.

To perform preliminary DPO fine-tuning, we generated three DPO datasets based on the three clinical text summarization datasets in our experiments using a subset (10%) of the original test dataset. These datasets were then passed to HuggingFace’s DPO trainer to train on top of their corresponding QLoRA fine-tuned Llama3-8B models. Completion of LLM-based DPO requires additional investigation into the tokenization process.

5 Experiments

We conduct three experiments using distinct clinical text summarization datasets and tasks to compare GPT-4o and QLoRA fine-tuned Llama3-8B and evaluate our `LLM-EVAL` metric.

Data We use three clinical text summarization datasets and tasks from Van Veen et al. (2024): MeQSum (`chq`), Open-i Radiology (`opi`), and Dialogue to Note (`d2n`). `chq` contains 1,200 commonly asked questions by patients, often written in a conversational tone, with the goal of reducing these into neutral single-sentence questions. `opi` contains 3,418 radiology reports detailing the findings of the doctor, with the goal of reducing these findings into a single observation. `d2n` contains 126 machine transcribed conversations between doctor and patient, with the goal of reducing the dialogue into a summarized assessment and plan. All datasets contain sample outputs. Example input-output pairs are provided in Section A.1.

Evaluation method We use five quantitative metrics for evaluation: `BLEU`, `ROUGE-L`, `BERTScore`, `MEDCON`, and `LLM-EVAL`. `BLEU` from Papineni et al. (2002) is a syntactic measurement, calculating the overlap between reference and generated text from 1-grams to 4-grams. `ROUGE-L` from Lin (2004) is a syntactic measurement, calculating syntactic similarity through the longest common subsequence. `BERTScore` from Zhang et al. (2020) is a semantic measurement, calculating semantic similarity via contextual BERT embeddings. `MEDCON` from Yim et al. (2023) is a medical concept measurement,

calculating semantic similarity of used medical concepts. LLM-EVAL is a novel summarization-specific quality measurement, calculating the completeness, correctness, and conciseness of the summary when compared to the reference text to be summarized. All five metrics range from $[0, 100]$. For BLEU, ROUGE-L, BERTScore, and MEDCON, higher scores indicate greater similarity between generated and example output texts. For LLM-EVAL, higher scores indicate greater completeness, correctness, and conciseness compared to the reference text.

We calculate BLEU, ROUGE-L, and BERTScore using the Evaluate library.[3] We calculate MEDCON using the QuickUMLS system (Soldaini (2016); Okazaki and Tsujii (2010)), which requires a Unified Medical Language System license. We calculate LLM-EVAL using our novel LLM-based evaluation methods that prompt GPT-4o to quantitatively evaluate our metric as described in Section 4.

Experimental details We inferenced GPT-4o (our baseline) and fine-tuned QLoRA Llama3-8B for our three clinical text summarization datasets and tasks, resulting in six sets of inferences. For chq and opi, we conduct approximately 250 inferences each. For d2n, which contains larger texts, we conduct approximately 100 inferences each. We then calculated our five evaluation metrics, BLEU, ROUGE-L, BERTScore, MEDCON, and LLM-EVAL for each set of inferences. Finally, we conducted qualitative evaluation on the inferenced responses.

Results Table 2 provides the resulting evaluation metrics for the chq, opi, and d2n experiments. Notably, fine-tuned Llama3-8B performed better on average than GPT-4o in all metrics except for LLM-EVAL for opi and d2n. This is expected, since fine-tuned Llama3-8B was additionally trained on the two datasets, allowing it to better summarize the radiology reports and transcribed doctor/patient conversations. Fine-tuned Llama3-8B performed similarly on average to GPT-4o in all metrics except for LLM-EVAL for chq. We expected fine-tuned Llama3-8B to perform better, but the similar performance could instead indicate that fine-tuning was not as necessary for this task (such that an LLM that was not fine-tuned can sufficiently summarize patient questions). This likely does not indicate that fine-tuning was less effective on this task's dataset, since our measured loss reduced over training and Llama3-8B without fine-tuning performed worse.

For all experiments, GPT-4o performed better on average than fine-tuned Llama3-8B for LLM-EVAL. This was unexpected, because in most cases for all other metrics, fine-tuned Llama3-8B performed better. Unlike the other four metrics, LLM-EVAL takes into account both syntactic and semantic aspects of the summarization during evaluation through completeness, correctness, and conciseness properties. Qualitatively, we find Llama3-8B summaries to be syntactically more similar to the reference text, aligned with the general BLEU and ROUGE-L metrics. However, we find GPT-4o summaries to be more semantically similar than implied by their BERTScore and MEDCON metrics at times. In particular, there are cases where GPT-4o summaries contain the semantic meaning of the reference text, but are written in a way that is substantially different from the syntax and structure used in the reference text. Using colloquially shortened text can potentially be missed by BERTScore and MEDCON, and LLM-EVAL can potentially be picking up on the semantic meanings in the summaries.

We note that there are large standard deviation values across all metrics except BERTScore, which is expected since there is a large variety in the test sets such that different test values could have substantially different scores. LLM-EVAL provides the largest standard deviations, which is not expected since GPT-4o was not prompted to scale the scores in a particular distribution or relative to particular scoring examples.

We additionally compare our results to the metrics for GPT-4 from Van Veen et al. (2024). We find that fine-tuned Llama3-8B outperforms GPT-4 on its own but that GPT-4 with sufficient in-context learning examples still outperforms fine-tuned Llama3-8B.

6 Analysis

Overall, we find that fine-tuned Llama3-8B performed better on average than GPT-4o in most metrics except for LLM-EVAL, our proposed LLM-based evaluation metric. However, from our qualitative results, we find that LLM-EVAL might be identifying semantic similarities missed by BERTScore and MEDCON. This could potentially explain why GPT-4o summaries have higher LLM-EVAL scores, but generally lower scores in all other metrics. Additionally, it's possible that GPT-4o could be biased

[3] https://pypi.org/project/evaluate/

chq

	BLEU	ROUGE-L	BERTScore	MEDCON	LLM-EVAL
GPT-4o (Baseline)	3.3 ± 10.3	$\mathbf{32.5} \pm 17.1$	$\mathbf{91.4} \pm 2.5$	47.8 ± 32.6	$\mathbf{97.7} \pm 46.0$
Fine-tuned Llama3-8B	$\mathbf{4.9} \pm 14.8$	31.3 ± 19.4	91.0 ± 3.3	$\mathbf{48.0} \pm 40.3$	77.3 ± 33.3

opi

	BLEU	ROUGE-L	BERTScore	MEDCON	LLM-EVAL
GPT-4o (Baseline)	1.4 ± 4.9	15.4 ± 16.0	86.7 ± 2.8	13.4 ± 22.4	$\mathbf{95.0} \pm 62.3$
Fine-tuned Llama3-8B	$\mathbf{12.3} \pm 30.5$	$\mathbf{41.2} \pm 33.7$	$\mathbf{90.0} \pm 5.3$	$\mathbf{25.6} \pm 41.0$	58.0 ± 49.7

d2n

	BLEU	ROUGE-L	BERTScore	MEDCON	LLM-EVAL
GPT-4o (Baseline)	5.8 ± 3.4	21.9 ± 5.3	85.1 ± 1.1	36.8 ± 12.8	$\mathbf{84.3} \pm 30.3$
Fine-tuned Llama3-8B	$\mathbf{18.2} \pm 14.0$	$\mathbf{30.8} \pm 13.7$	$\mathbf{88.3} \pm 2.5$	$\mathbf{48.5} \pm 18.1$	71.7 ± 19.3

Table 2: Average and standard deviation of scores for GPT-4o and Fine-tuned Llama3-8B across BLEU, ROUGE-L, BERTScore, MEDCON, and LLM-EVAL metrics for the chq, opi, and d2n tasks.

when evaluating text generated by GPT-4o itself, due to the generated text being more probable in its own language model word-generation distribution. By comparing the other quantitative metrics and qualitative evaluation with LLM-EVAL, it appears that LLM-EVAL shows potential for measuring clinical text summarization success beyond typical heuristic evaluations.

Similarly to the approach taken in the clinical summarization paper, we used a low temperature given the domain of healthcare and the low tolerance for hallucinations. This frequently resulted in outputs that were very similar or even repeated which impacts metrics such as the BLEU score as there is less opportunity to try different words that will be similar to the reference text.

In addition, due to the custom method the original paper used to tokenize their inputs, some examples in the test dataset were discarded due to exceptions that were presumably caused by some tokenization issue. The discarded examples were not manually selected, but were jumped over by the for-loop that processed the inferences.

Hardware limitations also made fine-tuning the d2n dataset difficult since it contained prompts that sometimes had thousands of words. Whenever, a prompt exceeded the max pad length specified during our fine-tuning process (1024), we simply truncated the prompt to fit which would have impacted our final performance. Nonetheless, Table 2 shows that fine-tuning Llama3 on the truncated inputs still produced reasonable outputs.

Given the specific nature of each task, we realize that DPO might not be necessary for these clinical summarization tasks. In most cases, when the model was considered incorrect, the generated text was more harmless than malicious, and it is obvious to the user that the model had some issues summarizing the prompt.

Another issue with DPO was that the low temperature used for generation led to similar responses. On occasion, Llama3 would generate the same response when queried multiple times which impacted our process for choosing the chosen and rejected prompts for the DPO dataset as GPT-4o's ranking of the responses would be meaningless. As a result, the generated DPO datasets sometimes contained chosen and rejected prompts that were exactly the same. We did consider raising the temperature to generate a wider variety of responses, but that often resulted in less accurate generations overall.

7 Conclusion

In this work, we investigate the use of GPT-4o and Llama3 for three clinical text summarization tasks. We evaluate the model outputs using four established heuristic methods–BLEU, ROUGE-L, BERTScore, and MEDCON–and a novel LLM-based evaluation metric, LLM-EVAL, where GPT-4o acts as a "human" evaluator. We find on average, fine-tuned Llama3 often outperforms GPT-4o without fine-tuning in the heuristic methods. However, GPT-4o often outperforms Llama3 in LLM-EVAL, which appears to be able to analyze semantic meaning beyond BERT embeddings and pre-specified medical concepts. These results show promise for using both GPT-4o and fine-tuned Llama3 for clinical text summarization over prior older models, as well as shows promise for LLM-EVAL and LLM-

based evaluation in general for potential use for evaluation clinical text summarization performance. We additionally explore using LLM-EVAL for Direct Preference Optimization and provide preliminary findings.

In general, we learned that clinical text summarization is a deviation from the typical task of language modeling. In these types of tasks, fine-tuned models for domain-specific applications often perform better than instruction-tuned models that can potentially conduct summarization, and instruction-tuned models often perform better than pretrained models without further instruction-tuning or fine-tuning. In most cases, there is a trade-off between breadth and depth, and instruction-tuned models are able to handle a variety of queries, but the fine-tuned model qualitatively better handled domain-specific medical terminology.

The primary limitations in this work include data and computing resources. Medical datasets are very hard to obtain due too strict privacy regulations, and for deep learning applications, datasets with sizes in the 1000s was likely not sufficient to generalize. While we did see good results, it is important to note that the distribution of language and context used is not very wide, and it is possible that inferencing prompts of a different format or from a different clinic would yield undesirable results. With respect to compute, due to using half of the VRAM used by the original paper, we had to use a quantized model with low batch sizes which increased training time and may have impacted the precision of the resulting generations.

Besides experimenting with further improving LLM models adapted for clinical text summarization, potentially by combining our state-of-the-art models with in-context learning examples from Van Veen et al. (2024), future work should systematically experiment with versions of LLM-EVAL with various prompting details and evaluate these results again human medical professional preferences, and future work should build on our preliminary investigation of LLM-based DPO to determine potential quantitative benefits over other methods by conducting more robust DPO fine-tuning and comparing LLM-based DPO Llama3-8B existing results on chq, opi, and d2n with QLoRA fine-tuned Llama3-8B. To more efficiently do these experiments, methods newer than QLoRA, such as GaLore which does not use extra memory for the gradients themselves allowing for more weights or larger batch sizes in the training process, could be utilized. Finally, in addition to DPO, RPO can be explored to consider the feedback on the responses to similar inquiries, rather than only those in the dataset. This can make a significant difference in the generation quality for inputs that are very out-of-distribution compared to the training datasets.

8 Ethics Statement

In general, the major ethical challenges and possible societal risks of using LLMs for clinical text summarization are concerns about (1) inferior summaries causing worse patient outcomes, (2) a lack of transparency in model decision-making, (3) data privacy for electronic health records. For LLM-based DPO, a specific ethical challenge based on these general challenges is ensuring that DPO fine-tuning on LLM evaluations does not result in inferior summaries compared to DPO fine-tuning on medical expert evaluations. To consider models developed using LLM-based DPO sufficient for clinical text summarization, safety checks and regulations must be put into place to ensure the quality of the summaries are at least equivalent to those of human medical experts. We attempt to mitigate this concern by doing thorough quantitative and qualitative evaluations (via LLM-evaluation) on our model outputs, and especially analyze any concerns for hallucination or understanding ambiguity. Further mitigation should involve medical experts in the field before approval for use.

A second ethical challenge is concerns about patient health information privacy. Compliance with HIPAA and other data handling standards is paramount, and there are additional measures and costs to ensure the proper anonymization and isolation of patient data. Even with the best efforts, it is possible for information to be left over in the datasets. For example, if the live chat session in the `d2n` dataset contained a patient's name, it is possible that it would not be filtered out, especially if it were in lowercase or was also a common noun. To mitigate these concerns, all datasets used to develop LLMs for clinical text summarization must follow proper data handling standards to prevent LLMs from even accessing improper data. Additionally, proactive work should be done to test and ensure that such LLMs do not hallucinate incorrect patient health information from filtered content to prevent LLMs from appearing to have patient health information even when they do not.

References

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Tim Dettmers. bitsandbytes. `https://github.com/TimDettmers/bitsandbytes`.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023a. Qlora: Efficient finetuning of quantized llms. `https://github.com/artidoro/qlora`.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023b. Qlora: Efficient finetuning of quantized llms.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, Beijing, China. Coling 2010 Organizing Committee.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Luca Soldaini. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A Appendix

A.1 Dataset Details and Example Input-Output Pairs

Dataset	Example Input	Example Output
Open-i Radiology: 3,418 radiology reports detailing the findings of the doctor. The goal of the dataset is to reduce those findings into a single observation. No planned preprocessing.	"Heart size is enlarged, pulmonary vascularity within normal limits. Cardiac defibrillator generator projects over the left mid lateral lung. No visible pneumothorax or pleural effusion. Minimal streaky airspace opacities in the lower lobes."	"1. Cardiomegaly without pulmonary edema. 2. Minimal bibasilar opacities which may reflect atelectasis or infiltrate."
Dialogue to Note: 126 machine transcribed conversations between doctor and patient. The goal of the dataset is to condense the dialogue into a summarized assessment. Preprocessing to only consider the 126 samples containing 'assessment and plan' sections, similar to Van Veen et al. (2024).	"[doctor] so bryan it's nice to see you again in the office today what's going on [patient] i was in my yard yesterday and i was raking leaves and i felt fine and then when i got into my house about two hours later my back started tightening up and i started getting pins and needles in my right foot ..." [truncated for space]	"ASSESSMENT \n \n Low back sprain. \n \n PLAN \n \n The examination findings and x-ray results were discussed with the patient and his partner today. I recommend we treat this conservatively with rest, meloxicam, and formal physical therapy. If he fails to improve, we can consider obtaining an MRI for further evaluation."
MeQSum: 1,200 commonly asked questions by patients, often written in a conversational tone. The goal of the dataset is to reduce these inputs to neutral single-sentence questions. No planned preprocessing.	"SUBJECT: Friedreich's ataxia MESSAGE: I have been told I have Friedreich's ataxia. I am looking for a treatment to reverse it. I has to do with the chromosome number 9 defective in both of my parents and possible damage to chromosome number 10. Would you be able to tell me if number 9 is defective in me."	"Where can I get genetic testing for friedreich's, and what are the treatments for it?"

Table 3: Dataset descriptions and example input-output pairs.

A.2 LLM-EVAL Prompt Text

Evaluation Introduction	You are quantitatively evaluating scores for multiple responses to the same reference text. We will provide you a Task Introduction, which is the description used to describe the desired response to the reference text; Evaluation Criteria, which describes how you should evaluate each response; Evaluation Steps, which describes how to score each response; Input Context, which is the reference text; and Input Targets, which are the responses you will be evaluating.
Task-Specific Instructions	(chq) Task Introduction: Summarize the patient health query into one question of 15 words or less. The reference text is the patient health query. Note that the response should summarize the query into one question, and it should not answer the query. (opi) Task Introduction: Summarize the radiology report findings into an impression with minimal text. The reference text is the radiology report findings. (d2n) Task Introduction: Summarize the patient/doctor dialogue into an assessment and plan. The reference text is the patient/doctor dialogue.
Evaluation Criteria	Evaluation Criteria: There are three evaluation criteria: completeness, correctness, and conciseness. Completeness evaluates which response most completely captures important information. Correctness evaluates which response includes the lease false information. Conciseness evaluates which response contains the least non-important information.
Evaluation Steps	Score each response as an integer from -10 to +10 for each of the three evaluation criteria: completeness, correctness, and conciseness. A score of -10 indicates the response is the worst possible response for that criterion, a score of 0 indicates the response is just acceptable, and a score of +10 indicates the response is the best possible response for that criterion. Each response should have three scores, one for each criterion. Return your evaluation in the exact format "Reference Text: reference_text, (inference_text_1, (completeness_value_1, correctness_value_1, conciseness_value_1)), (inference_text_2, (completeness_value_2, correctness_value_2, conciseness_value_2)), (inference_text_3, (completeness_value_3, correctness_value_3, conciseness_value_3)), (inference_text_4, (completeness_value_4, correctness_value_4, conciseness_value_4))" Replace inference_text_1, completeness_value_1, correctness_value_1, conciseness_value_1, inference_text_2, completeness_value_2, correctness_value_2, conciseness_value_2, inference_text_3, completeness_value_3, correctness_value_3, conciseness_value_3, inference_text_4, completeness_value_4, correctness_value_4, conciseness_value_4 with the actual values. Otherwise, use the exact formatting provided. Do not add newlines, tabs, or other unnecessary formatting characters.
Reference Text	[The original reference text to be summarized.]
Inferences	[A list of generated summaries as inferenced from the model.]

Table 4: Prompt section text used by an LLM to produce LLM-EVAL metrics.

A.3 Fine-tuning Training Loss and Samples per Second

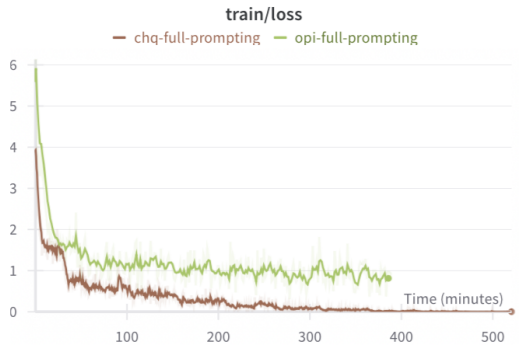

Figure 1: Training loss for `chq` and `opi`

`d2n` was trained using a separate process that did not allow for easy logging on Weights&Biases Biewald (2020). In addition, the small dataset size means that 50 epochs involves only a couple hundred optimization steps.

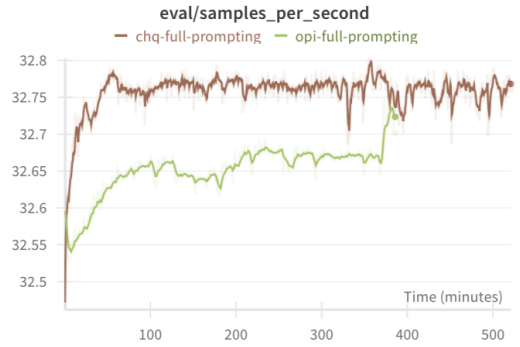

Figure 2: Evaluation samples per second for `chq` and `opi`

`opi` had a lower sampling rate due to generally larger prompt sizes.

A.4 Additional Experiment: Llama3 without Fine-Tuning

We conducted preliminary experiments on Llama3-8B without fine-tuning. We found that it produced summaries with low scoring metrics and qualitatively bad evaluations. However, adding instructions or fine-tuning with instructions (such as Alpaca Taori et al. (2023)) helped produce better summaries by both quantitative and qualitative metrics.

These preliminary results are expected, because the pretrained large language model's is trained to generate the next probable token, and in most text, if someone is asking a question, writing a report, or providing a dialogue, it is usually not followed by a summary. Using the prior tokens themselves, rather than the tokens that are likely to follow, as the distribution of data for generation, is unique to summarization. Furthermore, general pretrained LLMs are not necessarily accustomed to instruction. The instruction followed by the prompt, with the Llama3-8B model without fine-tuning resulted in seemingly random generation of text until the maximum length was reached.

A.5 Additional Experiment: Fine-tuning Llama3-8B on the Alpaca Dataset

We conducted a preliminary experiment on the Alpaca dataset (Taori et al. (2023)) to help establish our system for fine-tuning Llama3-8B. We fine-tuned both 4-bit and 8-bit quantized model versions of Llama3-8B on the Alpaca dataset for 10,000 iterations, taking around 12 hours and 16 hours respectively. For both 4-bit and 8-bit runs, we use a learning rate of 0.0002, a LoRA dimension of 64, and Adam parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The hardware used for training is an RTX 4090.