



Ensemble Model Building for Medical Dialog Applications


by Henry Gardner

[project repo](#)

A decorative graphic in the top-left corner featuring a network of thin, intersecting lines in purple and orange. Some lines terminate in small circular nodes, resembling a circuit board or a data network diagram.

Goal

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

A decorative graphic in the bottom-right corner featuring a grid of small blue dots. Overlaid on this grid are several wavy, flowing lines in orange and purple, along with some geometric shapes like triangles and rectangles, creating a futuristic, digital aesthetic.

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

Models Built in Order

01

BERT

02

GPT-2

03

**LLaMA
3.2-3B**

04

**GPT-3.5
Turbo**

05

**BERT +
GPT-3.5
Turbo**

Fine-tuning
approach

N/A

traditional

LoRA

OpenAI API

N/A

Highlights of Performance Metrics

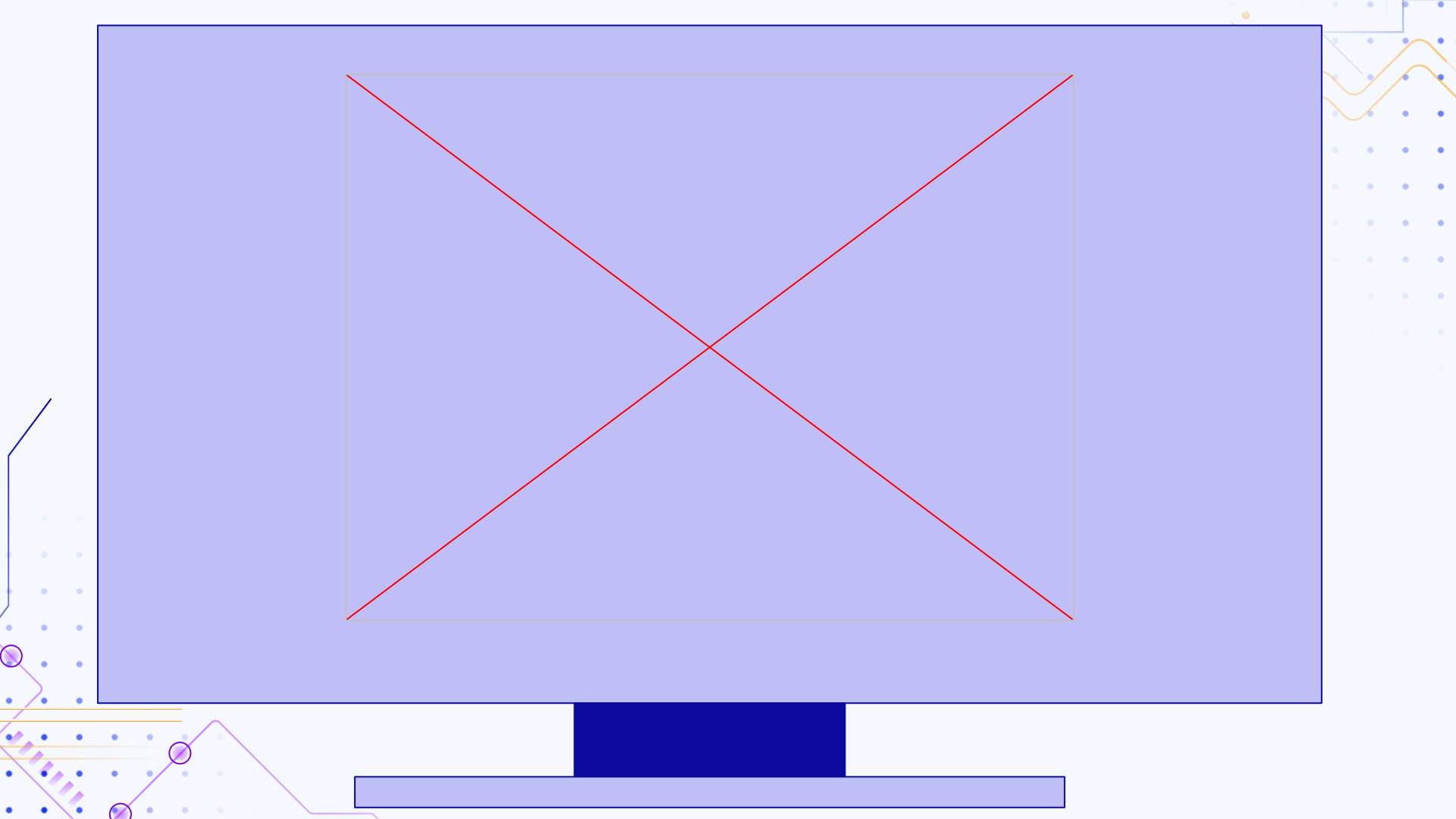
Baseline

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



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!



Thank you very much!

