

Ensemble Model Building for Medical Dialog Applications

Henry Gardner

henrygardner@berkeley.edu

Abstract

This paper presents a natural language processing approach to building a medical chat-bot based on real patient and doctor conversations. Utilizing these real conversations, I aim to build a generative model that can return accurate and informative medical advice to the everyday person. Although this is not aimed to replace a doctor or a medical professional, this research proves that language models can be tuned to further the progress of medical assistance through natural language processing.

1. Introduction

The integration of natural language processing (NLP) into healthcare has accelerated the development of intelligent systems capable of supporting both patients and medical professionals. Conversational AI models are increasingly being applied to simulate patient-doctor interactions, providing accessible health information and assisting with preliminary assessments. Despite these advancements, achieving a well-rounded and medically accurate response remains a significant challenge, particularly when balancing the generalization power of large language models (LLMs) with the specificity required in healthcare applications(8).

This research explores the effectiveness of 3 LLMs: GPT-2, LLaMA 3.2-3B, and GPT-3.5 Turbo, fine-tuned on a dataset consisting of real patient and doctor interactions. To enhance the relevance and contextual accuracy of generated responses, I integrated a BERT-based retrieval model, leveraging its strength in identifying similar conversational contexts. The final ensemble model combines the fine-tuned GPT-3.5 Turbo’s generative capabilities with BERT’s recommendation outputs, aiming to produce responses that are not only informative but also contextually aligned with previous medical conversations.

The study evaluates the models on key NLP metrics, such as the METEOR Score, BERTScore, Cosine Similarity, etc., while discussing the practical implications of deploying such models in healthcare applications. The project will be successful if these key metrics improve with the fine-tuning and ensemble approaches. The goal for this research

is to be deployed for world-wide accessibility, instead of being a tool just for medical professionals. Because of this, I created a virtual avatar and model API to run as if talking to a medical professional. I emphasize that these systems are not substitutes for professional medical advice, rather serve as tools for improving access to preliminary healthcare information. Always consult with a medical professional with any health related concerns.

2. Background

With the rapid advancements in large language models (LLMs) (11), significant research efforts have been directed toward their applications in the medical domain (8). Among these, Google introduced Med-PaLM and Med-PaLM 2, medical-specific models extensively studied through fine-tuning and ensemble modeling techniques. Leveraging ensemble refinement and conducting rigorous human evaluations with both physicians and laypeople, Med-PaLM 2 achieved state-of-the-art performance on various medical benchmarks, including MedQA and PubMedQA, by enhancing medical reasoning and aligning responses with clinical best practices (13).

Research into efficient fine-tuning techniques has shown that scaling transformer-based models can yield significant improvements when tailored to their size and architecture (14).

Accordingly, this study selected three transformer-based models of small, medium, and large sizes, each fine-tuned using strategies optimized for their respective scales. The goal is to build off the fine-tuning research and apply a different perspective of Google’s ensemble modeling to further medical interactions with language models.

3. Data

This research utilized a subset of the MedDialog dataset, comprising approximately 24,000 anonymized healthcare dialogues sourced from online doctor platforms such as Icliniq, HealthcareMagic, and HealthTap (7). Each entry in the dataset represents a unique interaction between a patient and a doctor, formatted as a patient question followed by a medical professional’s response. An example of such an interaction is illustrated in figure 1. As shown in figure 2,

the distribution of text and character lengths reveals that patient questions, averaging 38.4 words, are typically shorter than doctor responses, which average 101.83 words.

For preprocessing, formatting inconsistencies and unexpected interaction diction were retained to ensure the model could handle messy user input. The dataset was split into training (80%), testing (10%), and validation (10%) subsets for effective model evaluation. All text was tokenized using the corresponding tokenizer for each model to ensure input compatibility (11).

Although referenced in the Limitations section, it is important to acknowledge that since the dataset consists of online medical consultations, it might not fully represent the diverse healthcare contexts of global patient queries. This serves as a start to prove the applications of natural language processing for medical applications.

4. Methods

This research used a hybrid approach, combining both retrieval-based and generative models to maximize the accuracy, relevance, and contextual appropriateness of responses in patient-doctor conversations. Unlike Google’s approach, which focuses on evaluating medical-specific language models using multiple-choice-style medical tests (13), this project aims to adopt a more conversational methodology for medical Q&A. To achieve this, widely available language models with existing conversational training were selected and fine-tuned specifically for the medical domain. Then, utilizing a BERT model to feed in similar conversations as context, the model’s output could significantly improve.

BERT (Bidirectional Encoder Representations from Transformers) was selected as the recommendation model to retrieve similar historical patient-doctor interactions. Its strong contextual understanding and ability to model bidirectional context make it ideal for identifying relevant responses in healthcare dialogues, as apposed to a standard approach such as base cosine similarity.

Generative models were selected to address the task of producing fluent, coherent, and contextually appropriate responses based on patient queries. Each model was fine-tuned to specialize in medical dialogue generation.

GPT-2 was chosen for its lightweight architecture to provide a baseline for generative performance against more advanced models. LLaMA 3.2-3B was selected as a cutting-edge open-source medium-sized model to balance the computational efficiency and response quality in the specific healthcare domain. Finally, GPT-3.5 Turbo was selected due to its large model size, offering more contextually aware responses.

A unique contribution of this study is the hybrid approach, combining the strengths of BERT’s retrieval capabilities with the generative power of fine-tuned GPT-

3.5 Turbo. This retrieval-augmented generation (RAG) approach would allow for further context in aiding the model’s response(5). As further discussed in the Limitations section, this should theoretically lower the times the model needs to be re-trained with new medical research being released. BERT ensures the generative model starts with high-quality, relevant input, improving response accuracy.

4.1. Fine-Tuning

The approach to fine-tuning was different for each of the generation models due to their distinct architectures(14). For speed, all the open-source models were fine-tuned with a NVIDIA GeForce RTX 4070 GPU.

For GPT-2, the training process utilized a traditional approach, where the model was fine-tuned using a standard AdamW optimizer and gradient accumulation steps to manage large batch sizes. The focus was on efficient training through gradient scaling and a learning rate scheduler to gradually decrease the learning rate during training (11). Additionally, a checkpoint system was implemented to save model states periodically, and memory management techniques were applied, including gradient clipping and automatic mixed precision (AMP) to optimize GPU memory usage. The initial experiments showed that a simple generative model like GPT-2 provided a useful baseline but lacked the contextual depth needed for effective medical dialogue. This is evident in Google’s research as well as their model needed to grow to Med-PaLM 2 before they were able to hit their impressive benchmarks.

For the larger models, full tuning on this scale is not doable in a reasonable amount of time, as discovered when attempting a similar approach as GPT-2. Therefore parameter-efficient fine-tuning would offer a different approach to fine-tuning with similar response advancements(3). The LLaMA model leveraged a LoRA (Low-Rank Adaptation) approach, which introduces a more memory-efficient fine-tuning technique by adapting only a small subset of the model’s parameters, rather than updating all of them(6). This was especially crucial given the large size of the LLaMA model - 3 billion parameters, leaving the training parameters to only 2,293,760. LoRA’s low-rank decomposition was applied specifically to the query and value projections in the attention layers, which are key to model performance. The optimizer for LLaMA used a custom configuration to handle the low-memory setup and involved more advanced GPU memory management, such as dynamically clearing the cache, ensuring smooth training with limited resources.

Finally, the OpenAI model GPT3.5-Turbo was fine-tuned through OpenAI’s API, making the fine-tuning process significantly faster and easier. The difficulty came in the setup of the data for training preparation, as OpenAI is very particular about the format. For this approach, I

Patient Query:

“Can I go out in the sun while taking 875 mg dose of amoxicillin?”

Doctor Response:

“Amoxicillin is one of the antibiotics that is not sun sensitive.”

Figure 1. One example of a patient query and doctor response.

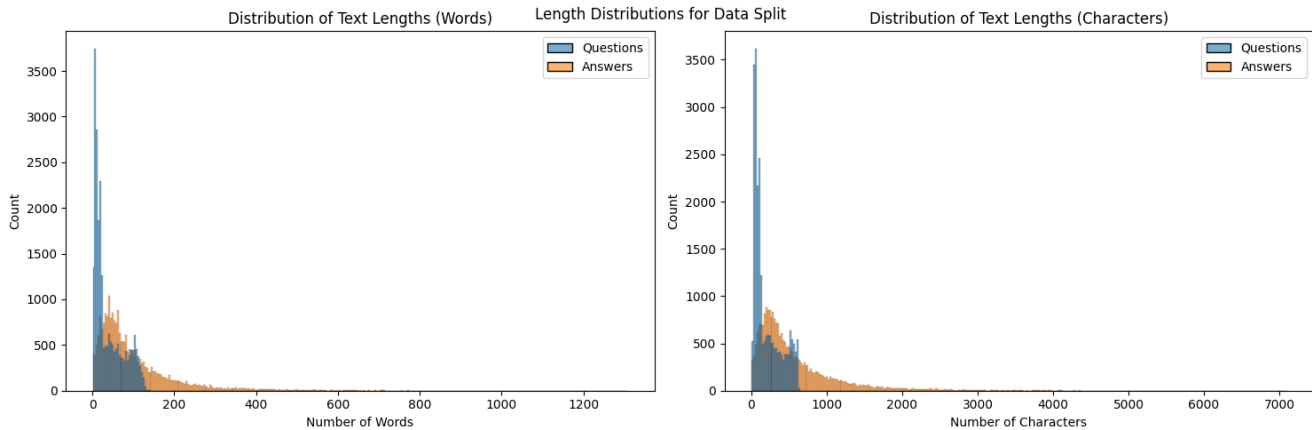


Figure 2. Distribution of text and character lengths split by both patient questions and doctor responses.

used the default hyperparameters that OpenAI suggested. The only customization for this fine-tuning process was the customization of the contexts. Although I didn’t generate a model to update these prompts in the training process, the following paper(10) proves the importance of prompting to help frozen language models. I played with different context roles in the training process with an even split of the following 4 messages: “You are a medical expert trained on giving advice to patients that are in need of assistance”, “You are an empathetic medical expert here to help with health concerns”, “You are a medical expert particularly experienced with interacting with patients and guiding them through their concerns”, and “You are a medical expert tasked with conversing with patients and helping them discover solutions to their concerns.” The single fine-tuning job was just under \$30.

Of all the training, the most unusual loss and accuracy metrics came from the fine-tuning of the GPT-3.5 Turbo model, for information on the other models refer to the project repository. The graph of training accuracy, shown in figure 3, shows significant fluctuations over the training iterations, ranging roughly between 0.45 and 0.75. The overall trend is fairly stable, with the mean (red line) hovering around 0.57 and the median (green line) close to the mean. The large variance in accuracy suggests that the model’s predictions vary widely between batches. This could indicate that the dataset contains a high degree of variability, or that the batches themselves are diverse in difficulty, caus-

ing the accuracy to swing from one batch to another. An unusual observation is that the accuracy does not show a clear upward trend over time. This might suggest either a learning rate that is too high, causing instability and preventing consistent improvement, or the model has potentially reached a plateau in learning. It could further mean that it is struggling to generalize on the training data.

The graph of training loss also exhibits considerable oscillations, with values ranging approximately between 1.0 and 3.0, shown in figure 4. However, unlike the accuracy graph, there is a slight downward trend in loss initially, followed by more stable oscillations around a mean (red line) of approximately 1.8 and a median (green line) very close to the mean. An interesting observation here is that while the loss decreased early on, it did not show significant further improvement, suggesting that the model quickly converged to a local minimum. The persistent oscillations could be due to batch variability, where different batches contain varying levels of difficulty or imbalance, or the use of a large or inconsistent learning rate, leading to fluctuating gradients. A potential need for further fine-tuning of hyperparameters like learning rate, batch size, or optimizer type could be evident. This means that the hyperparameters suggested from OpenAI would need re-thinking in order to see improvement. The training instability in the GPT-3.5 Turbo model can be further seen in the Results section.

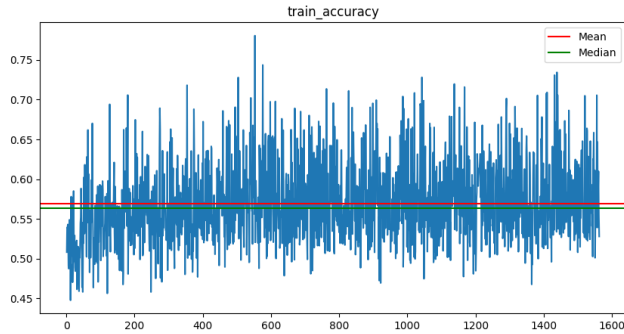


Figure 3. Fine-tuning training accuracy for the OpenAI GPT-3.5 Turbo model.

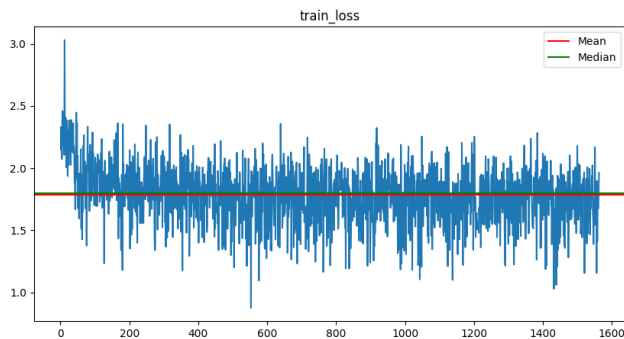


Figure 4. Fine-tuning training loss for the OpenAI GPT-3.5 Turbo model.

4.2. Ensemble Model Building

Following fine-tuning, the ensemble model was constructed. Evident from the limitations of GPT-2 fine-tuning, this hybrid approach was used for contextual grounding when leveraging generation models. The process was relatively straightforward: a test example or user query was first processed by the BERT model to retrieve highly similar conversations from the training data, based on the similarity scores that BERT would generate. Although modifiable, this was producing the 3 most relevant conversations. These retrieved conversations were then used as additional context for the fine-tuned GPT-3.5 Turbo model, which generated a response to the user query. This integration aimed to enhance response quality by leveraging the contextual information provided by BERT.

5. Results and Discussion

The performance of each model was evaluated using key metrics to assess their ability to generate contextually accurate, fluent, and relevant responses in patient-doctor dialogues. The evaluation focused on metrics that emphasize semantic alignment, language fluency, and contextual coherence.

BERTScore was selected for its ability to measure the similarity between generated responses and references using contextual embeddings rather than relying on exact word matching. This approach is particularly valuable in healthcare dialogues, where semantic accuracy is more important than lexical precision. By evaluating precision, recall, and F1, BERTScore provided a comprehensive view of how well the models captured both the accuracy and completeness of the information in the dialogues (15).

METEOR was included to capture semantic matching beyond surface-level similarity. Its emphasis on synonymy, stemming, and paraphrasing made it particularly useful for evaluating the flexibility and natural variation in language expression, which are essential in healthcare communication. This metric highlighted the models' ability to produce responses that conveyed the intended meaning even when phrasing differed from the reference (1).

ROUGE metrics offered insight into the overlap of n-grams between the generated and reference responses, focusing on both individual word matching and longer sequences. ROUGE-1, ROUGE-2, and ROUGE-L were used to assess how closely the models replicated key terms and content structures, providing a more surface-level evaluation of the generated text. However, in open-ended dialogue tasks, these metrics often revealed the trade-off between exact replication and the need for varied, natural language output (2).

Cosine similarity provided an overall measure of semantic alignment by comparing the embeddings of the generated and reference responses. This metric was particularly important for understanding how closely the meaning of the generated text aligned with the reference, regardless of the specific words used. It served as a strong indicator of the models' ability to maintain consistency in meaning while allowing for flexibility in language (4).

Overall, the results demonstrated that semantic metrics like BERTScore and cosine similarity were the most relevant for evaluating generative models in the context of healthcare dialogues. These metrics emphasized the importance of producing responses that were contextually appropriate and semantically accurate, rather than focusing solely on exact word overlap. The ROUGE and METEOR scores, while informative, underscored the challenges of capturing exact n-gram sequences or synonyms in open-ended generation tasks, reinforcing the need to prioritize semantic coherence. Additional metrics and detailed results are available in the project's repository for further reference.

The median metrics for each of the generation models can be found in tables 1 and 2 for the baseline and fine-tuned results, respectively.

The results reveal significant improvements across key metrics following fine-tuning, highlighting the impact of task-specific training on the generation models' perfor-

mance. METEOR shows notable gains for each model, with LLaMA 3.2-3B exhibiting the most substantial increase from 0.0925 to 0.2114. This indicates that fine-tuning enhanced the models' ability to capture nuanced variations in language. It is important to note, however, the overall low scores for each model in the METEOR category. This is typically because of the heavily weighted lexical choices that METEOR emphasizes. Meaning, that the low score indicates low word-level exact matches or synonyms. This is okay as long as the overall meaning (described in the other evaluation metrics) stays the same.

BERTScore metrics also demonstrate consistent improvements. LLaMA 3.2-3B and GPT-2 show enhanced BERTScore scores, suggesting better retrieval of relevant information while maintaining precision. Notably, GPT-3.5 Turbo exhibited a minor decrease in each BERTScore metric post-fine-tuning. However, due to the size of the model, this suggests that refined hyperparameters are needed improve it.

ROUGE F1 scores, which capture surface-level n-gram overlap, improved for all models, with LLaMA 3.2-3B showing the most significant increase from 0.1663 to 0.2655. This suggests that fine-tuning contributed to better lexical similarity, even in an open-ended dialogue setting where exact matches are less common.

Cosine similarity, which assesses semantic coherence, exhibited the most substantial improvement for LLaMA 3.2-3B, increasing from 0.9204 to 0.9544. This highlights the model's enhanced ability to generate responses that align closely with the reference embeddings. This is evident in the initial LoRA paper as well, indicating the success of the fine-tuning approach (6). Interestingly, the ensemble model, combining BERT retrieval with GPT-3.5 Turbo fine-tuned, demonstrated a minor increase in overall performance, suggesting integration of retrieval-based and generative approaches(5). Although it was not as significant as expected, improving the fine-tuning of such a large model could offer better results.

Fine-tuning not only improved the models' ability to generate contextually appropriate and semantically rich responses but also showcased the versatility of combining retrieval-based and generative approaches in the ensemble model, refer to figure 5 for an example output. Looking through all the metrics analyzed, it is clear that the LLaMA 3.2-3B fine-tuned model was the best at delivering consistent and effective responses and showed the most improvement upon tuning. More model output examples and full metric details can be found in the project repository, attached at the end.

6. AI Avatar

To make this research accessible to a global audience, merely presenting model outputs and statistics was insuffi-

cient. Therefore, I created a website featuring an AI avatar powered by the fine-tuned GPT-3.5 Turbo model, chosen for its integration via the OpenAI API. The avatar, developed using HeyGen technology, presents a virtual persona for interactive conversations. A demo showcasing my interaction with the avatar, where I discuss concerns about a sore throat, is available in the project repository. Additionally, figure 6 includes a screenshot of the interactive platform.

7. Limitations

While the model results demonstrate promising improvements following fine-tuning, there are several limitations to consider. First, the dataset used in this study, though diverse, was curated from specific sources such as Icliniq, HealthcareMagic, and HealthTap. As a result, the models may not generalize well to conversations from other healthcare platforms or less formal patient-doctor interactions, limiting the external validity of the findings.

Another limitation lies in the reliance of the evaluation metrics. Although these metrics provide valuable insights into surface-level similarity and semantic alignment, they may not fully capture the nuances of healthcare dialogues, such as the importance of actionable advice, which is crucial in real-world patient-doctor interactions. Furthermore, the models are assessed on their ability to generate contextually relevant responses, but the absence of patient-specific factors, such as emotional state or medical history, might hinder their ability to provide truly personalized feedback.

While fine-tuning the models improved their performance on the testing data, it is unclear how they would perform on real-world, out-of-distribution data. The dataset does not include disease labels or medical expertise annotations, which means the models may struggle with providing medically accurate advice, making them unsuitable for critical healthcare applications.

Lastly, with new research, the models will need to be re-trained to give the most up-to-date accurate advice. Although the additional computational cost required for re-training these models may limit their scalability in real-time applications, especially in environments with resource constraints. Although the ensemble model seeks to address this, by feeding in relevant information as context to the model, until that information would be non-conversational and include medical facts, re-training would be needed.

8. Next Steps

Improvement of each of the generation models can come with a grid-search over the hyperparameters, tweaking the temperature output parameter to be more or less creative, collecting more data, including training on medical research outside of just conversations, prompt engineering for the context/input to the model, and a medical fact checker

Model	METEOR	BS Precision	BS Recall	BS F1	ROGUE F1	Cosine Similarity
GPT-2	0.0830	0.8503	0.8174	0.8340	0.1429	0.8962
LLaMA 3.2-3B	0.0925	0.8355	0.8156	0.8254	0.1663	0.9204
GPT-3.5 Turbo	0.1533	0.8536	0.8380	0.8434	0.2313	0.9252

Table 1. Median baseline metrics on each of the generation models. Note that BS stands for BERTScore. Additionally, the ensemble model was only run with the fine-tuned version of GPT-3.5 Turbo and therefore, there is no baseline version of it.

Model	METEOR	BS Precision	BS Recall	BS F1	ROGUE F1	Cosine Similarity
GPT-2	0.1241	0.8504	0.8293	0.8392	0.1956	0.9194
LLaMA 3.2-3B	0.2114	0.8379	0.8414	0.8412	0.2655	0.9544
GPT-3.5 Turbo	0.1795	0.8201	0.8356	0.8301	0.2173	0.9330
Ensemble	0.1894	0.8201	0.8357	0.8252	0.1997	0.9310

Table 2. Median fine-tuned metrics on each of the generation models. Note that BS stands for BERTScore.

through medical research in sites hosted online such as PubMed (12) to verify any information before any output.

The avatar was used as a human-like approach to conversation, and can be improved upon by training on real world medical professionals. With significant improvement, computer vision can be investigated to have the patient upload images or ask for analysis into medical details, visually. Recently an article showcasing generation of synthetic medical images to improve medical classifiers could prove useful for this type of advancement, as training a large medical classifier would be necessary(9).

9. Conclusion

This research explored the development of several natural language processing models, including BERT, GPT-2, LLaMA 3.2-3B, GPT-3.5 Turbo, and an ensemble approach, with the goal of generating accurate and contextually relevant responses based on real patient-doctor interactions. The goal was to enhance conversational AI in the healthcare domain, accessible by anyone with an internet connection. Each model was fine-tuned using a subset of the MedDialog dataset(7) and evaluated for semantic understanding to assess improvements in sentence-level comprehension. The model that showed the most improvement and offered consistently accurate responses was LLaMA 3.2-3B, improving in every metric significantly and resulting in an impressive average score of 0.9544 on cosine similarity. Additionally, although minimal improvements, the ensemble approach proved effective at generating responses with contextual hints at doctor responses to similar queries. The improvements needed to make this project more stable lie in the ability to measure and train medical research and facts, improving this approach from conversational to medically precise.

As stated before, this project does not seek to replace medical professionals, but aims to further the medical application of language models for preliminary medical conver-

sations usable by anyone! All of the code for the project and a demo showcasing the avatar can be found in the project repository.

References

- [1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [2] M. Barbella and G. Tortora. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317*, 2022.
- [3] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5:220–235, 2023.
- [4] D. Gunawan, C. Sembiring, and M. A. Budiman. The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of physics: conference series*, volume 978, page 012120. IOP Publishing, 2018.
- [5] S. Hofstätter, J. Chen, K. Raman, and H. Zamani. Fidlite: Efficient and effective retrieval-augmented text generation. *arXiv preprint arXiv:2209.14290*, 2022.
- [6] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [7] Kabatubare. Medical conversations dataset, 2023. <https://huggingface.co/datasets/Kabatubare/medical>.
- [8] M. Karabacak and K. Margetis. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15(5):e39305, May 2023.
- [9] I. Ktena, O. Wiles, I. Albuquerque, S.-A. Rebuffi, R. Tanno, A. G. Roy, S. Azizi, D. Belgrave, P. Kohli, T. Cemgil, A. Karthikesalingam, and S. Gowal. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30:1166–1173, 2024.
- [10] C. Li, M. Zhang, Q. Mei, W. Kong, and M. Bendersky. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference*. ACM, 2024.
- [11] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2024.
- [12] PubMed. Pubmed, 2024. <https://pubmed.ncbi.nlm.nih.gov/>.
- [13] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.
- [14] Y. Tay, M. Dehghani, J. Rao, W. Fedus, S. Abnar, H. W. Chung, S. Narang, D. Yogatama, A. Vaswani, and D. Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv:2109.10686*, January 2024.
- [15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

A. Additional Material

For more information on the results, metrics, and setup refer to the project repository’s notebook folder. An example output of the ensemble model can be seen in figure 5. A screenshot showing an interaction with the AI avatar yielding the model response can be seen in figure 6.

Patient Query:

“What is hypotension?”

Model Response:

“Hypotension is low blood pressure. Very low blood pressure to the point that not enough pressure to maintain blood flow to vital organs. There are different levels of low blood pressure; the most severe form is called shock.”

Expected Response:

“Hypotension means low blood pressure.”

Figure 5. Example output from an entry in the test set from the ensemble model - BERT and GPT-3.5 Turbo. The output is clearly aligned to the expected response, and even gives more detail than the doctor. Although it is possible that the originating model was trained on something similar, it is important to note the improving metrics that fine-tuning yielded.

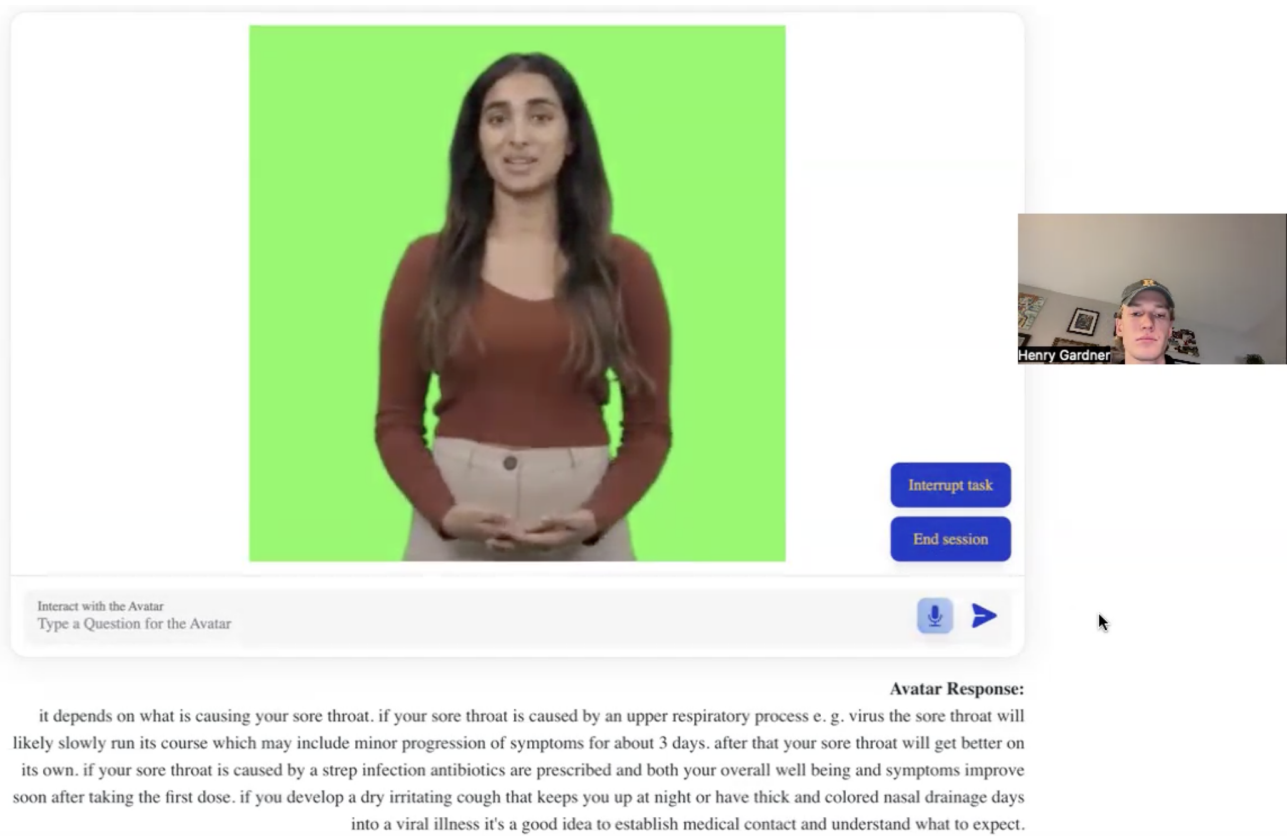


Figure 6. Screenshot showing the interactive avatar, where you can interact with the model, I asked it a question regarding what can happen if a sore throat gets worse. Below the avatar, you can see the response from the model.