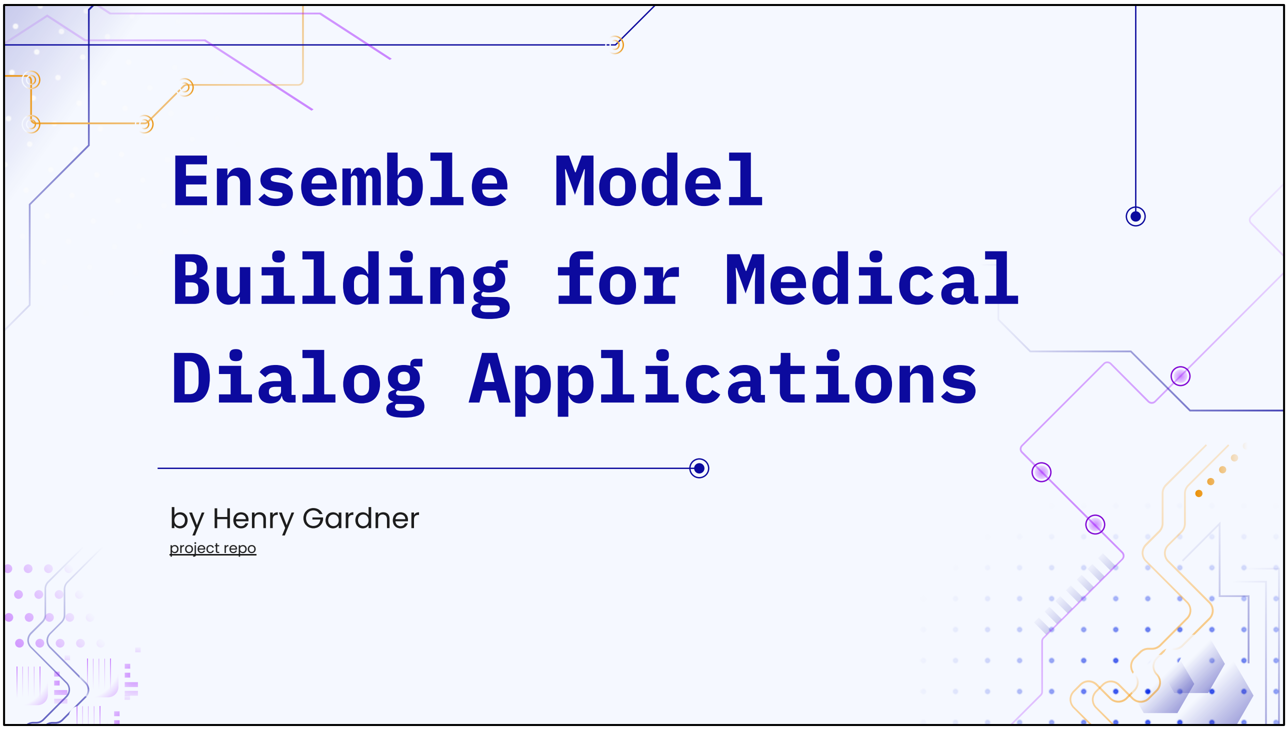
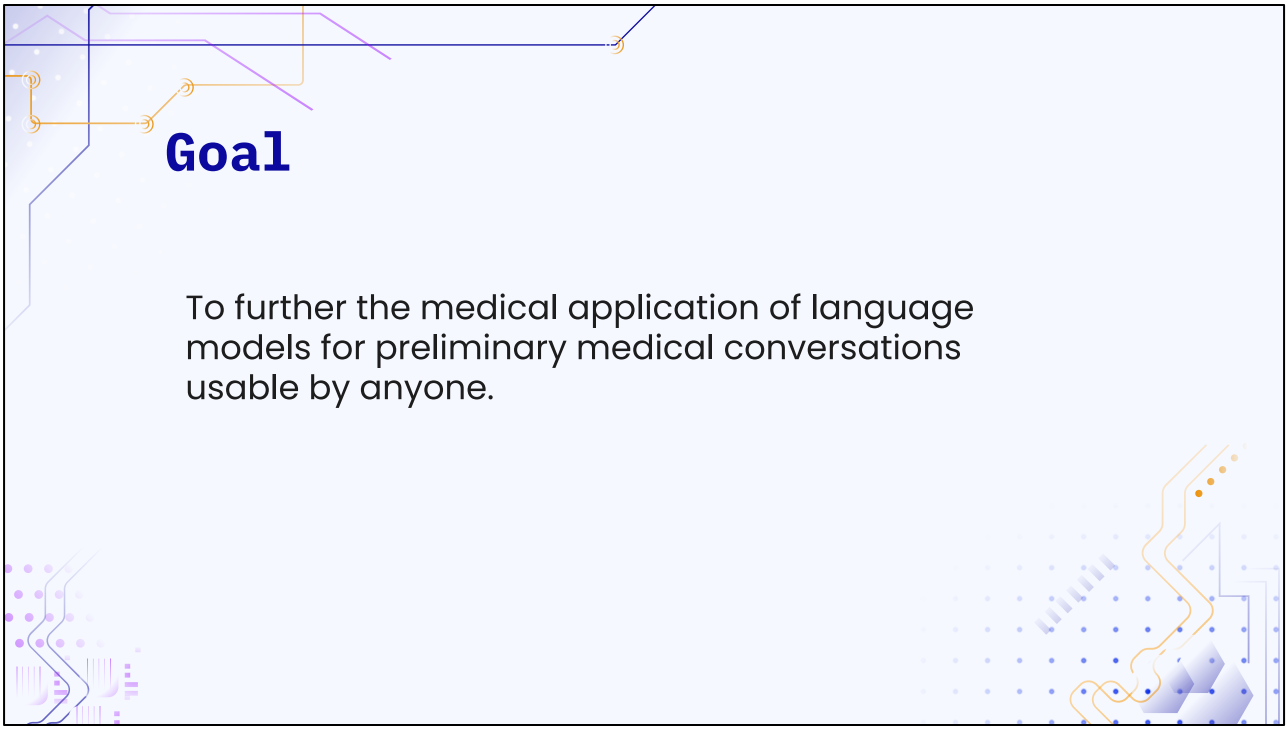


Ensemble Model Building for Medical Dialog Applications

by Henry Gardner

project repo

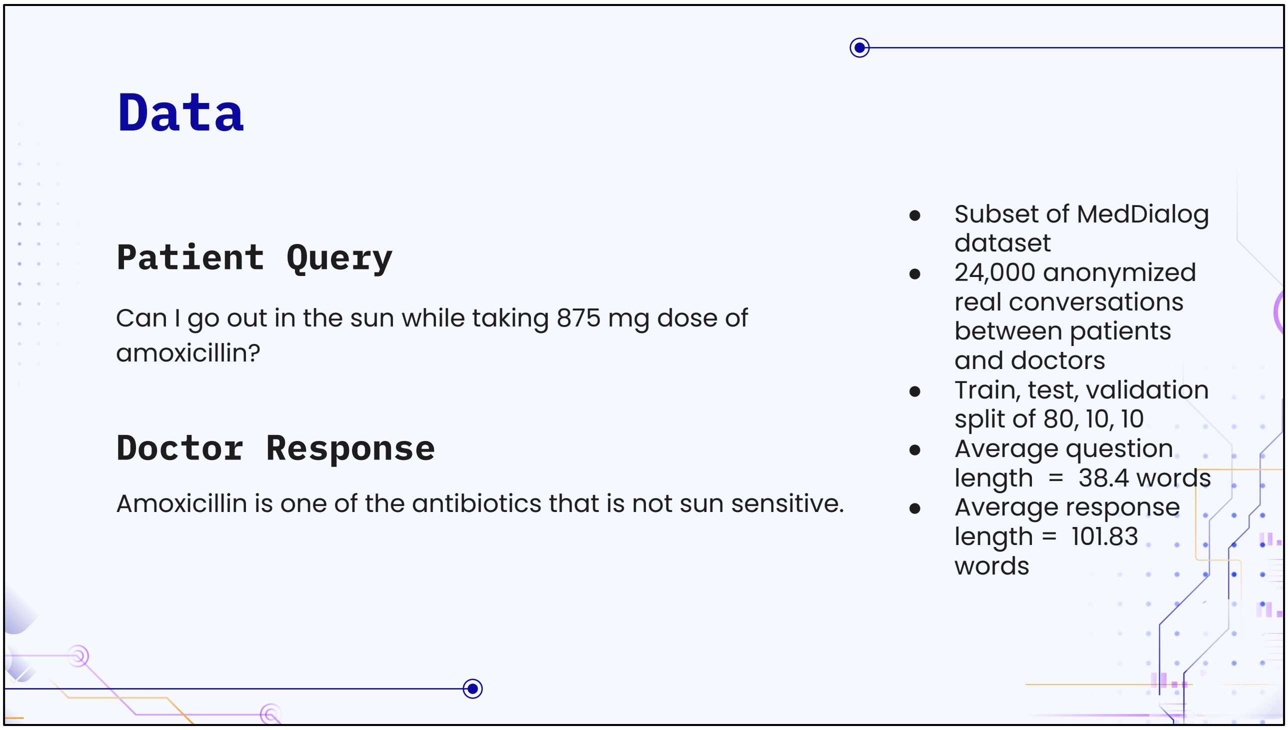
I know I only have 5 minutes, so I’ll try and keep this efficient as there is a lot of info to cover!



Goal

To further the medical application of language models for preliminary medical conversations usable by anyone.

The main goal for this project was to further the development of medical applications of language models for world-wide accessibility instead of just being a tool for medical professionals. These systems will not be substitutes for professional medical advice, rather serve as tools for improving access to preliminary healthcare information.



Data

Patient Query

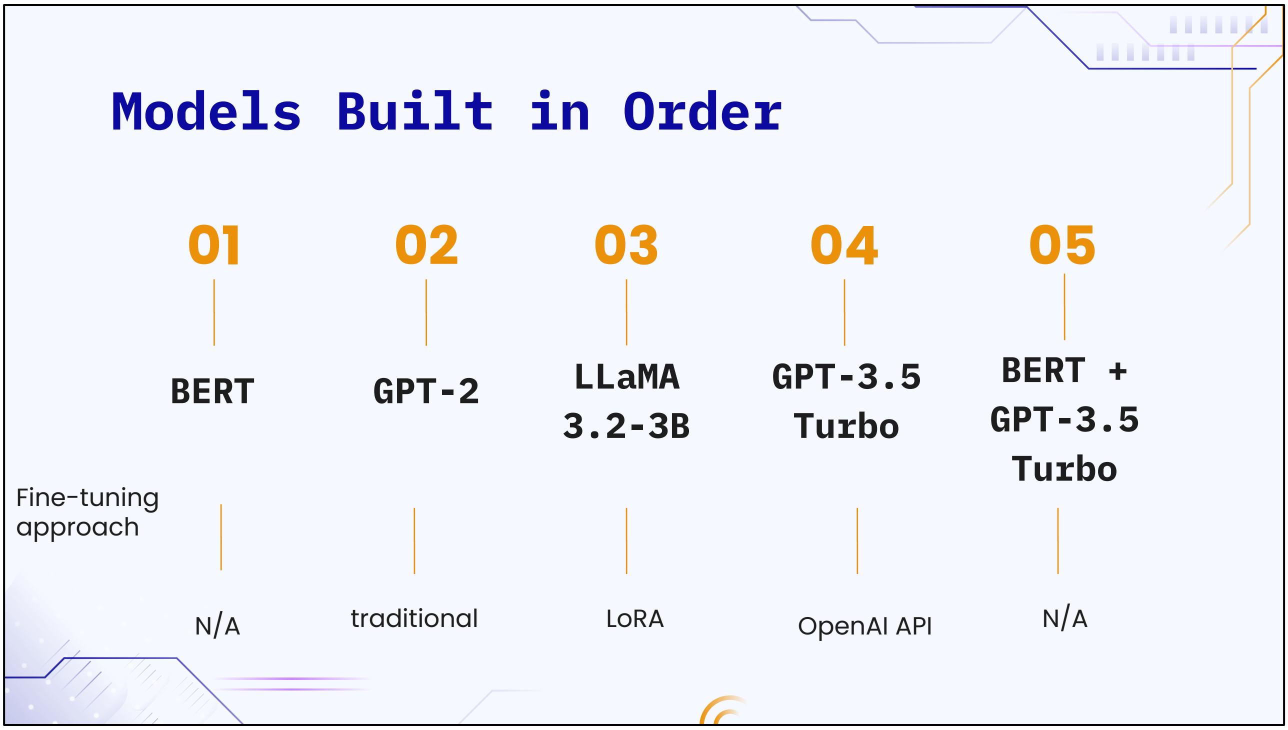
Can I go out in the sun while taking 875 mg dose of amoxicillin?

Doctor Response

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

- Subset of MedDialog dataset
- 24,000 anonymized real conversations between patients and doctors
- Train, test, validation split of 80, 10, 10
- Average question length = 38.4 words
- Average response length = 101.83 words

- The data included a subset of the MedDialog dataset, comprising approximately 24,000 anonymized healthcare dialogues sourced from online doctor platforms
- Each entry in the dataset represents a unique interaction between a patient and a doctor
- I created a 80, 10, 10 split for training, testing, and validation
- For preprocessing, grammar and word inconsistencies were kept to help the model work on unpredictable data, but all diction was tokenized



Models Built in Order

	01	02	03	04	05
	BERT	GPT-2	LLaMA 3.2-3B	GPT-3.5 Turbo	BERT + GPT-3.5 Turbo
Fine-tuning approach	N/A	traditional	LoRA	OpenAI API	N/A

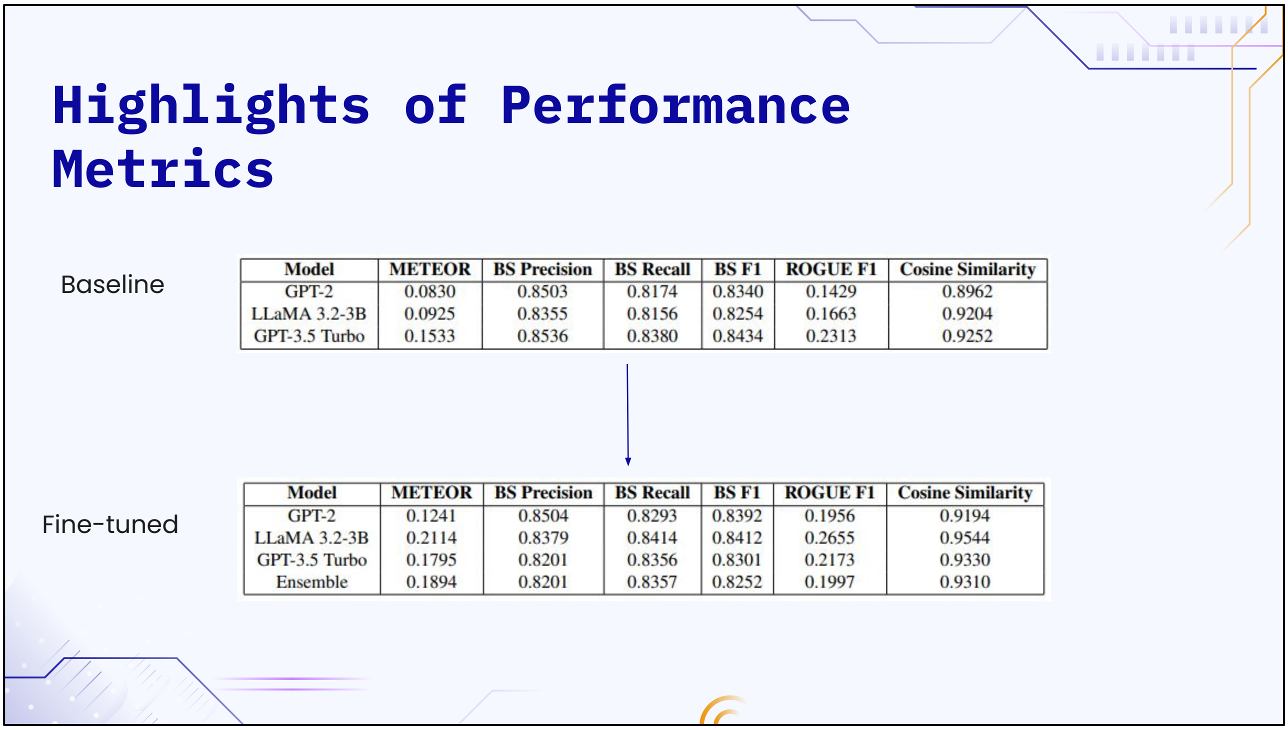
This project involved 3 generative models of sizes small, medium, and large: GPT2, LLaMA 3.2-3B, and GPT-3.5 Turbo

Additionally, I created what I am calling an ensemble model where I attached BERT to the GPT-3.5 Turbo model for a RAG implementation.

The approach to fine-tuning was different for each of the generation models due to the their distinct architecture.

For GPT-2 I did a traditional approach, LLaMA required a parameter-efficient approach so I chose LoRA due to the large number of parameters - 3 billion, and lastly GPT-3.5 turbo was fine-tuned through the OpenAI API, making it the easiest.

All open-source models were fine-tuned with a 4070 gpu.



Highlights of Performance Metrics

Baseline

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

Fine-tuned

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

My evaluation focused on metrics that emphasize semantic alignment, language fluency, and contextual coherence.

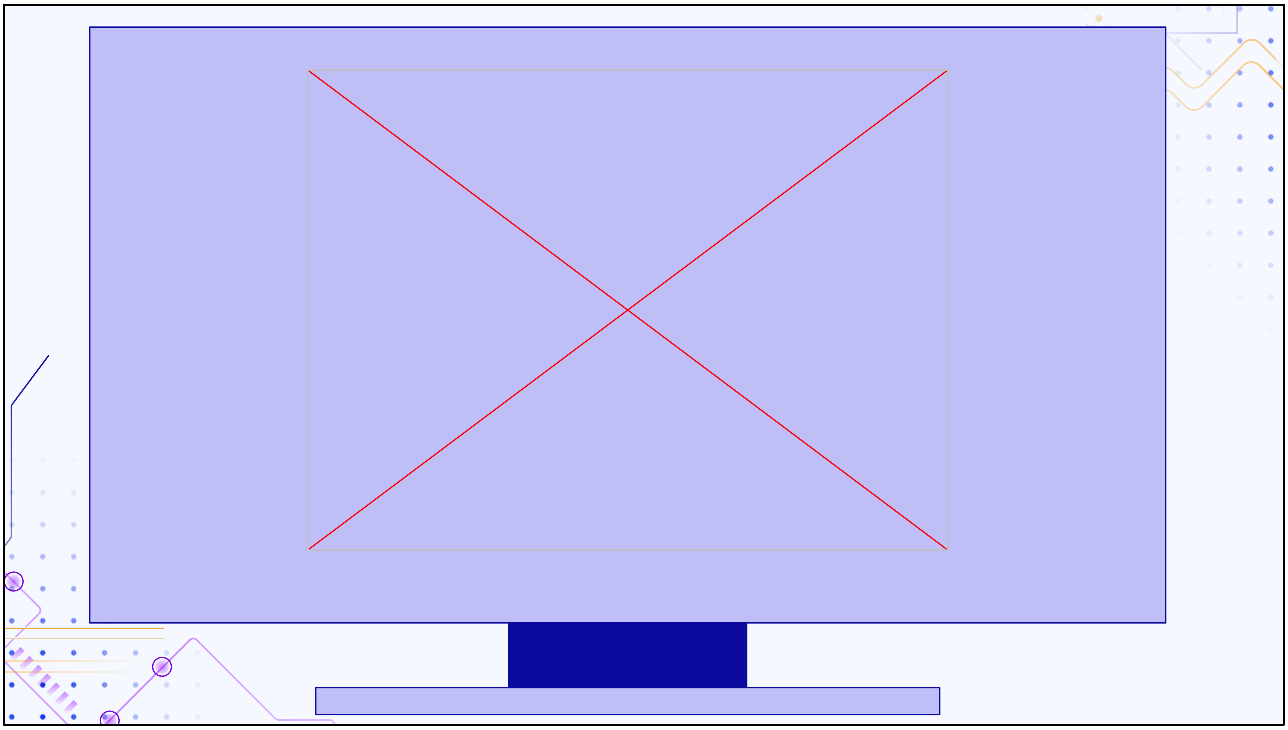
BERTScore was selected for its ability to measure the similarity between generated responses and references using contextual embeddings rather than relying on exact word matching

METEOR was included to capture semantic matching beyond surface-level similarity

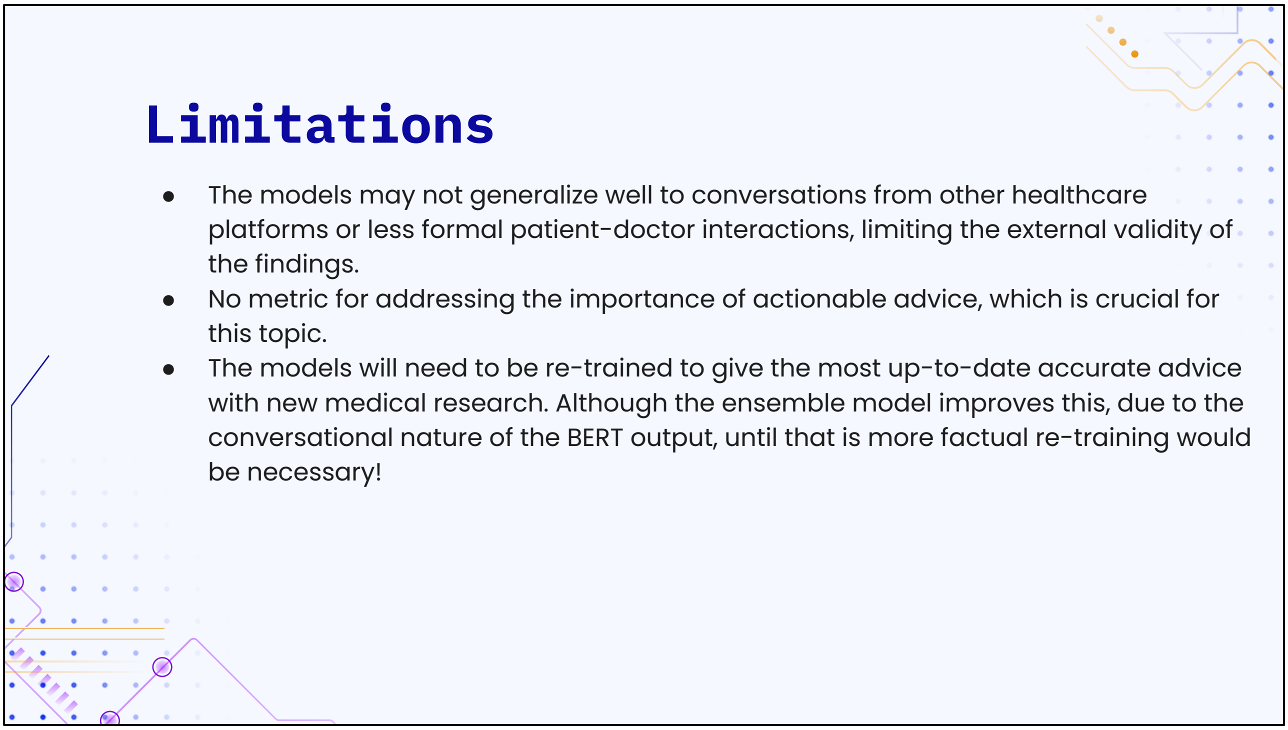
ROUGE metrics offered insight into the overlap of ngrams between the generated and reference responses, focusing on both individual word matching and longer sequences.

Cosine similarity provided an overall measure of semantic alignment by comparing the embeddings of the generated and reference responses

Due to time, I will only comment on the general findings showing that fine-tuning improved the metrics for each evaluator, where the most significant improvement came from LLaMA. The ensemble model did demonstrate improvements as well over the standard GPT3.5 turbo model, although those improvements were less than expected.



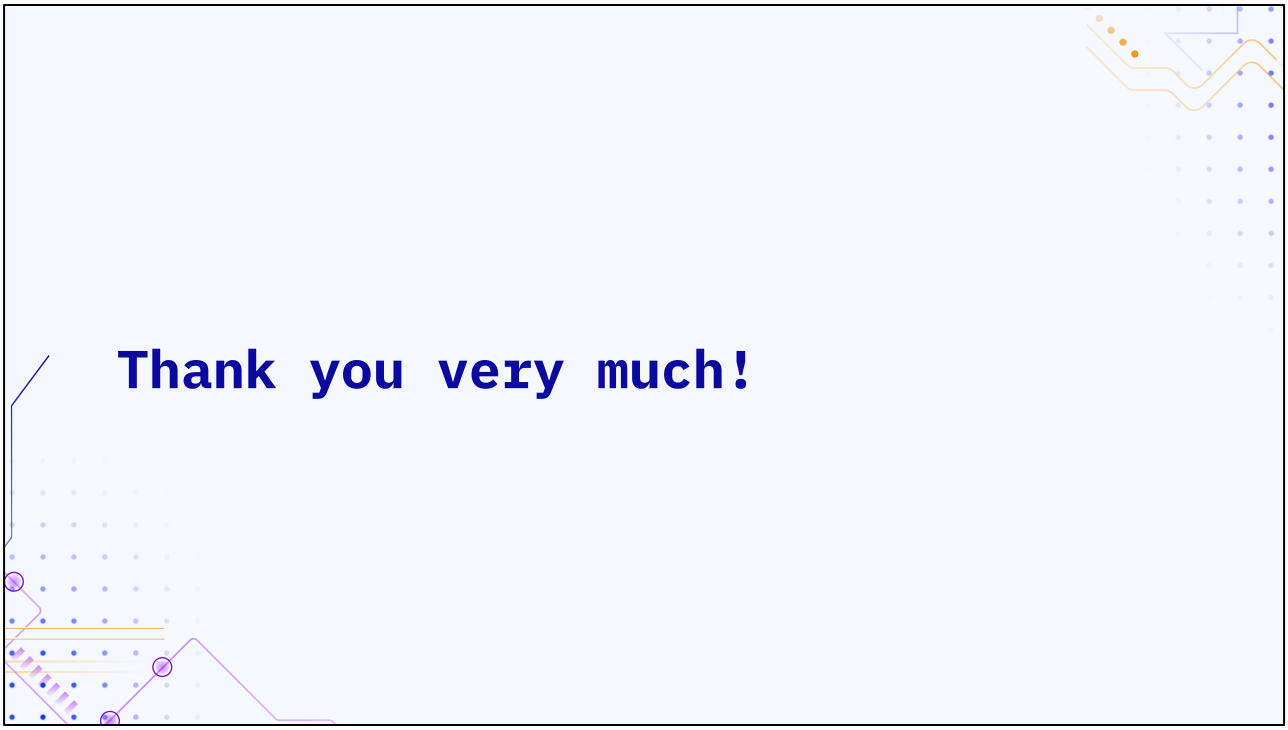
To make this research accessible to a global audience, merely presenting model outputs and statistics was insufficient. Therefore, I created an AI avatar and model API so you can interact with it. Here is a quick demonstration of it!



Limitations

- 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.
- No metric for addressing the importance of actionable advice, which is crucial for this topic.
- The models will need to be re-trained to give the most up-to-date accurate advice with new medical research. Although the ensemble model improves this, due to the conversational nature of the BERT output, until that is more factual re-training would be necessary!

Lastly, it is important to note that the models may not generalize well to outside conversations as these came from online platforms. There is also no metric addressing the importance of actionable advice, which would need to be developed to really test the validity of the outputs. And finally the models will need to be re-trained with new medical research.



Thank you very much!