Title

Closing the gap between open source and commercial large language models for medical evidence summarization.

Journal

npj Digital Medicine, 7(1)

Authors

Zhang, Gongbo

Jin, Qiao

Zhou, Yiliang

et al.

Publication Date

2024-09-09

DOI

10.1038/s41746-024-01239-w

Peer reviewed

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01239-w

Closing the gap between open source and commercial large language models for medical evidence summarization

Gongbo Zhang[1], Qiao Jin[2], Yiliang Zhou[3], Song Wang[4], Betina Idnay[1], Yiming Luo[5], Elizabeth Park[5], Jordan G. Nestor[5], Matthew E. Spotnitz[6], Ali Soroush[7,8,9], Thomas R. Campion Jr.[3,10], Zhiyong Lu[2], Chunhua Weng[1] ✉ & Yifan Peng[3,10] ✉

Large language models (LLMs) hold great promise in summarizing medical evidence. Most recent studies focus on the application of proprietary LLMs. Using proprietary LLMs introduces multiple risk factors, including a lack of transparency and vendor dependency. While open-source LLMs allow better transparency and customization, their performance falls short compared to the proprietary ones. In this study, we investigated to what extent fine-tuning open-source LLMs can further improve their performance. Utilizing a benchmark dataset, MedReview, consisting of 8161 pairs of systematic reviews and summaries, we fine-tuned three broadly-used, open-sourced LLMs, namely PRIMERA, LongT5, and Llama-2. Overall, the performance of open-source models was all improved after fine-tuning. The performance of fine-tuned LongT5 is close to GPT-3.5 with zero-shot settings. Furthermore, smaller fine-tuned models sometimes even demonstrated superior performance compared to larger zero-shot models. The above trends of improvement were manifested in both a human evaluation and a larger-scale GPT4-simulated evaluation.

Medical evidence plays a critical role in healthcare decision-making. In particular, systematic reviews and meta-analyses of randomized controlled trials (RCTs) are considered the gold standard for generating robust medical evidence[1,2]. However, systematically reviewing multiple RCTs is laborious and time-consuming[3]. It requires retrieving relevant studies, appraising the evidence quality, and synthesizing findings. Meanwhile, systematic reviews frequently become obsolete upon publication, primarily due to protracted review processes. The delay is exacerbated by the exponential increase in scientific discoveries, exemplified by over 133,000 new clinical trials registered at ClinicalTrials.gov since 2020[4]. As such, it is imperative to establish an efficient, reliable, and scalable automated system to streamline and accelerate systematic reviews.

Typically, systematic reviews include quantitative and qualitative reports[5], the former being a statistical meta-analysis of relevant clinical trials, and the latter being a concise narrative explanation of the quantitative results[6,7]. Language generation technologies could potentially be employed to auto-generate such narratives, yet they have not been widely applied for medical evidence summarization[7]. Text summarization has attracted research attention for decades. Earlier techniques relied on extracting key phrases and sentences through rules or statistical heuristics like word frequency and sentence placement[8–12]. These methods, however, struggled with comprehending context and generating cohesive summaries. A significant shift occurred with the adoption of neural network-based methods, enhanced by attention mechanisms[13–19]. These mechanisms enable the model to concentrate on various input segments and understand long-range connections between text elements. This advancement allows for a deeper grasp of context, leading to smoother and more precise summaries.

Recent advancements in generative Artificial Intelligence, notably large language models (LLMs), have shown tremendous potential in comprehending and generating natural language[1,20]. While these generalist models perform well across diverse tasks, they fail to capture in-depth domain-specific knowledge, particularly in biomedicine[21,22]. Furthermore, despite

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA. [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [3]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [4]Cockrell School of Engineering, The University of Texas at Austin, Austin, TX, USA. [5]Department of Medicine, Columbia University, New York, NY, USA. [6]Office of the Director, National Institutes of Health, Bethesda, MD, USA. [7]Division of Data-Driven and Digital Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [8]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [9]Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [10]Clinical & Translational Science Center, Weill Cornell Medicine, New York, NY, USA.
✉e-mail: cw2384@cumc.columbia.edu; yip4002@med.cornell.edu

the relevant superior performances compared to open-source alternatives[7,23–26], the disadvantages of closed-source models were rarely discussed or even mentioned[27,28]. The lack of transparency of closed-source models makes it challenging to understand the model behavior and troubleshoot customized variants. Moreover, reliance on closed-source models raises the risks associated with changes in service terms or discontinuation of services, which pose a critical threat to long-term projects. Open-source models provide a promising solution to mitigate the above risks.

While multiple open-source model architectures were proposed for either general-purpose foundation models or specifically for text summarization, few have been optimized to synthesize medical evidence. Gutierrez et al. compared the in-context learning performance of GPT-3 with fine-tuned BERT models on named-entity recognition tasks in the biomedical domain[29]. Tang et al. assessed zero-shot GPT-3.5 and ChatGPT on summarizing Cochrane review abstracts[7]. The above-related studies, however, did not focus on fine-tuning open-source models for medical evidence summarization. It's also unclear to what degree the optimization strategies, e.g., few-shot learning or fine-tuning, can help bridge the performance gap between open-source models and cutting-edge closed-source alternatives. To quantitively assess fine-tuning technologies to enhance open-source LLMs for medical evidence summarization, we experimented with three broadly used open-sourced LLMs: PRIMERA[15], LongT5[14], and Llama-2[25], including both architectures designed specifically for text summarization and architectures of generalist foundation models. Fine-tuning these models is challenging due to the substantial requirement for computational resources and the risk of catastrophic forgetting, a phenomenon in machine learning where the performance degrades on tasks where the LLM initially performed well[30]. To address these issues, we employed low-rank adaptation (LoRA)[31], which is a parameter-efficient fine-tuning method focusing on updating only a minimal amount of model parameters during the fine-tuning process.

To facilitate future studies on leveraging LLMs for medical evidence summarization, we present a benchmark dataset, MedReview, consisting of 8161 pairs of meta-analysis results and narrative summaries from the Cochrane Library[32], published on 37 topics between April 1996 and June 2023. The dataset consists of training, validation, and test sets (see details in the "Method" section and Supplementary Table 1). This collection is an extension of our previous study of evaluating LLMs for evidence summarization[7] and covers a wider range of specialties and writing styles, which highlight common text summarization challenges (Fig. 1a).

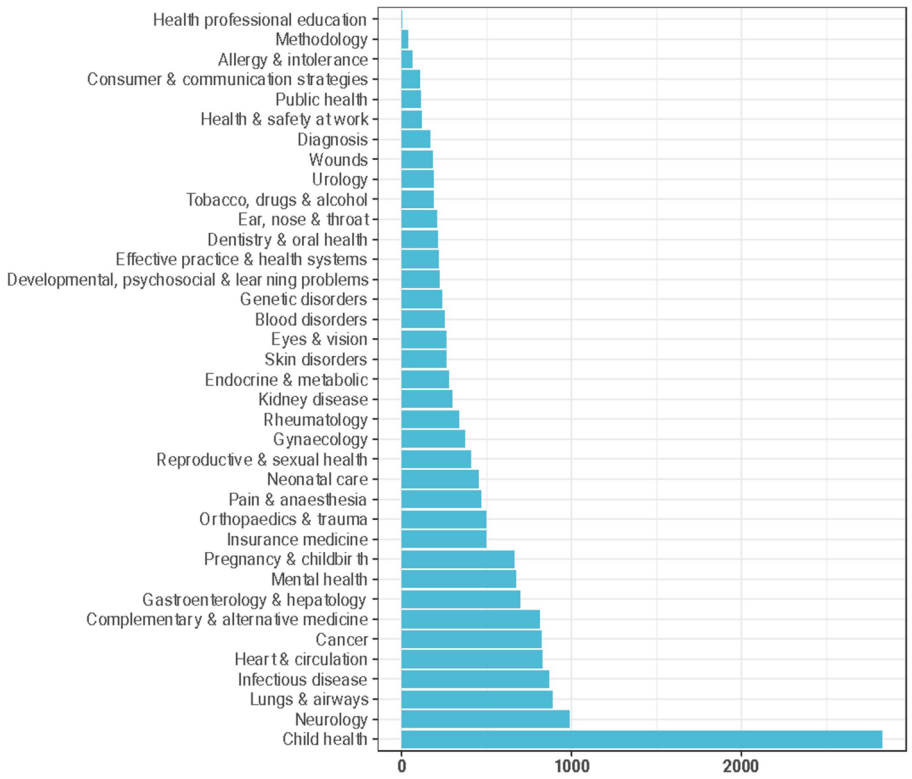
Results

Comparison of different LLMs in automatic evaluations

First, we fine-tuned PRIMERA, LongT5, and Llama-2 using the LoRA method (Fig. 1b). We observed that fine-tuning considerably improved the performance of most models ($p < 0.01$, Fig. 2). Specifically, LongT5 models benefited the most from fine-tuning, which led to an increase from 14.72 to 24.61, 15.06 to 28.27 in METEOR, 15.15 to 38.81 in CHRF, and 36.05 to

Fig. 1 | Overview of topic distribution of the MedReview dataset and LLMs in this study. **a** Topic distribution of the MedReview dataset. **b** Choice of LLMs in this study.

a



b

Model	Developer	Context Window Size	Open Source	Summarization Task Specific	# of Trainable Params (LoRA r=8)	Total # of Params
PRIMERA	AI2	4,096	Y	Y	1.18M (0.26%)	448.40M
LongT5-base	Google	16K	Y	Y	3.54M (1.41%)	251.13M
LongT5-xl	Google	16K	Y	Y	4.72M (0.17%)	2.85B
Llama-2	Meta	4,096	Y	N	16.38M (0.02%)	69.0B
GPT 3.5	OpenAI	≥ 4K	N	N	N/A	175B

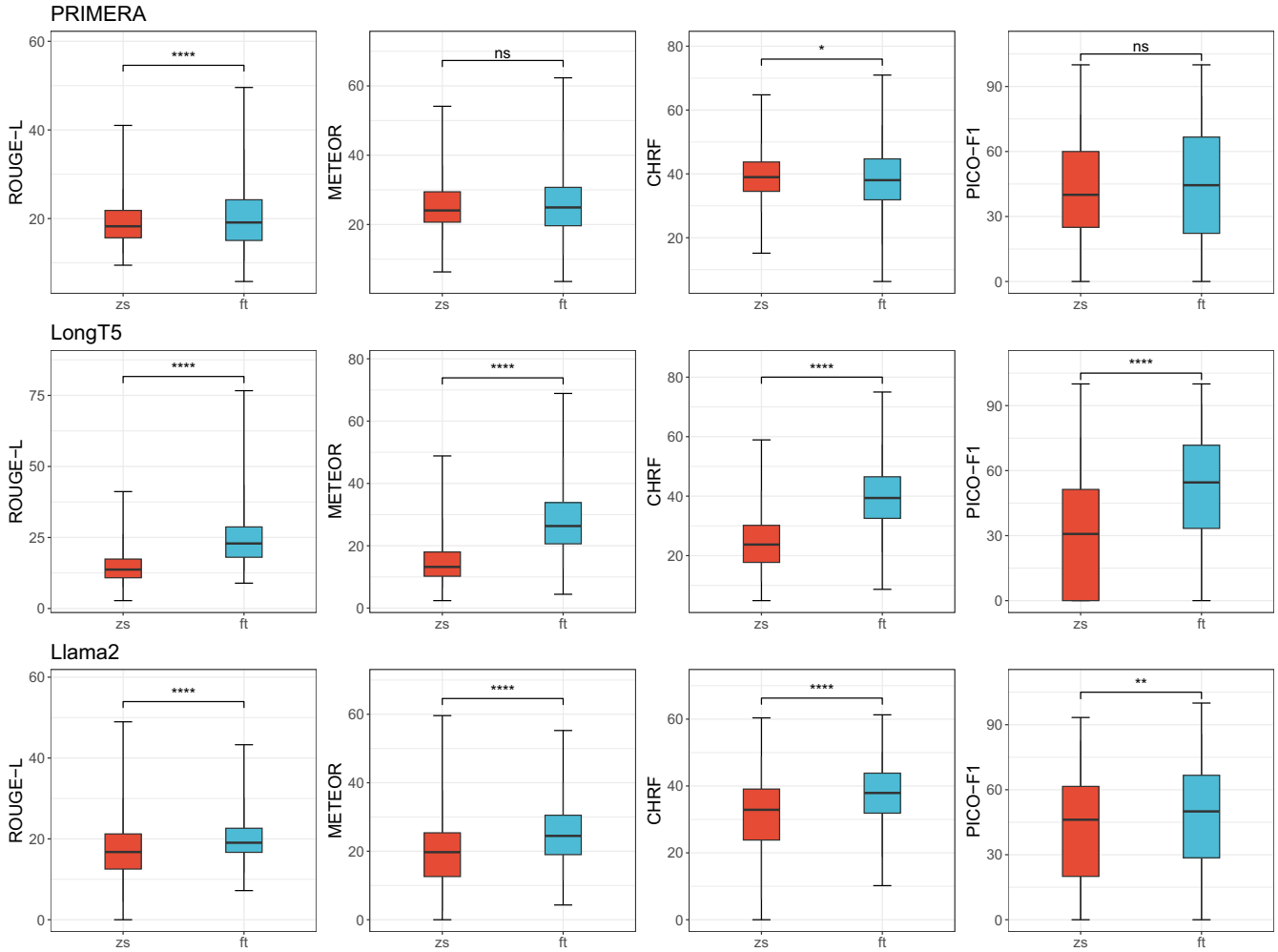


Fig. 2 | Performance of different medical evidence summarization systems in automatic evaluations. The p-value was calculated using a paired t-test to determine the statistical significance of the difference between the models. FT fine-tuning, ZS zero-shot learning, ns not significant; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; $^{****}p < 0.0001$.

51.43 in PICO-F1 (Supplementary Table 2). In contrast, PRIMERA demonstrated a relatively moderate improvement with a ROUGE-L increase from 18.90 to 20.48, METEOR increase from 25.15 to 26.50, and PICO-F1 increase from 43.22 to 49.47. However, there was a slight decrease in the CHRF from 39.25 to 37.84. Overall, the fine-tuned LLMs improved the ROUGE-L score by an absolute of 9.89% (95% confidence interval of improvement: 8.94–10.81), the METEOR score by 13.21 (95% confidence interval of improvement: 12.05–14.37), and the CHRF score by 15.82 (95% confidence interval of improvement: 13.89–16.44). These recently released models all outperformed the fine-tuned variant of BART, the previous SoTA. The fine-tuned BART achieved 17.74, 27.49, and 40.54 in ROUGE-L, METEOR, and CHRF, respectively. We compared the fine-tuned models with GPT-3.5-turbo, one of the most widely known and cutting-edge closed-source LLMs. Zero-shot GPT-3.5-turbo achieved 23.15 in ROUGE-L, 28.83 in METEOR, and 39.74 in CHRF scores. The performance gaps between the open-source models and GPT-3.5-turbo were reduced after fine-tuning. The fine-tuned LongT5 achieved similar results as GPT-3.5-turbo (Supplementary Table 2). We also conducted a pilot study using GPT-4. The summaries generated by GPT-3.5-turbo and GPT-4 are not significantly different; as such, we only use GPT-3.5-turbo for comparison.

Furthermore, while few-shot learning helped reduce the performance gap, LLMs fine-tuned with the entire training data still demonstrated better performance. We constructed two sets of few-shot learning baselines, one based on few-shot prompting and the other based on few-shot fine-tuning (see details in "Methods"). Under 1-, 2-, and 5-shot prompting, Mixtral-8x7B achieved 24.87/24.53/24.99 in Rouge-L, 27.82/25.78/27.59 in METEOR, 37.63/35.61/37.42 in CHRF, respectively. After being fine-tuned using 100 randomly selected samples, PRIMERA achieved 19.11 in ROUGE-L, 24.98 in METEOR, and 39.01 in CHRF; LongT5 was moderately improved by few-shot fine-tuning, resulting in 15.06 in ROUGE-L, 16.17 in METEOR, and 24.81 in CHRF.

We also calculated Pearson correlation coefficients among different automatic metrics using the entire test set (Supplementary Fig. 2). Most metrics are strongly positively correlated with each other; with a few exceptions, most coefficients are larger than 0.7. The only exceptional metric is the PICO coverage, which is only weakly positively correlated to the others, with coefficients ranging between 0.33 and 0.46. As discussed above, zero-shot models tend to verbatim replicate the input, which is already abundant in PICO concepts, since we selected only the objective and main results section of the review abstracts.

Comparison between zero-shot LongT5-xl and fine-tuned longT5-base

Next, we investigated whether fine-tuned smaller models have the capability to outperform zero-shot larger models. To this end, we fine-tuned the LongT5-base model, which has 10% fewer parameters than LongT5-xl. Figure 3 shows that fine-tuned LongT5-base outperforms zero-shot LongT5-xl. We also found that this observed trend holds across different LLM architectures. For instance, the performance of fine-tuned PRIMERA and LongT5 exceeded that of zero-shot Llama-2 (Supplementary Table 2), even though the latter model comprises at least 20 times more parameters.

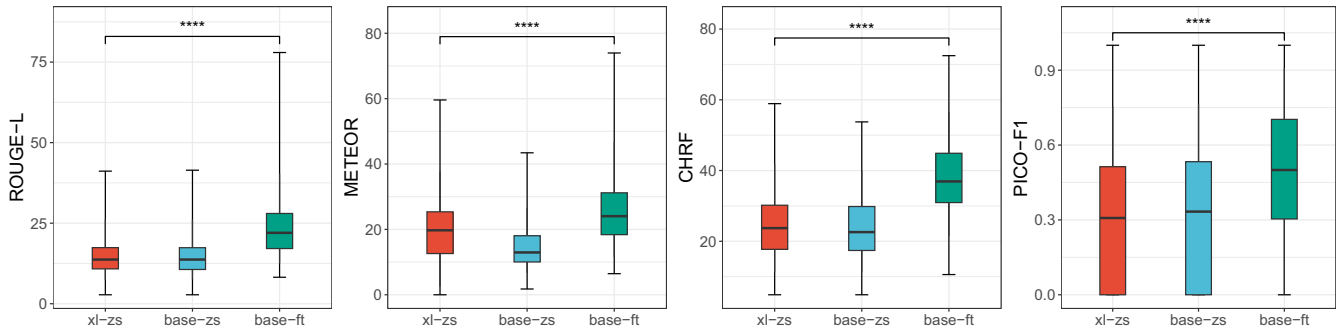


Fig. 3 | Comparison between zero-shot LongT5-xl and fine-tuned longT5-base. FT fine-tuning, ZS zero-shot learning, ns not significant; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$.

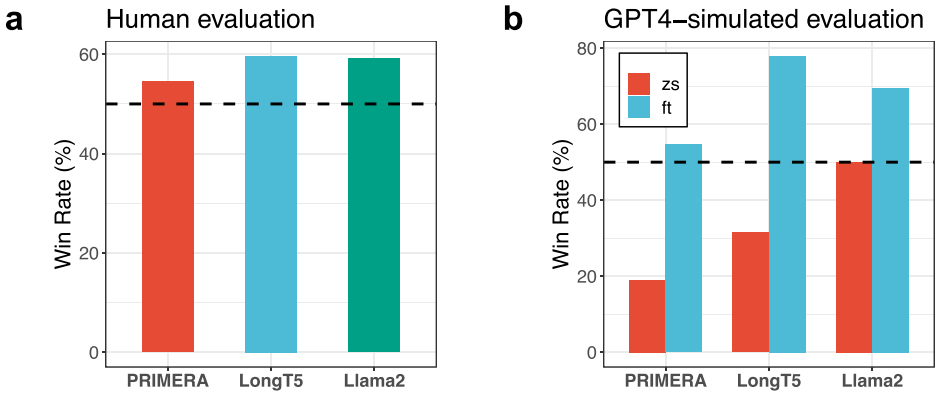


Fig. 4 | Human and GPT4-simulated evaluation of LLM-generated summaries. a Performance of different summarization systems in human evaluations using win-rates against zero-shot Llama-2 (Llama-2-zs). The dotted line represents the default 50% win rate of the Llama-2-zs. **b** Performance of different summarization systems in GPT4-simulated evaluation using win-rate. The dotted line represents the default win rate of Llama-2-zs. zs zero-shot learning, ft fine-tuning.

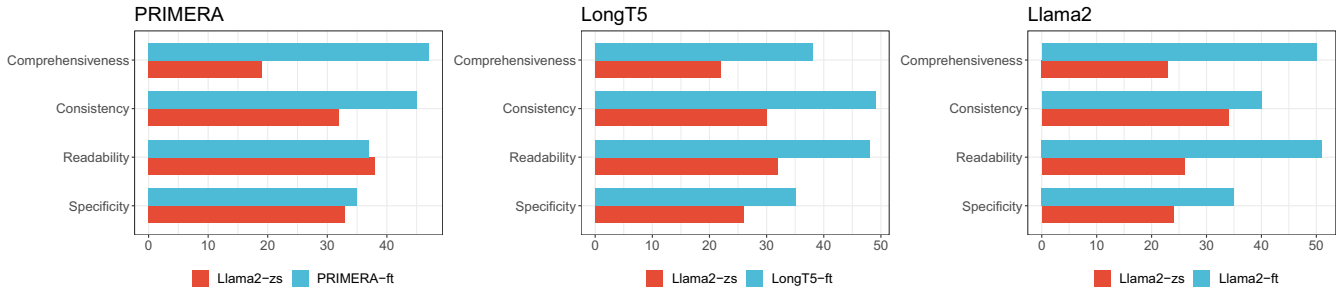


Fig. 5 | The number of summaries where zero-shot LLama-2 generated better summaries (left/red), in contrast to the cases where the fine-tuned models generated better summaries (right/blue). As compared to zero-shot LLama-2, fine-tuned models produced more comprehensive, readable, consistent, and specific summaries in general. Despite PRIMERA and LongT5 having much smaller model architectures, they significantly outperformed zero-shot LLama-2 after fine-tuning. LLama-2 was also improved in all aspects via fine-tuning.

Qualitative evaluation

Finally, we conducted a comprehensive human evaluation and a GPT-4 simulated[26] evaluation of machine-generated summaries. Our baseline, zero-shot Llama-2, is one of the latest and largest open-sourced LLMs by the time of writing. In both evaluations, we requested clinical experts or GPT-4 to select the better summary from a pair of candidates—one generated using the baseline and the other generated using our fine-tuned models. The win rate is the ratio of machine-generated summaries evaluated as better than the baseline. The baseline win rate is 50%, given that (1) zero-shot Llama-2 is being compared to itself, (2) all pairs of summaries to be compared are identical, and (3) ties are broken randomly[24].

According to the human evaluation, fine-tuned Llama-2 was preferred to zero-shot, with the win rate increased from 50% to 59.20% (Fig. 4a and Supplementary Table 3). Fine-tuned PRIMERA and LongT5 models also achieved 54.47% and 59.68% win-rate against the zero-shot Llama-2.

We further asked the evaluators to share the rationale behind their preference for the chosen summaries. The fine-tuned models were compared with zero-shot Llama-2 on multiple dimensions that have been established as desired properties of summaries[7,33]. Figure 5 shows the number of cases where zero-shot LLama-2 generated better summaries (left/red), in contrast to the cases where the fine-tuned models generated better summaries (right/blue). With few exceptions, LLMs were improved in all aspects after fine-tuning (Supplementary Table 5). By manually comparing the summaries generated by zero-shot and fine-tuned models, we found that zero-shot models tend to present a detailed background of the summarized studies but do not provide any findings or conclusions, i.e., they present a high resemblance of leading sentences in the paragraphs. Recall that word embeddings are combined with positional embeddings to represent each token in a document in transformer architecture. In general summarization tasks, the key information is typically presented in leading or concluding

sentences. The ordering of key information and other redundant information can impact the positional embeddings during pre-training. The zero-shot open-source summarization models tend to extract the leading sentences instead of key information. This indicates that positional embeddings significantly impacted the summary more than word embeddings in zero-shot open-source models. On the other hand, fine-tuned models align more closely with ground-truth summaries, which can provide supportive evidence or identify the lack of sufficient evidence for intervention outcomes.

The GPT-4 simulated evaluation also indicates a significant improvement in all models after fine-tuning (Fig. 4b). In addition, 257 out of 378 simulated evaluation results concord with the judgment of human experts (68% accuracy).

We also evaluated all models on two distinct test sets of review articles (Supplementary Table 4). One group, denoted as "after the cutoff," was published after the latest knowledge cutoff date of every model, i.e., no articles in this group were used in pre-training the foundation models. The other group, denoted as "before cutoff", was published before the knowledge cutoff date; the articles in this group may be used for pre-training purposes. On both the "after cutoff" and the "before cutoff" test sets, fine-tuned models demonstrated improved performances. Recap that all articles used for fine-tuning were published "before cutoff". This demonstrates the generalizability of fine-tuned models "before cutoff" data on "after cutoff" test data.

Discussion

In this study, we focused on comparing open-source and closed-source LLMs in medical evidence summarization. While closed-source models, exemplified by GPT families and others, demonstrated superior performance as compared to open-source alternatives, the risks associated with using closed-source models are not negligible, which include lack of transparency, reliance on external dependency as a single-point-of-failure, potentially high-cost in migration to other vendors. Furthermore, it's still unclear whether patients would consent to have their information utilized in LLMs, especially when they were unaware of and unable to understand what the LLMs will be used. However, such disadvantages were not paid enough attention. To mitigate the above risks, we investigated recently released open-source LLMs for medical evidence summarization. To our best knowledge, open-source models still fall behind in natural language understanding, as compared to closed-source ones. Based on the observation that fine-tuning can enhance, we validate principles of model optimization within evidence summarization and quantitatively measure the performance boost via fine-tuning.

To facilitate future studies in this direction, we first introduced MedReview, a collection of meta-analysis and summary pairs to assist in fine-tuning LLMs for medical evidence summarization. We further showed that open-source LLMs, including LongT5, PRIMERA, and Llama-2, demonstrated improved summarization performance after fine-tuning with MedReview, even close to GPT-3.5-turbo. This observation also confirms that these generalist models perform well across diverse tasks but need more in-depth domain-specific knowledge, e.g., in biomedicine. While few-shot learning has been effective in other NLP tasks, the applicability is still limited by the context window within lengthy document summarization. Fine-tuning is a robust approach to bridge this performance gap between open-source models and closed-source alternatives, while maintaining the advantages of transparency, easy customization, maintenance, and migration. However, the manual evaluation results show that fine-tuning does not guarantee truthful and accurate summaries. This issue was also reported in the previous evaluation of zero-shot GPT models[7]. This highlights that trustworthy summarization of medical evidence remains challenging and unresolved.

In addition, the usage of large models is limited by the high demand for computing resources. Our experiments indicated that smaller models, when fine-tuned, can sometimes outperform zero-shot larger models on specific tasks. This is observed in two experiments. First, the fine-tuned LongT5-base performs better than the zero-shot LongT5-xl (Fig. 3). Second, PRIMERA and LongT5-xl, despite being smaller in size than Llama-2 (70B), demonstrated enhanced performance than the latter one upon fine-tuning (Fig. 5 and Supplementary Table 2). In the cases of limited computing resources, PRIMERA demonstrated robust performance in medical evidence summarization in the few-shot fine-tuning setup, which confirms the findings of the original reports of the PRIMERA paper. However, few-shot fine-tuned models underperformed as compared to those fully fine-tuned on the entire training data.

The evaluation of LLM-generated summaries presents another challenge in this task. Automatic metrics frequently used for text summarization can, at best, measure the similarity between word distributions of references and LLM-generated summaries. These metrics do not correlate strongly with properties of desired summaries, such as factual comprehensiveness and consistency. As such, human evaluation, especially those from clinical experts, is still critical but not scalable. Therefore, we used GPT-4 as a simulated evaluator and found that 68% of GPT4-simulated evaluation results aligned with human judgments. These findings suggest that GPT-4 holds the promise to not only facilitate summary evaluation, but also to provide feedback to align the summarization model with simulated expert feedback. The simulated evaluation results further confirmed that fine-tuned smaller models can surpass larger zero-shot models (Fig. 4b). This observation is consistent with the improvement discussed in the automatic evaluation.

Summarization systems have diverse applications, particularly benefiting researchers and healthcare professionals. For researchers engaged in systematic reviews, these systems can analyze clinical trial reports efficiently, pinpointing relevant studies and distilling essential findings without requiring exhaustive document review. This significantly speeds up the research process. Healthcare professionals and policymakers, on the other hand, can utilize these systems to efficiently grasp the latest clinical trials' outcomes and implications, which is vital for informed decision-making concerning patient care, treatment guidelines, and healthcare policies, especially under the urgent pandemic conditions. Furthermore, by condensing the latest clinical trial information, summarization systems provide concise, current data to clinical decision support systems, assisting healthcare providers in making evidence-based treatment decisions customized to their patient's specific needs.

This study has certain limitations. Due to the restricted time and resources, we cannot extensively explore fine-tuning models such as Claude, GPT-3.5-turbo, and GPT-4. Instead, we mainly focused on open-sourced LLMs. This restriction, however, does not impact the applicability of our study in the clinical domain. We believe that open-sourced LLMs foster wider collaboration and transparency, thus forwarding research progress as other researchers can enhance, modify, and refine these models. Plus, by advocating for the democratization of technology, these LLMs encourage a more extensive use and a potentially higher rate of innovation. Another limitation is that the LLMs were not fine-tuned to summarize clinical trial publications but the manually curated "main results" of review abstracts. This study design was initially aimed at simplifying and expediting the development and testing of our summarization algorithms. A future direction is to deploy LLMs to directly synthesize information from clinical trials.

To summarize, our findings underscore the utility of fine-tuning as a robust technique for bridging the performance gap between open-source LLMs and closed-source ones, reinforcing its applicability across a spectrum of model architectures and sizes, and setting the stage for more nuanced investigations into the efficiency and effectiveness of model optimization strategies. However, additional research is warranted to fully uncover the potential of LLM in this context.

Methods

Data collection

We collected 8161 abstracts of systematic reviews from the Cochrane Library[32]. Unlike abstracts of the biomedical literature, which are highly condensed summaries, systematic review abstracts provide a structured overview that enables readers to quickly determine the validity and

applicability of the review. These abstracts typically follow a common structure, detailing preferred reporting items[5]. The Cochrane review abstracts present background, objectives, search methods, selection criteria, data collection and analysis, main results, and authors' conclusions. Within such a self-contained structure, the authors' conclusion presents a narrative summary of the most salient details of the included clinical studies[7]. This section is one of the first to consult when healthcare providers seek answers to clinical questions. Given the meta-analysis results as input, we aim to automatically reproduce this narrative summary. The collected reviews cover a wide range of topics, including but not limited to neurology, gastroenterology, rheumatology, nephrology, and radiology. The publication dates of the reviews range from April 1996 to June 2023.

We split the dataset into distinct training (91.56%), validation (4.83%), and test (3.61%) sets, ensuring that all of the samples appear in one set (Supplementary Table 1). All LLMs were prepared with a large extent of public textual data collected until a certain moment, known as the cutoff. To maintain a legitimate comparison between LLMs, we put all articles published after September 2022 as test data (because most LLMs studied in this study used the data up to September 2022). Articles published prior to this are primarily used for training and validation. The division of training and validation was stratified according to the time of publication.

Few-shot baselines

Few-shot learning has been proven an effective and sample-efficient strategy for optimizing task-specific LLMs. We construct two few-shot learning baselines, one based on prompting and the other on fine-tuning.

For few-shot prompting, we use Mixtral-7x8B for the few-shot prompting foundation model. Due to the limit on token numbers, other open-source models cannot fit demonstrations of long document summarization in the context windows. We randomly selected 1, 2, and 5 samples from the training set as demonstrations. For few-shot fine-tuning, we followed the findings of the PRIMERA report that LLMs can be reasonably adapted to domain-specific tasks with a limited number of labeled samples. We used the same setup as the few-shot experiments in the PRIMERA, where we randomly selected 100 samples from the training set and fine-tuned LLMs.

Fine-tuning LLMs

We investigated several LLM architectures that have recently surfaced for tasks related to summarization tasks or as foundation models. In this study, we only consider models that satisfy the following two conditions. First, models need to be publicly accessible and open-sourced to ensure their transparency and accountability. Second, context windows need to be long enough to digest input without requiring condensation or truncation. Bearing all these factors in mind, we included PRIMERA, LongT5, and Llama-2 in our studies. PRIMERA deploys a pretraining strategy named Entity Pyramid to select and aggregate salient information focusing on document summarization[15]. LongT5 is an extension of T5 architecture that adopts a summarization pretraining strategy to scale up the input length[14]. Llama-2 is one of the recently released open-source, scalable foundation models with 7B, 13B, and 70B parameters[25]. Since Mixtral-8x7B demonstrated similar benchmark performance as Llama-2, we did not fine-tune Mixtral-8x7B[23]. Instead, we report the few-shot prompting performance of Mixtral-8x7B.

Following previous work[7], we selected the objective and main results sections of a review abstract as input and used the authors' conclusion as a reference to fine-tune models. We applied the LoRA method, which keeps the original parameters frozen and adjusts only a relatively small number of extra parameters via matrix decomposition[31]. The exact number of parameters depends on the rank hyperparameter in the LoRA method. We refer the readers to the original LoRA paper for technical details[31].

Our implementation uses the following libraries: transformers[34], torch[35], and PEFT[36]. Most fine-tuning jobs were completed on AWS and our local lab servers. Llama-2 models were fine-tuned on the SageMaker platform. All models were fine-tuned for 1 epoch, within which the validation loss had already stopped decreasing. We set the learning rates to 3e-5[15], 1e-4[25], and 1e-3[14] for PRIMERA, LLama-2, and LongT5 models, respectively. We set the rank hyperparameter of LoRA to 8[31].

Evaluation metrics

We first use the natural language generation (NLG) metrics to evaluate the quality of the generated summary. These metrics include ROUGE-L (recall-oriented understudy for gisting evaluation) and METEOR (metric for evaluation of translation with explicit ordering) scores. We also include the CHRF (CHaRacter-level *F*-score), which was reported to correlate highly with the readability of generated text[33]. Their values range from 0.0 to 100.0, with a score of 100.0 indicating that the generated summaries are identical to the reference summary. The model performance on our test is approximately normally distributed and the performance of each model is independent of others. As such, we calculated the p-value using a paired t-test to determine the statistical significance of the difference between the two models.

PICO metrics

NLG metrics are known to be inadequate for evaluating factual completeness and consistency[33]. We therefore propose to use a PICO (participants, interventions, comparison, and outcomes) extraction system to evaluate the accuracy of the generated summaries. More specifically, we fine-tuned a BERT-based[13,37] model to extract PICO concepts and then score a generated summary by comparing the values of these PICO elements obtained from the reference. We consider a PICO element to be a true positive, if it satisfies two conditions: (1) the text in the reference overlaps with the text in the generation and (2) the two entity types should have the same PICO category. The micro averages for precision, recall, and F1 scores are all computed over the PICO components.

Human evaluation

We conducted a review of the summary quality via human evaluation. The quality is measured from four aspects: consistency, comprehensiveness, specificity, and readability, which were established as essential factors for measuring machine summary quality. Consistency indicates whether the summary contradicts the input source. Comprehensiveness measures coverage of key information of input. Specificity measures the preciseness and conciseness of the summary. Readability indicates a machine summary is fluent and free of grammatical errors that hinder understanding.

To evaluate the machine summaries, we invited seven clinical experts, each specializing in one or two of the following specialties, including Gastroenterology, General Surgery, Internal Medicine, Nephrology, Neurology, Radiology, and Rheumatology. All experts have obtained MD training and currently provide direct patient care. Following a recent LLM study[24], we request the experts to compare the quality of different machine summaries (interface shown in Supplementary Fig. 1). Specifically, according to their domain knowledge, each expert was assigned review abstracts along with three summaries: (1) the authors' conclusion section, (2) zero-shot Llama-2, and (3) one of the fine-tuned models. From the zero-shot baseline and the fine-tuned models, the experts will select which one generates a better summary. To reduce the potential order-related bias, the order of summaries generated by zero-shot Llama-2 and fine-tuned summaries was randomized. We further asked the experts to choose the reasons for their choices.

GPT-4 evaluation

Like many other annotation scenarios, collecting experts' feedback is not scalable with respect to the samples to be labeled. In addition to our manual review, we explored the use of GPT-4[26] as a simulated expert to answer the same questions as those assigned to human experts. Instead of selecting a sample of test data as in the manual review, we used the model summaries for all test articles for GPT-4 evaluation. We also analyzed the percentage of questions that human judgments agree with the GPT-4 evaluation.

Data availability

The data underlying this article will be available upon request.

Code availability

The codes underlying this article will be available upon request.

Received: 15 February 2024; Accepted: 29 August 2024;
Published online: 09 September 2024

References

1. Peng, Y., Rousseau, J. F., Shortliffe, E. H. & Weng, C. AI-generated text may have a role in evidence-based medicine. *Nat. Med.* **29**, 1593–1594 (2023).
2. Concato, J., Shah, N. & Horwitz, R. I. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* **342**, 1887–1892 (2000).
3. Borah, R., Brown, A. W., Capers, P. L. & Kaiser, K. A. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* **7**, e012545 (2017).
4. ClinicalTrials.gov. U.S. National Library of Medicine. Available at: https://clinicaltrials.gov (Accessed: 4 September 2024).
5. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Rev. Esp. Cardiol.* **74**, 790–799 (2021).
6. Wallace, B. C., Saha, S., Soboczenski, F. & Marshall, I. J. Generating (Factual?) narrative summaries of RCTs: experiments with neural multi-document summarization. *AMIA Jt. Summits Transl. Sci. Proc.* **2021**, 605–614 (2021).
7. Tang, L. et al. Evaluating large language models on medical evidence summarization. *NPJ Digit. Med.* **6**, 158 (2023).
8. Barzilay, R. & Elhadad, N. Inferring strategies for sentence ordering in multidocument news summarization. *J. Artif. Intell. Res.* **17**, 35–55 (2002).
9. Pivovarov, R. & Elhadad, N. Automated methods for the summarization of electronic health records. *J. Am. Med. Inform. Assoc.* **22**, 938–947 (2015).
10. Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K. B. Frontiers of biomedical text mining: current progress. *Brief. Bioinform.* **8**, 358–375 (2007).
11. Li, F. et al. Structure-aware review mining and summarization. In *Proc. 23rd International Conference on Computational Linguistics (Coling 2010)* (eds. Huang, C.-R. & Jurafsky, D.) 653–661 (Coling 2010 Organizing Committee, Beijing, 2010).
12. Demner-Fushman, D. & Lin, J. J. Answering clinical questions with knowledge-based and statistical techniques. *Comput. Linguist.* **33**, 63–103 (2007).
13. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2021).
14. Guo, M. et al. LongT5: Efficient Text-To-Text Transformer for Long Sequences. Findings of the Association for Computational Linguistics: NAACL 2022, 724–736. https://doi.org/10.18653/v1/2022.findings-naacl.55 (2022).
15. Xiao, W., Beltagy, I., Carenini, G. & Cohan, A. PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. In *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol 1: Long Papers) 5245–5263 (ACL 2022).
16. Zhang, J., Zhao, Y., Saleh, M. & Liu, P. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proc. 37th International Conference on Machine Learning* (eds. Iii, H. D. & Singh, A.) 11328–11339 (PMLR, 2020).
17. Lewis, M. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics* 7871–7880. *(ACL 2020)*, https://doi.org/10.18653/v1/2020.acl-main.703.
18. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* 4171–4186*. (NAACL-HLT 2019)*, https://doi.org/10.18653/v1/n19-1423.
19. Mrabet, Y. & Demner-Fushman, D. HOLMS: alternative summary evaluation with large language models. In *Proc. 28th International Conference on Computational Linguistics* (eds. Scott, D., Bel, N. & Zong, C.) 5679–5688 (International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020).
20. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
21. Zack, T. et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).
22. Jin, Q., Yang, Y., Chen, Q. & Lu, Z. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinform.* **40**, btae075 (2024).
23. Jiang, A. Q. et al. Mixtral of experts. Preprint at https://doi.org/10.48550/arXiv.2401.04088 (2024).
24. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst* **35**, 27730–27744 (2022).
25. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://doi.org/10.48550/arXiv.2307.09288 (2023).
26. OpenAI, R. Gpt-4 technical report. Preprint at https://doi.org/10.48550/arxiv.2303.08774 (2023).
27. Nosek, B. A. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
28. Zhang, G. et al. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. *J. Biomed. Inform.* **153**, 104640 (2024).
29. Gutierrez, B. J., et al. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. Findings of the Association for Computational Linguistics: EMNLP 2022, 4497–4512. https://doi.org/10.18653/v1/2022.findings-emnlp.329 (2022).
30. Tadros, T., Krishnan, G. P., Ramyaa, R. & Bazhenov, M. Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. *Nat. Commun.* **13**, 7742 (2022).
31. Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. The Tenth International Conference on Learning Representations (ICLR 2022).
32. The Cochrane Library. https://www.cochranelibrary.com/.
33. Fabbri, A. R. et al. SummEval: Re-evaluating Summarization Evaluation. *Trans. Assoc. Comput. Linguistics* **9**, 391–409 (2021).
34. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds. Liu, Q. & Schlangen, D.) 38–45 (Association for Computational Linguistics, Online, 2020).
35. Paszke, A. et al. Automatic differentiation in PyTorch (2017).
36. Mangrulkar, S. et al. *PEFT: State-of-the-Art Parameter-Efficient Fine-Tuning Methods*. https://github.com/huggingface/peft (2022).
37. Zhang, G. et al. A span-based model for extracting overlapping PICO entities from randomized controlled trial publications. *J. Am. Med. Inform. Assoc.* **31**, 1163–1171 (2024).

Acknowledgements

Funding/support: this project was sponsored by the National Library of Medicine grant R01LM009886, R01LM014344, T15LM007079, the National Human Genome Research Institute grant R01HG012655, and the National Center for Advancing Clinical and Translational Science awards UL1TR001873 and UL1TR002384. Q.J. and Z.L. are supported by the NIH Intramural Research Program, National Library of Medicine. We also want to thank Amazon Web Services (AWS) for providing the computational resources used in our research. Role of the funder/sponsor: the funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author contributions

Study concepts/study design: G.Z., C.W., and Y.P.; manuscript drafting or manuscript revision for important intellectual content: G.Z., Q.J., Y.Z., S.W., B.I., Y.L., E.P., J.G.N., M.E.S., A.S., T.C., Z.L., C.W., and Y.P.; approval of final version of the submitted manuscript: G.Z., Q.J., Y.Z., S.W., B.I., Y.L., E.P., J.G.N., M.E.S., A.S., T.C., Z.L., C.W., and Y.P.; agrees to ensure any questions related to the work are appropriately resolved: G.Z., Q.J., Y.Z., S.W., B.I., Y.L., E.P., J.G.N., M.E.S., A.S., T.C., Z.L., C.W., and Y.P.; literature research: G.Z. and Y.P.; experimental studies: G.Z., Q.J., Y.Z., and S.W.; human evaluation: Q.J., B.I., Y.L., E.P., J.G.N., M.E.S., and A.S.; data interpretation and statistical analysis: G.Z. and Y.P.; and manuscript editing: G.Z., Q.J., Y.Z., S.W., B.I., Y.L., E.P., J.G.N., M.E.S., A.S., T.C., Z.L., C.W., and Y.P.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01239-w.

Correspondence and requests for materials should be addressed to Chunhua Weng or Yifan Peng.