CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Fine-tuning generative pre-trained transformers for clinical dialogue summarization
Authors	Ronan, Isabel;Tabirca, Sabin
Publication date	2025-01-17
Original Citation	Ronan, I. and Tabirca, S. (2025) 'Fine-tuning generative pre-trained transformers for clinical dialogue summarization', 2024 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 9-10 December 2024, pp. 1-6. https://doi.org/10.1109/FIT63703.2024.10838420
Type of publication	Conference item;proceedings-article
Link to publisher's version	10.1109/fit63703.2024.10838420
Rights	© 2025, IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Download date	2026-11-16 14:56:12
Item downloaded from	https://hdl.handle.net/10468/17083



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Fine-Tuning Generative Pre-Trained Transformers for Clinical Dialogue Summarization

1st Isabel Ronan
School of Computer Science
University College Cork
Cork, Ireland
i.ronan@cs.ucc.ie

2nd Sabin Tabirca
School of Computer Science
University College Cork
Cork, Ireland
tabirca@cs.ucc.ie

***Abstract*—Automated clinical dialogue summarization can help make health professional workflows m ore e fficient. Wi th the advent of large language models, machine learning can be used to provide accurate and efficient s ummarization t ools. Generative Pre-Trained Transformers (GPT) have shown huge promise in this area. While larger GPT models, such as GPT-4, have been used, these models pose their own problems in terms of precision and expense. Fine-tuning smaller models can lead to more accurate results with less computational expense. In this paper, we fine-tune a GPT-3.5 model to summarize clinical dialogue. We use both default hyperparameters along with manual hyperparameters for comparison purposes. We also compare our default model to past work using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScores. We find o ur m odel o utperforms G PT-4 a cross all measures. As our fine-tuning process is based on the smaller GPT-3.5 model, we show that fine-tuning l eads t o m ore a ccurate and less expensive results. Informal human observation also reveals our notes to be of acceptable quality.**

***Index Terms*—Machine translation, Transformers, Data augmentation, Synthetic data generation, Parameter tuning**

I. Introduction

Healthcare is slowly evolving into a personalised care model in which the patient as an "hyper-individual" is capable of accessing a lot of their healthcare needs online [1]. This new health landscape enables online consultation, along with increased dialogue recording capabilities [2]. Dialogue recording in many clinical situations, both in-person and online, can reduce the workload placed on busy healthcare professionals, while also allowing them to maximise on time spend with their patients [2], [3].

With the advent of large language models, many have turned to transformer-based models, including Generative Pre-Trained Transformers (GPT), in order to effectively summarize large bodies of text [2], [4]. Michalopoulos et al. [5] used a transformer model to summarize patient-doctor conversations. They use medical guidance signals to introduce domain-specific knowledge into the model for accuracy improvements. These signals enable the model to focus on medical terms and concepts. The authors use a proprietary dataset consisting of real medical encounters to aid their research. This work is significant in its use of real-world data and domain-specific guidance. Their evaluation using ROUGE-1 and ROUGE-L scores indicated that their approach reduced errors when compared against a baseline. However, they also highlight that their approach is quite computationally intensive, and thus may not be suitable for low-resource scenarios.

Navarro, Dras and Berkovsky [6] trained 10 transformer models to assess their performance with regards to fine-tuning. They found that models pre-trained on general dialogues performed better than baseline models. They also found that fine-tuning models on medical datasets improved summarization outputs. Their BART-based models achieved relatively high scores when assessed using various metrics, such as ROUGE-1, ROUGE-2 and ROUGE-L scores.

While BART-based models have shown promise, these models have a limited input sequence length. This problem is highlighted by Suri, Saha and Singh [7], who proposed a two-stage prompt approach involving initial dialogue summarization followed by clinical summarization. They show that fine-tuning transformer models can lead to results as good as in-context learning using OpenAI's GPT-3. They also highlight their use of BERTScore, describing it as a suitable metric for such research.

Tang, Tran, Tan and Gerstein [4] investigated the use of a a GPT-4 model along with a fine-tuned Davinci model (GPT-3) to summarize doctor-patient conversations. Their results revealed GPT-4 to be the most acceptable model, while fine-tuned GPT-3 also performed relatively well. Their work reveals the potential for the use of GPT models in such summarization tasks.

Liu, Ju and Wang [2] have investigated the use of GPT-3.5 with inter-context learning to generate medical dialogue summarizations. These summarizations were compared against a BART model. While BART outperformed GPT-3.5 regarding ROUGE-1, ROUGE-2, ROUGE-L and BERTScore results, GPT-3.5 was favoured by medical experts during manual evaluation. This paper highlights the capabilities of GPT-3.5 in such contexts.

While other GPT models have been fine-tuned in similar contexts [4], [8], GPT-3.5 offers a less-resource intensive

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6222. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

alternative to GPT-4, while also being much improved from GPT-3. Ouyang et al. [9] have highlighted that larger models, such as GPT-4, are not always the optimal solutions. Large language models are created for a broad range of tasks; utilising them in certain use-cases can lead to inappropriate output. Additionally, larger models are more computationally demanding than smaller models. Ouyang et al. [9] argue that using smaller fine-tuned models to generate results in specific domains is more effective than using larger and more generalized models. They fine-tuned a GPT-3 model on a commercial dataset to show that such fine-tuning and model alignment can effectively balance computation and correctness.

In this paper, we focus on furthering clinical dialogue summarization research by fine-tuning a GPT-3.5 Turbo model on a clinical dialogue dataset. Our contribution to the field is the first use of a fine-tuned GPT-3.5 model on a dataset of this type. Additionally, the different hyperparameter options used during fine-tuning have not been reported; we intend to test both default and manual hyperparameter tuning methods in order to discern the reliability default parameters. We will discuss our methodology in Section II and our results in Section III, before summarizing our findings in Sections IV and V.

II. Methodology

In order to effectively train out models in a consistent manner, we used the OpenAI API [10]. Additionally, we used the `aci-bench` dataset [11] to train, validate and test our models. This dataset has already been tested using standard GPT-4, GPT-3.5 and GPT-3 models [11]. However, fine-tuned GPT-3.5 models have not yet been tested [4], [11]. In this section, we describe the steps we took in our fine-tuning and subsequent testing processes. We will describe training and test data along with our fine-tuning and evaluation procedures. Figure 1 provides an overview of our methodology; each box represents a different step in the training and testing process. The yellow box in Figure 1 represents the model training and validation process. The green box in this figure represents subsequent data generation using our fine-tuned models along with the test dataset. Additionally, the pink and blue boxes in the same represent the comparison between model and human output using ROUGE scores, BERTScore, and informal human evaluation.

A. Dataset

We used the `aci-bench` dataset [11] to train and test our models. We chose this dataset as it is composed entirely of synthetic conversations between medical annotators and patient actors; therefore, no real patients are involved in the generation process. This circumstance circumvents ethical issues surrounding the use of anonymized records. Additionally, this dataset has already been used to test the performance of standard GPT-4 and GPT-3.5 models in clinical dialogue summarization tasks; therefore, the evaluation of our model should be comparable to this test.

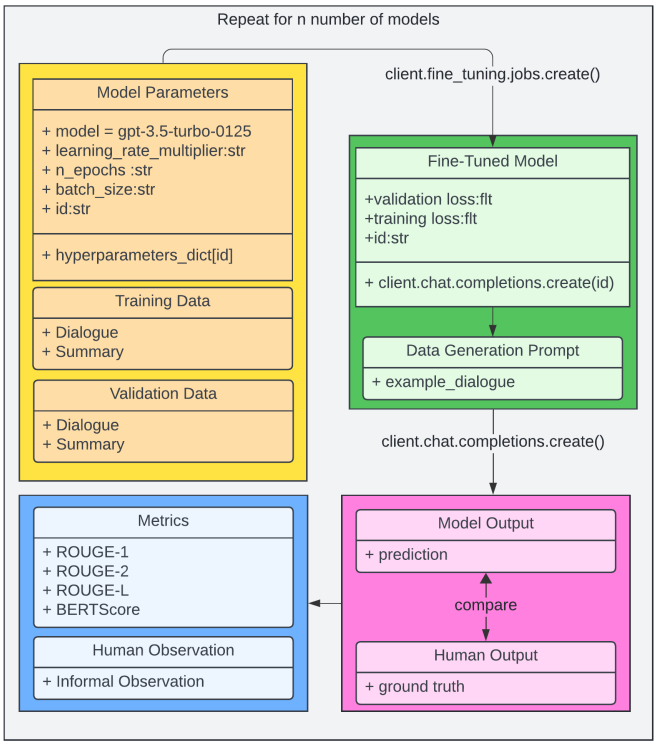


Fig. 1. Fine-Tuning Research Process

The dataset was downloaded from Microsoft's `clinical_visit_note_summarization_corpus` [12]. This dataset is openly available for research under a Creative Commons Attribution 4.0 International Licence.

We used a total of 207 data points, with each point consisting of a clinical dialogue transcription and the resultant clinical notes. Data from training and validation was obtained from the train and test files available in Microsoft's repository containing 67 (32.37%) and 20 (9.66%) different data points respectively. An additional 120 (57.97%) data points were used for testing. These 120 points were derived from 3 test files associated with the same dataset. The average number of sentences per dialogue was 80 and the average number of sentences per clinical note was 49 [11].

B. Fine-Tuning

Our fine-tuning process was undertaken using the OpenAI API [10]. This API allows researchers to access new AI models developed by OpenAI, while also enabling the development of programmatic workflows. Although it is possible to fine-tune GPT models using a UI interface, we chose to train our models using Python in order to maintain consistency between our different processes. Additionally, using Python ensures the replicability of our results.

1) Base Model: The base model chosen for experimentation was `gpt-3.5-turbo-0125`. This model was selected based on OpenAI documentation recommendations [13]. GPT-3.5 models have 175 billion parameters, making them highly suited to text-based tasks [14]. Of these models, GPT-3.5

```
{"messages": [
    {"role": "system",
    "content": "You are an expert medical professional. Given a clinical dialogue,
    create a clinical note outlining key dialogue aspects such as 'CHIEF COMPLAINT' (or
    'CC'), 'HISTORY OF PRESENT ILLNESS' ('HPI'), 'REVIEW OF SYSTEMS', 'PHYSICAL
    EXAMINATION', 'VITALS REVIEWED', 'RESULTS', 'ASSESSMENT AND PLAN', 'INSTRUCTIONS',
    'CURRENT MEDICATIONS', 'PAST MEDICAL HISTORY', 'EXAM', 'IMPRESSION', 'PLAN',
    'ASSESSMENT', 'PAST HISTORY', 'ALLERGIES', 'SOCIAL HISTORY', 'PHYSICAL EXAM',
    'PROCEDURE', 'FAMILY HISTORY', 'MEDICATIONS', 'VITALS', 'MEDICAL HISTORY', 'SURGICAL
    HISTORY'. You will not use all of these aspects in every dialogue, vary it from
    dialogue to dialogue."},
    {"role": "user",
    "content": "Dialogue:[doctor] hi , martha . how are you ?\n[patient] ..."},
    {"role": "assistant",
    "content": "CHIEF COMPLAINT\n\nAnnual exam.\n\nHISTORY OF ..."}
]}
```

Fig. 2. Prompt Template

Turbo is considered the most capable and least computationally intensive [14].

2) Inputs: When using OpenAI models, a specific training and validation data format is required when submitting a tuning job. This format includes the use of `.jsonl` files, along with dictionary-style training input. We used a chat-style prompt input for training and testing purposes. This style is chosen due to the ability to use prompt engineering techniques in our input. We used direct language and system role prompts to prevent the GPT-3.5 model misinterpreting instructions. Our prompt template is described in Figure 2. This figure shows the three different components of each data point in the training and validation data; the system, user, and assistant components. The system message remains the same for every data point, whereas the user and assistant messages change for every data point. There were 67 and 20 data points in our training and validation sets respectively.

3) Training: OpenAI allows for three hyperparameters to be changed during model fine-tuning: `n_epochs`, `learning_rate_multiplier`, and `batch_size`. The `n_epochs` parameter refers to the number of epochs the model will be trained for; an epoch refers to one full training dataset cycle. The `learning_rate_multiplier` property refers to the learning rate scaling factor. Finally, the `batch_size` variable refers to the number of samples in each batch.

We trained a model using OpenAI's default hyperparameters before we made any manual adjustments to the same. We used a random seed of 42 for all model training processes. Upon training this model, we found the `n_epochs`, `learning_rate_multiplier`, and `batch_size` to be set to 3, 2, 1 respectively. We used the default hyperparameters determined by the OpenAI API to inform all other hyperparameter testing combinations. We undertook manual hyperparameter testing in order to provide a comprehensive overview of the fine-tuning space. Therefore, we chose to test 0.5, 1 and 2 for the `learning_rate_multiplier` parameter, 3 and 10 for the `n_epochs` parameter, and 1, 32 and 67 for the `batch_size` parameter. We trained a total of 19 different models: 18 versions with manually-created hyperparameters and one default version.

C. Evaluation Process

We evaluated our fine-tuned models using 120 test notes available in the `aci-bench` dataset. These test notes were not used at any point during the training or validation processes. We tested all of our trained models. These tests were run by creating completions for each note and comparing the model notes with the corresponding human notes. In order to maintain consistency with the literature, we chose the same evaluation metrics as those used in previous work [11]. The metrics we report in this paper are ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore. We report our ROUGE-L metrics using the ROUGE-L summary method as reported in the original ROUGE and `aci-bench` papers [11], [15]. All ROUGE metrics were computed using the `rouge-score` Python library [16]. We used the `bert-score` Python library to calculate BERTScore [17]. ROUGE-1 and ROUGE-2 calculate the overlap of single and two consecutive words respectively [15]; these metrics are based off ROUGE-N, which is outlined in Equation 1.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{1}$$

In Equation 1, n refers to the length of the n-gram $gram_n$, and the maximum number of n-grams co-occurring in candidates and references is represented by $Count_{match}(gram_n)$. ROUGE-L was also used; this metric captures the longest common subsequence at a summary level, making it a holistic measure of the human vs model output [15].

We used BERTScore to calculate contextual similarity between the human and model summaries; this score is based of BERT embeddings and has been shown to correlate well with human judgements [18]. In equations 2, 3 and 4, we outline the necessary calculations for BERTScore, where x refers to a tokenized reference sentence and $\hat{x}$ refers to a tokenized candidate sentence. Equation 2 outlines the recall calculation for BERTScore. Equation 3 describes BERTScore's precision calculation. These two metrics are combined together to create the F1 score, as shown in Equation 4. F1 score is primarily used in our analysis as recommended by Zhang, Kishore, Wu, Weinberger and Artzi [18].

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \tag{2}$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j \tag{3}$$

$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{4}$$

We calculated the average scores for each model and metric for all notes during the evaluation process. We also used informal human observation to provide further insight into the model outputs.

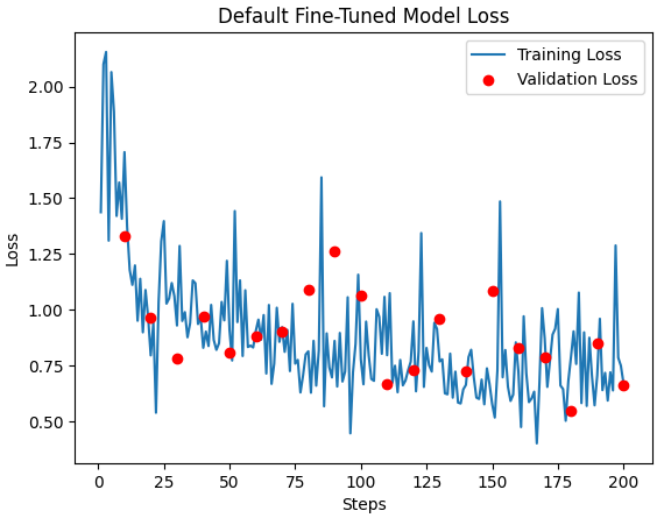


Fig. 3. Default Hyperparameter Training

TABLE I
MODEL TRAINING PERFORMANCE

ID	Tokens Trained	Learning Rate Multiplier	Epochs	Batch Size	Validation Loss
id0	472,395	0.5	3	1	0.9285
id1	472,395	0.5	3	32	0.9966
id2	472,395	0.5	3	67	1.1496
id3	1,574,650	0.5	10	1	0.9144
id4	1,574,650	0.5	10	32	0.9212
id5	1,574,650	0.5	10	67	0.9473
id6	472,395	1	3	1	0.8946
id7	472,395	1	3	32	0.9494
id8	472,395	1	3	67	1.0236
id9	1,574,650	1	10	1	0.9825
id10	1,574,650	1	10	32	0.8851
id11	1,574,650	1	10	67	0.9081
id12*	472,395	2	3	1	0.8778
id13	472,395	2	3	32	0.9205
id14	472,395	2	3	67	0.9812
id15	1,574,650	2	10	1	1.1277
id16	1,574,650	2	10	32	0.8623
id17	1,574,650	2	10	67	0.8676

* default fine-tuning parameters

III. RESULTS

A. Training Results

Using default parameters, we achieved a training loss value of 0.6734 and a validation loss value of 0.8788. Figure 3 outlines the default fine-tuned model (`default-aci-bench`) training process.

When fine-tuned using OpenAI's default parameters, our `default-aci-bench` model achieved a relatively low validation loss when compared against the high loss reported in the original GPT-3 paper [19]. Table I outlines the loss function results of our manual hyperparameter tuning. These models also achieved appropriately low validation losses. The default training model performed well, demonstrating the best balance of fidelity and efficiency. As this model is recommended for use by OpenAI [13], we will discuss the default hyperparameters in subsequent sections of this paper.

1) Evaluation Metrics: Figure 4 shows the results of our metric-based tests. From these results we can see that our `default-aci-bench` model achieved higher results than both the GPT-3.5 and GPT-4 implementations used in previous work [11]. Our model achieved ROUGE-1 scores 7.81% and 3.49% higher than standard GPT-3.5 and GPT-4 models respectively. Similarly the model exhibited 4.64% increase in ROUGE-2 score when compared to the standard GPT-3.5 model. ROUGE-L, our holistic summarization measure, shows that our model achieved scores 11.37% higher than GPT-4, and 7.87% higher than GPT-3.5. The BERTScore computed for this model is also improved from GPT-3.5 and GPT-4 results.

We also compared our model performance on synthetic data to the work of Michalopoulos et al. [5], who used both real and synthetic data to assess performance. Our model outperformed their model's aggregated performance on real data by 1.34% and 8.16% for ROUGE-1 and ROUGE-L scores respectively. Our model also outperformed their model's performance on synthetic data by 27.48% and 27.78% for ROUGE-1 and ROUGE-L scores respectively.

Informal evaluation of the default model's generated notes reveals the same to be quite consistent and accurate in nature. However, there were some instances where our model added additional overly-helpful chat-style phrases such as "Let me know if you need anything." and "Is there anything else you would like to add?". Additionally, the model "hallucinated" information in some notes. For example, in one case, the human note referred to the patient as "Angela Powell", whereas our model called the same lady "Angela Williams". However, in most of these cases, the dialogue related to these notes does not confirm or deny that either the human or the model is correct. As in the aforementioned dialogue, the woman's surname is not mentioned during the conversation; she is only referred to as "Angela". Therefore, unknown human knowledge may give the model an unfair disadvantage; alternatively, both the model and the human may be "hallucinating", as has happened in previous studies [20]. However, overall, the observed notes were coherent and context appropriate.

IV. DISCUSSION

Our fine-tuned model exhibits the potential for smaller, more accurate models to be trained for specific tasks. In the following section we will compare our fine-tuned model with non-tuned models, discuss limitations of our research, and future directions to explore.

A. Overall Findings

We trained several different models using a selection of hyperparameter tuning options available within the OpenAI API. Results from our work reveal that the default hyperparameters provided by the API are the optimal parameters for our task.

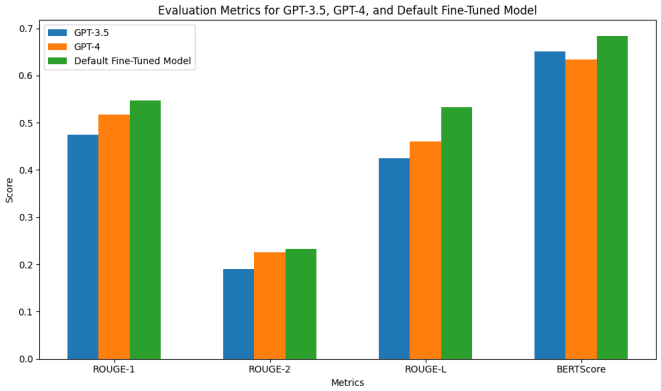


Fig. 4. Model Metrics

These hyperparameters allowed for a balance of exactitude and computational expense. We would recommend that future researchers also use the API's default parameters.

Our `default-aci-bench` model performs better than past non-tuned GPT models; we compared our model with those reported in the original `aci-bench` dataset paper [11] using standard ROUGE and BERTScore evaluation metrics. Additionally, informal human evaluation of the model output reveals acceptable medical summarizations. As our findings show, non-tuned models exhibit acceptable results, but could be further improved using fine-tuning. Our model not only outperforms the standard GPT-3.5 model, but also outperforms GPT-4, its more computationally expensive counterpart. Therefore, hyperparameter-tuned models may outperform more expensive, generalized models within specific contexts.

B. Limitations

Like any research, our study comes with its own set of limitations.

Firstly, the `aci-bench` dataset is a synthetic dataset which is not based on real patient encounters. While this dataset offers advantages over real-world data in terms of data privacy, the results of our model may not hold true when used on a real-world dataset. Additionally, while the model proposed by Michalopoulos et al. [5] is evaluated using both real and synthetic data, our model is only evaluated using synthetic data. Therefore results reported from the comparison of our different methods may not be directly comparable. Future work could focus on evaluating our model with real-world data for a more valid comparison between techniques.

Our model is not perfect. While, the default hyperparameters chosen result in a model with improved performance, this model can also "hallucinate" false information. Additionally, the model occasionally returned output which was deemed to be overly "chat-like" and inappropriate for the context. However, these problems are present in past models also, marking the need for future research to address the same before AI-based summarization tools are deployed in critical areas, such as healthcare [11].

There exists no standard evaluation metrics for synthetically generated text [21]. We chose the same metrics reported in past work, such as ROUGE and BERTScore. However, these metrics are not guaranteed to convey a meaningful information in all scenarios. Additionally, metrics such as BERTScore can suffer from `truncation`, wherein input longer than the token limit is discarded in the evaluation process. We did not subdivide our model output into sections as in the original paper, as there is a loss of context in this process. Therefore, in this paper we chose to focus our findings mainly on the ROUGE score. However, future research is needed to establish more meaningful evaluation metrics for long-text summarization tasks, while also remaining true to human opinion.

V. Conclusions

In this paper, we presented a hyperparameter tuning approach to medical dialogue summarization using GPT-3.5 and the OpenAI API. We compare both manual and default hyperparameter tuning methods and find that the default parameters are optimally positioned within the context of both the loss function and subsequent ROUGE and BERTScore evaluation metrics. Our model outperforms standard GPT-3.5 and GPT-4 models. We reported evaluation metrics used in previous work to clearly show the potential benefits of our method. Our model is based on the GPT-3.5 architecture and exhibits higher scores across metrics when compared against standard GPT-3.5 and GPT-4 models. In order to avoid unnecessarily large language model usage for specific tasks, fine-tuning can be used to create the optimal balance of efficiency and accuracy in future medical dialogue summarization systems.

Notes

The GitHub repository housing all code related to this study can be found via the following link: https://github.com/isabel-ronan/gpt-tuning.

References

[1] L. J. Finney Rutten, K. D. Blake, A. J. Greenberg-Worisek, S. V. Allen, R. P. Moser, and B. W. Hesse, "Online Health Information Seeking Among US Adults: Measuring Progress Toward a Healthy People 2020 Objective," *Public Health Reports (Washington, D.C.: 1974)*, vol. 134, no. 6, pp. 617–625, 2019.

[2] Y. Liu, S. Ju, and J. Wang, "Exploring the potential of ChatGPT in medical dialogue summarization: a study on consistency with human preferences," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 75, Mar. 2024. [Online]. Available: https://doi.org/10.1186/s12911-024-02481-8

[3] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A. S. Chaudhari, "Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts," *Research Square*, pp. rs.3.rs–3 483 777, Oct. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10635391/

[4] X. Tang, A. Tran, J. Tan, and M. Gerstein, "GersteinLab at MEDIQA-Chat 2023: Clinical Note Summarization from Doctor-Patient Conversations through Fine-tuning and In-context Learning," May 2023, arXiv:2305.05001 [cs]. [Online]. Available: http://arxiv.org/abs/2305.05001

[5] G. Michalopoulos, K. Williams, G. Singh, and T. Lin, "MedicalSum: A Guided Clinical Abstractive Summarization Model for Generating Medical Reports from Patient-Doctor Conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4741–4749. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.349

[6] D. F. Navarro, M. Dras, and S. Berkovsky, "Few-shot fine-tuning SOTA summarization models for medical dialogues," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, D. Ippolito, L. H. Li, M. L. Pacheco, D. Chen, and N. Xue, Eds. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, Jul. 2022, pp. 254–266. [Online]. Available: https://aclanthology.org/2022.naacl-srw.32

[7] K. Suri, S. Saha, and A. Singh, "HealthMavericks@MEDIQA-Chat 2023: Benchmarking different Transformer based models for Clinical Dialogue Summarization," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, and A. Rumshisky, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 472–489. [Online]. Available: https://aclanthology.org/2023.clinicalnlp-1.50

[8] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," *Natural Language Processing Journal*, vol. 6, p. 100048, Mar. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949719123000456

[9] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback."

[10] "OpenAI API." [Online]. Available: https://openai.com/index/openai-api/

[11] W.-w. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, and M. Yetisgen, "Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation," *Scientific Data*, vol. 10, no. 1, p. 586, Sep. 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-023-02487-3

[12] "clinical_visit_note_summarization_corpus," Jul. 2024. [Online]. Available: https://github.com/microsoft/clinical_visit_note_summarization_corpus

[13] "OpenAI Platform." [Online]. Available: https://platform.openai.com

[14] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models," Dec. 2023, arXiv:2303.10420 [cs]. [Online]. Available: http://arxiv.org/abs/2303.10420

[15] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[16] G. LLC, "rouge-score: Pure python implementation of ROUGE-1.5.5." [Online]. Available: https://github.com/google-research/google-research/tree/master/rouge

[17] Artzi, Yoav, T. Zhang, V. Kishore, F. Wu, and K. Q. Weinberger, "bert-score: PyTorch implementation of BERT score." [Online]. Available: https://github.com/Tiiiger/bert_score

[18] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 2020, arXiv:1904.09675 [cs]. [Online]. Available: http://arxiv.org/abs/1904.09675

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," Jul. 2020, arXiv:2005.14165 [cs]. [Online]. Available: http://arxiv.org/abs/2005.14165

[20] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A. S. Chaudhari, "Adapted large language models can outperform medical experts in clinical text summarization," *Nature Medicine*, vol. 30, no. 4, pp. 1134–1142, Apr. 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41591-024-02855-5

[21] A. Figueira and B. Vaz, "Survey on Synthetic Data Generation, Evaluation Methods and GANs," *Mathematics*, vol. 10, no. 15, p. 2733, Jan. 2022, number: 15 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2227-7390/10/15/2733