MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

PNW

PURDUE UNIVERSITY NORTHWEST

@twitterhandle

Answering List-Type Questions in Health Domain with a Pretrained Large Language Model: A Case for COVID-19 Symptoms

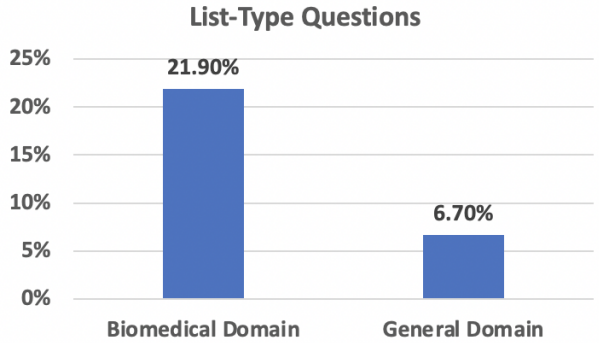
K Jiang[1], M Mujtaba[1], GR Bernard[2]

[1] *Purdue University Northwest*

[2] *Vanderbilt University*

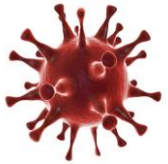
List-Type Questions

- Questions having *a varying number* of answers
- Disease => Symptoms
- Treatment => Options
- Medication => Side Effects



- More list-type questions in biomedical domain (21.9%) than general domain (6.9%) (Yoon, 2022)

The COVID-19 Pandemic



- Users shared their personal experience of the viral infection on social media
- Gathering social media users' symptomatic experience can help supplement/enrich our understanding of the deadly disease
- Annotation is a time-consuming, laborious process requiring domain experts
- Social media data: Noisy, informal writing, out-of-vocabulary short text, layman's terms

COVID-19-Related Tweet

Please share this:

My Experience:

Day 1: Body aches and cold chills. (Tested for Covid19/positive)
Day 2: Minor Body Aches, no chills
Day 3: No aches, but loss of sense of smell and taste
Day 4 thru Day 11: No sense of smell or taste
Day 12: Regained both smell and taste

12:31 AM · Mar 25, 2020

Prior Efforts

- For formal writing:
  - Wasim et al. utilized dependent tree to parse biomedical passages (2018)
  - Yoon et al. developed a BioBERT-based sequence tagging method (2022)
- For Twitter posts:
  - Guo et al. annotated 30732 tweets using n-grams (2020)
  - Krittanawong et al. labelled 14698 tweets (2020)
  - Saker et al. utilized semiautomatic filtering on 7495 tweets (2020)
  - Jiang et al. manually annotated 699 tweets (2022)

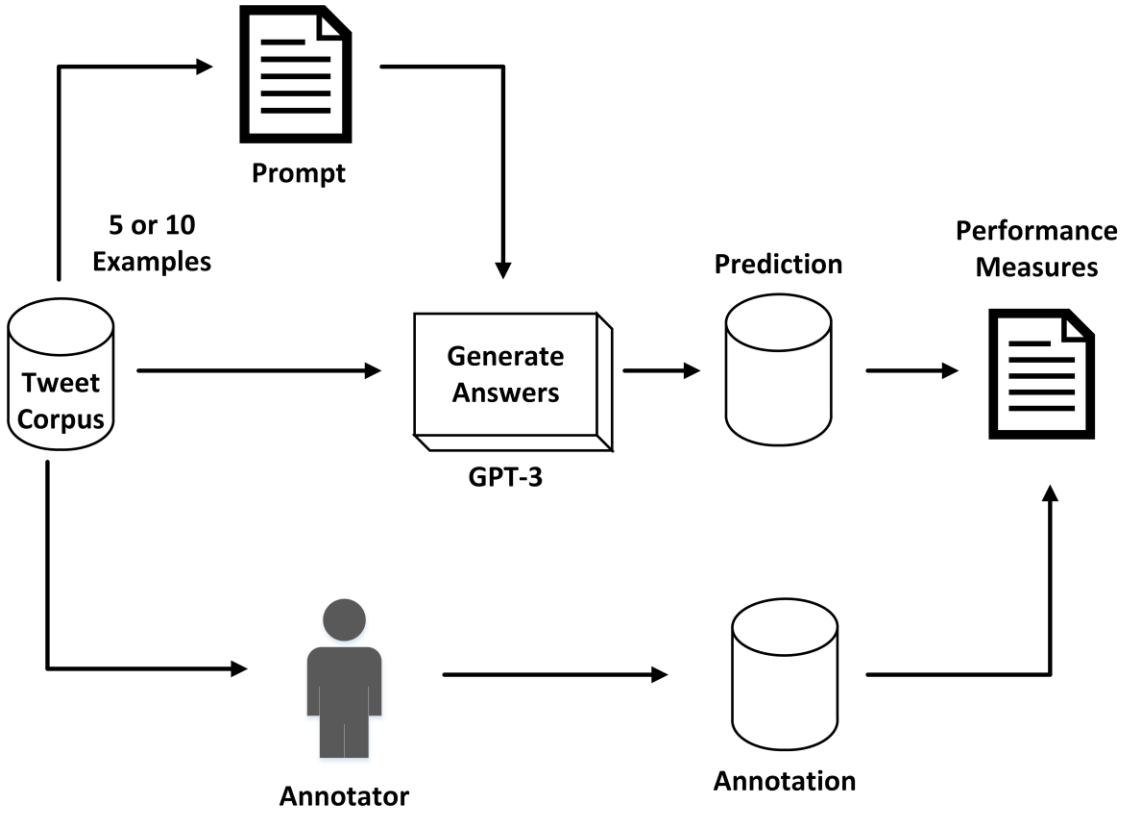
Method

GPT-3 (Brown et al. 2020)

- The 3rd generation of Generative Pre-trained Transformer
- Based on GPT-2
- 8 sizes
- The largest model: 175B parameters
- Trained with ~500 billion tokens
- STOA performance on many NLP tasks

Few-Shot

- Small # of instances
- 655 tweets for testing
- Out-of-box GPT-3



Prompt

- Instruction to LLM
- An important step
- Natural language
- Trial and error

List all the symptoms from the tweet.

Tweet: In hindsight I had Wuhan virus this past January. I had all the symptoms that are now known. I was a little sick for 5 or 6 days: achy, sore throat, dry cough and low grade fever. I went to work every day because I didn't know what I had. I was perfectly fine on day 6.
Symptoms: ['achy', 'sore throat', 'dry cough', 'fever']

… more examples of Tweets with symptoms …

Tweet: Please share this: . . My Experience:. . Day 1: Body aches and cold chills. (Tested for Covid19/positive). Day 2: Minor Body Aches, no chills. Day 3: No aches, but loss of sense of smell and taste. Day 4 thru Day 11: No sense of smell or taste. Day 12: Regained both smell and taste
Symptoms:

Predicted: ['body aches', 'cold chills', 'minor body aches', 'loss of sense of smell', 'loss of sense of taste']
Annotated: ['aches', 'chills', 'aches', 'loss of sense of smell and taste', 'No sense of smell or taste']

Performance Measures

- Exact Match (EM)
  - A ***strict*** measure, character-based match on each question
  - A single mismatch between the prediction and ground truth ➔ Non-EM
  - Not considering *partial match* (sub-string) and *semantic match*
  - We propose Total Match (TM)
  - TM = Exact Match + Partial Match + Semantic Match
- F1 score
  - A ***loose*** measure for answers to all questions

Result

	Exact Match	Total Match	Precision	Recall	F1
0-shot	0.109	0.212	0.779	0.856	0.816
5-shot	0.191	0.377	0.844	**0.890**	0.866
10-shot	**0.264**	**0.429**	**0.912**	0.846	**0.877**

Discussions

- Total match makes more sense than exact match
- Semantic match was done manually
- GPT-3 appears to be a powerful language model
- More examples provided, the better the total match and F1
- Negation
  - *No fever* vs. *No sense of taste and smell*

Conclusion

- Utilization of pretrained Large Language Models like GPT-3 appears to be a powerful approach for our task
- Significantly reducing the laborious annotation effort
- A valuable to tool for new diseases, medications and so forth
- Applicable to many other biomedicine NLP tasks

Thank you!