An Analysis of Large Language Models in the HealthCare Domain

THE COLLEGE OF WOOSTER

Bolanle Oladeji | Computer Science | Advised by Kowshik Bhowmik

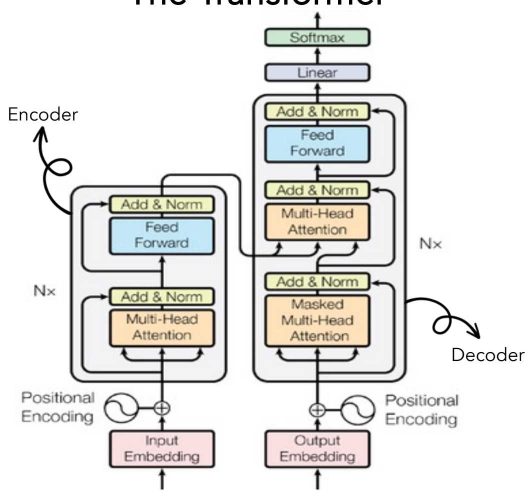
Research Problem

Healthcare workers are constantly being overworked due to inadequate resources (both personnel-wise and device-wise) and an overwhelmingly large number of patients. We explore the possibility of using Large Language Model (LLM) driven Conversational Agents, otherwise known as chatbots, as a tool to answer common medical questions and for global health provisioning. We also question the effectiveness of LLMs especially with regards to AI misinformation and bias.

Goals

- Design and implement AI-based medical chatbots using different types of Transformer models
- Evaluate the best model and build a front-end user interface
- Compare models' metrics (perplexity, BLEU score) and qualitatively analyze models' responses

The Transformer



Transformer Architecture

- Neural network architecture proposed by Vaswani et al. (2017) based on a concept called Attention [1].
  Attention is the concept of assigning more weights to specific parts of an input sequence.
- Transformer consists of two parts: the Encoder and the Decoder.
- Transformers form the basis of many models like OpenAI's GPT series and Google's BERT.

Transformer Based Models

DialoGPT: This model uses only the decoder-portion of the Transformer. DialoGPT is based on the GPT-2 architecture but is trained on conversational data gathered from Reddit.

T5: This model uses the standard encoder-decoder architecture of the original Transformer with only some slight vocabulary and functional changes.

BERT: BERT which stands for Bidirectional Encoder Representations from Transformers, uses the encoder stack of the Transformer with some modifications for language modelling.

Dataset

Models are fine-tuned on data sourced from four websites: eHealth Forum, iCliniq, Question Doctors, and WebMD up until May 2017 [2].

Dataset has attributes: "Question" (the medical based question), "Answer" (medical expert's answer), and "Context."

Methodology

All models are fine-tuned for question-answering downstream task with the medical using the same hyper parameters for comparative purposes. We evaluate our model on the following metrics:

$$BLEU = BP * exp(\sum_{k=1}^{n} w_k log(p_k))$$

$$BP = e^{min(1-\frac{len(reference)}{len(prediction)},0)}$$

$$\text{Perplexity}(M) = M(s)^{-1/n} = \sqrt[n]{\prod_{k=1}^{n} \frac{1}{M(w_k|w_0 w_1 \cdots w_{k-1})}}$$

Perplexity of a language model M

- The higher the BLEU Score the better the model
- The lower the perplexity the better the model

Results

DialoGPT	Perplexity 5.82	BLEU Score: 0.352
T5	Perplexity 8.58	BLEU Score: 0.722

Table 7.4: Model Ratings

Model	Minimum (Points)	Maximum (Points)	Mean (Points)
DialoGPT	1.00	5.00	3.77
T5	3.00	5.00	3.80
BERT	1.00	5.00	3.60

Table 7.5: Preference of Model

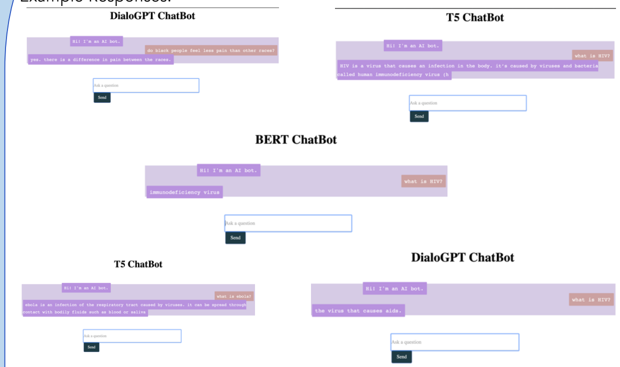
Model	Percentage of Preference
DialoGPT	60%
T5	40%

Table 7.3: Naturalness of the Models

Model	Minimum	Maximum	Mean	S.D Deviation	Variance
DialoGPT	3.00	9.00	6.57	1.84	3.38
T5	2.00	10.00	6.37	2.07	4.30
BERT	1.00	10.00	6.13	2.50	6.25

Analysis and Insights

Example Responses:



- The results do not indicate that a particular model was significantly better than the other. A majority of the evaluators, however, selected DialoGPT as the better model.
- The results also show that our models can generate inaccurate information and biased responses.

Improving Our Chosen Model

We integrate our preferred model (DialoGPT) with a heuristic-based model to improve its conversational abilities. We also connect its context to the internet. Our results show an improvement in BLEU Scores.

Table 7.6: BLEU Scores

Model	BLEU Score
DialoGPT (original)	0.352
DialoGPT (improved)	2.228

Conclusion

The results show that LLMs could be of tremendous use in the healthcare industry. However, the results also indicate a lack of readiness to be deployed in real-world settings, much less as a tool for global health provisioning, due to misinformation and bias.

We recommend the following:

- Researchers should source more representative and accurate training data
- The process of training these models be made transparent so that users are aware of their limits
- AI and humans should work together complimenting each other instead of relying on one solely over the other.

Further work is required to improve these models.

Acknowledgment

I would like to thank my advisor, Kowshik Bhowmik, for his guidance, the College of Wooster's Mathematical and Computational Science Department, my family, and my friends for their support during this project.

References:

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (p./pp. 5998--6008), .
2. LasseRegin. Lasseregin/medical-question-answer-data: Medical question and answer dataset gathered from the web url: https://github.com/LasseRegin/medical- question- answer- data