Large Language Models are Zero-shot Oncology Information Extractors

Madhumita Sushil, Brenda Miao, Divneet Mandair, Travis Zack, Atul J. Butte

Bakar Computational Health Sciences Institute, University of California, San Francisco

madhumita.sushil@ucsf.edu

MOTIVATION

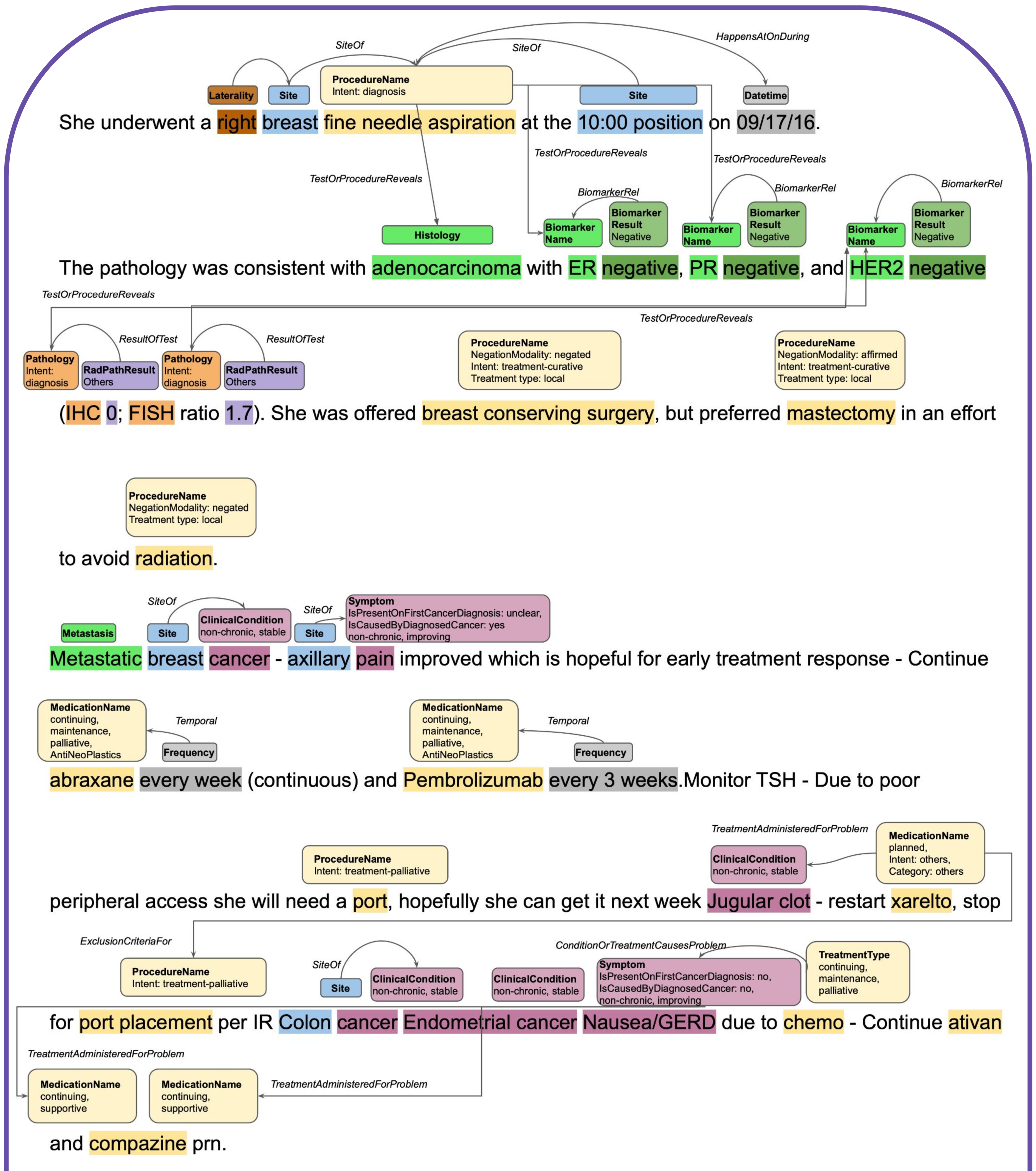
No advanced, public cancer data models for MedOnc text

Existing schemas are either cancer- / note-type-specific, or only represent a subset of information.

No comprehensively annotated benchmarks for oncology information extraction

Lack of benchmark datasets limit further studies.

ANNOTATIONS



She underwent a right breast fine needle aspiration at the 10:00 position on 09/17/16.

The pathology was consistent with adenocarcinoma with ER negative, PR negative, and HER2 negative

(IHC 0; FISH ratio 1.7). She was offered breast conserving surgery, but preferred mastectomy in an effort

to avoid radiation.

Metastatic breast cancer - axillary pain improved which is hopeful for early treatment response - Continue

abraxane every week (continuous) and Pembrolizumab every 3 weeks.Monitor TSH - Due to poor

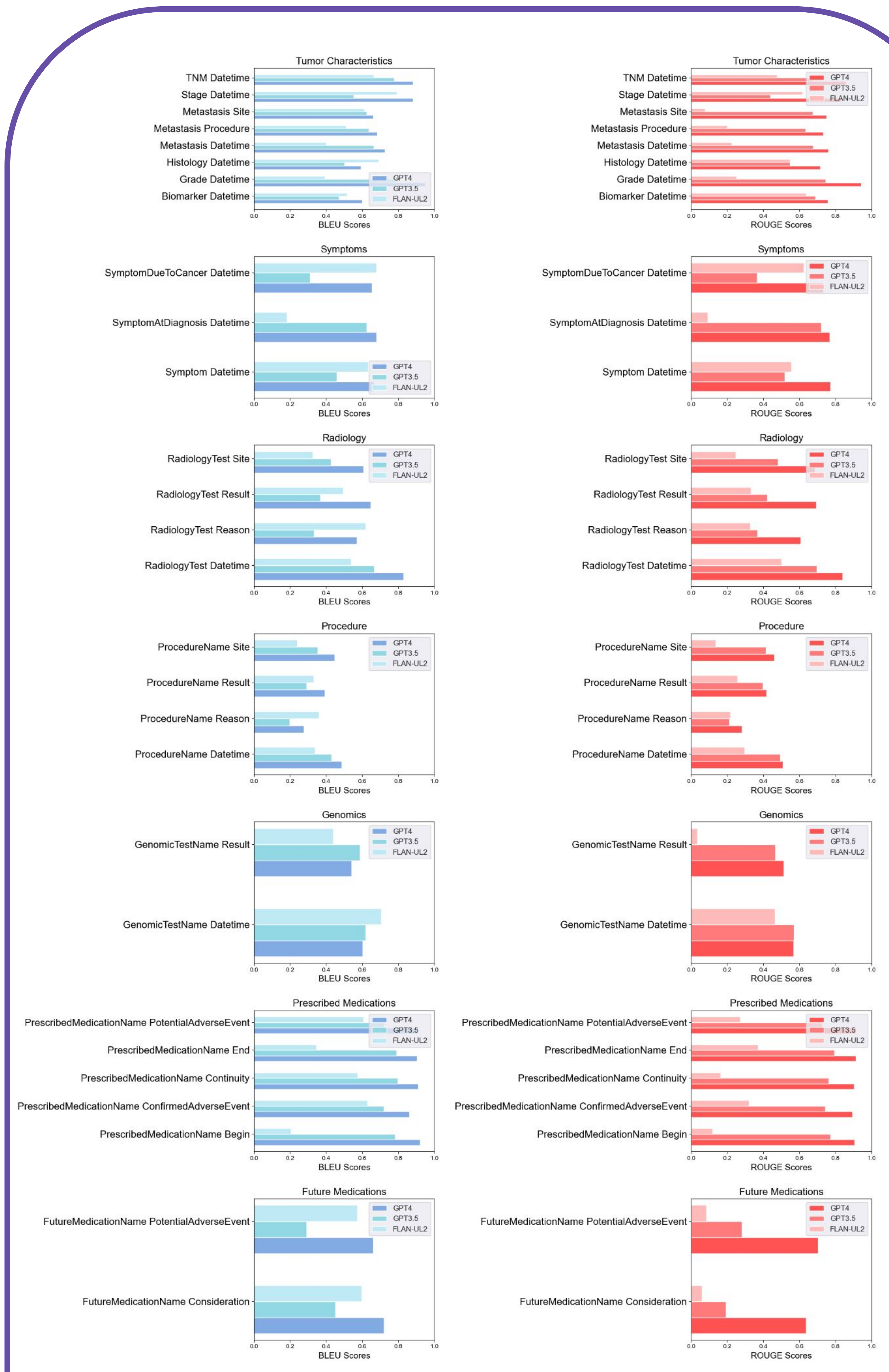
peripheral access she will need a port, hopefully she can get it next week Jugular clot - restart xarelto, stop

for port placement per IR Colon cancer Endometrial cancer Nausea/GERD due to chemo - Continue ativan

and compazine prn.

- 20 breast cancer and 20 pancreatic cancer notes annotated in this manner
- Annotations comprised 9028 entities, 9986 attributes or modifiers, and 5312 relationships between entities.
- Dataset will be made publicly available following manuscript publication.
- Resulting expert annotations used to benchmark LLMs in zero-shot extraction of paired entities and relations:
    - Symptom, tumor characteristics, radiology, procedure, genomics, and medication extraction tested.

INFORMATION EXTRACTION RESULTS



- GPT-4 outperformed GPT-3.5 and FLAN-UL2. Clinical T5, llama/llama2 13B were not usable.
    - Given higher precision, low recall, FLAN-UL2 is promising for further clinical fine-tuning.
- Mean BLEU score (precision) for GPT-4 was 0.68, and mean ROUGE score (recall) was 0.71.
- Despite no previous known clinical training, GPT-4 is excellent at tumor characteristic and prescribed medication extraction.
- Scope of improvement for more advanced reasoning-driven extraction. Potential to improve with in-context learning and advanced prompt engineering.

CONCLUSIONS AND DISCUSSION

- Our schema captured nuanced rhetoric in medical oncology narratives, including family history, SDOH factors, clinical data, causality between diagnoses, treatments, and symptoms, and treatment intent and response including potential and current adverse events.
- The study demonstrated impressive zero-shot capability of the GPT-4 model in synthesizing oncologic history from the HPI and A&P sections, including tasks requiring advanced linguistic reasoning, such as extracting adverse events for prescribed medications.
- GPT-4 is promising for scaling manual phenotyping efforts to facilitate NLP-based research and applications, for real-world drug and device safety monitoring, automatically populating cancer registries, and generating larger annotated datasets for supervised machine learning.
- However, open source models like FLAN-UL2 are also promising for use with further clinical fine-tuning for a privacy-preserving counterpart.

https://arxiv.org/abs/2308.03853