UCSF Bakar Computational Health Sciences Institute

Expert-curated dataset for inference of advanced oncology concepts and relations with large language models

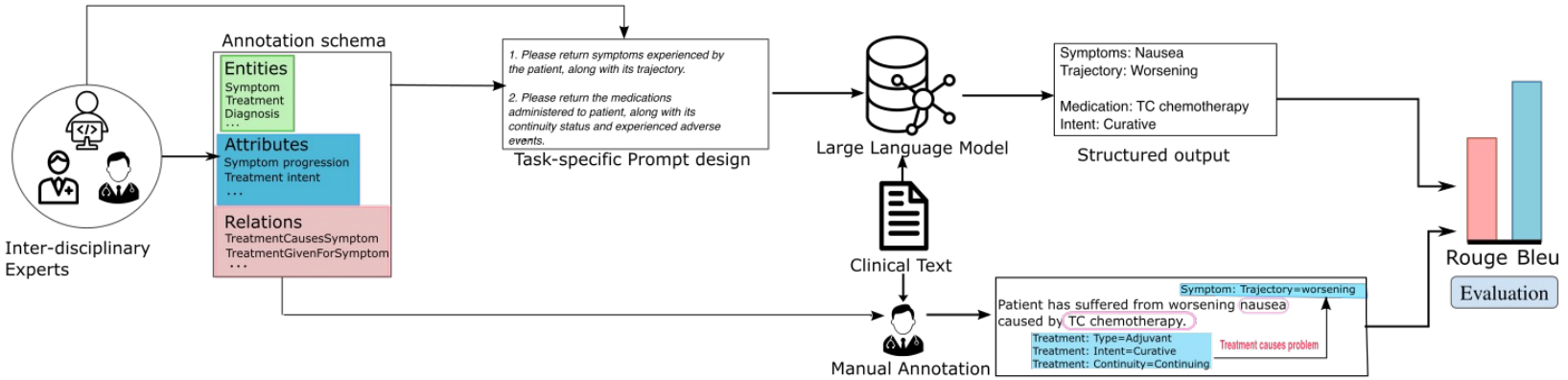
February 2nd, 2024

Madhumita Sushil, PhD
University of California, San Francisco
Madhumita.Sushil@ucsf.edu
@madhumitasushil

Can large language models like GPT-4 help oncology?

- Large language models are trained to generate text seen on the internet and to follow human instructions and preferences, which makes them remarkable in their ability to chat with humans.
- These models are being investigated for their promising capability in medicine, including:
  - Diagnostic assistance
  - Patient phenotyping
  - Clinical report summarization
- Can the GPT-4 model also extract cancer phenotypes to assist observational studies, clinical care, and clinical documentation?

Zero-shot oncology information extraction with GPT-4

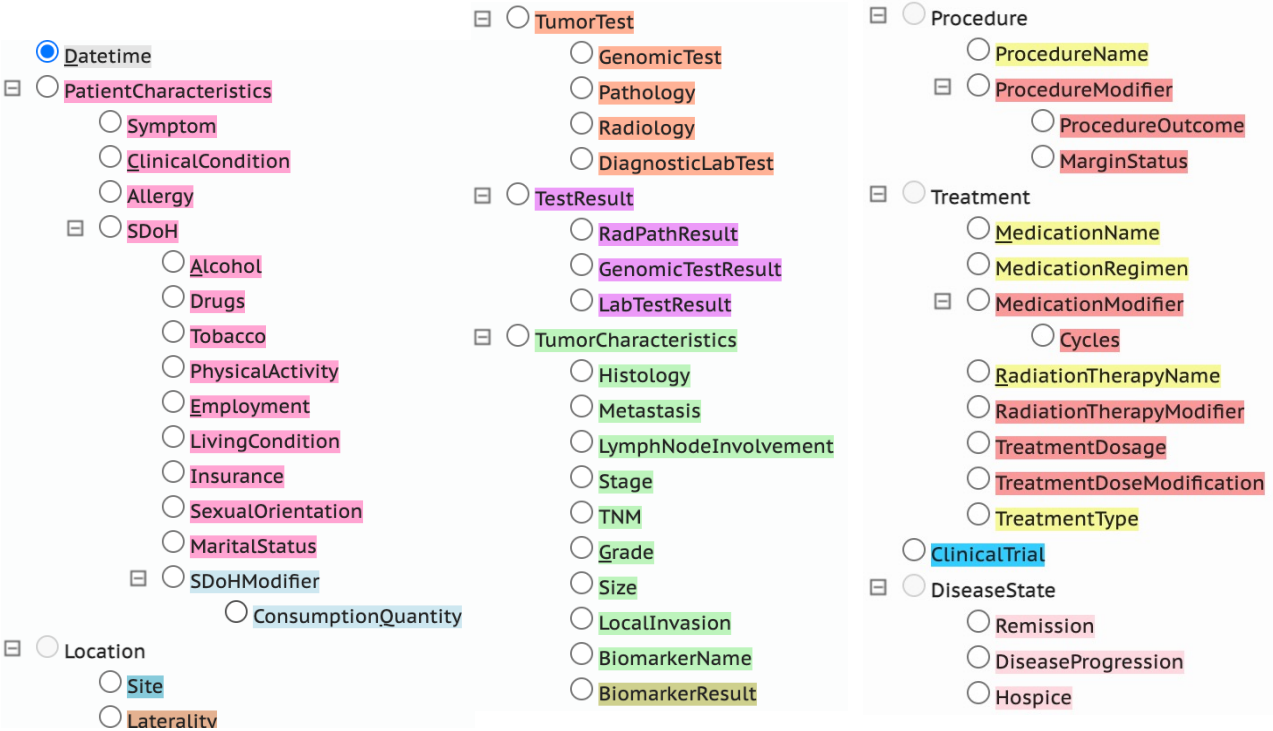


- Oncology fellows at UCSF synthesized detailed clinically relevant information in breast and pancreatic cancer progress notes
- GPT-4 model was asked to provide details of symptoms, tumor characteristics, radiology, procedures, genetic findings, and medications from these progress notes.
- Model performance was benchmarked against clinician performance.

Sushil, Madhumita, et al. "CORAL: Expert-Curated Medical Oncology Reports to Advance Language Model Inference." *New England Journal of Medicine - AI* (2024).

Oncology data representation schema

- Oncology-specific information grouped into logical buckets of information
- Information annotated as entities, attributes, and relations
- BRAT software used to create the schema



http://brat.nlplab.org/manual.html

+ Attributes

- Modifiers of the entities. Examples:
  - Negation
  - Experiencer
  - Is a symptom caused due to cancer diagnosis
  - Temporality (history/new; stable/improving/worsening; meds finished/discontinued early?)
  - Cancer episodes
  - Intent of a test or a procedure
  - Type of treatment (adjuvant/neo-adjuvant/maintenance/local, curative/palliative, anti-neoplastic/others)
  - …

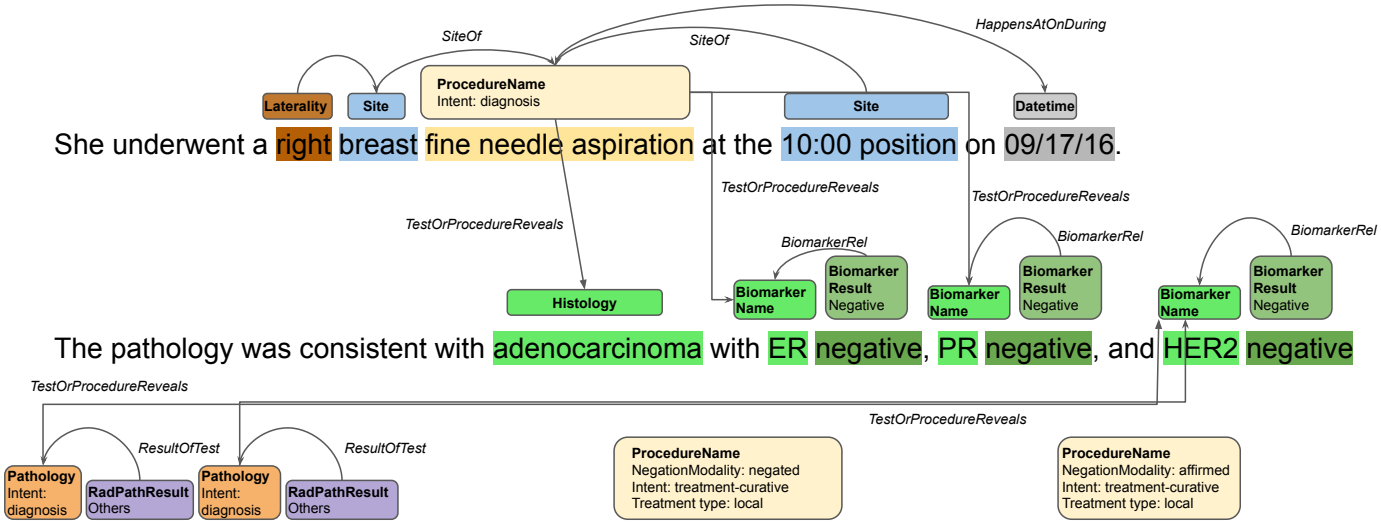
+ Relations

- Temporal relations between a datetime entity and any other entity
- Descriptive relations, for example, relation between a test and its result
- Advanced relations: those requiring implicit or explicit inference
  - TestOrProcedureConductedForProblem
  - TreatmentDiscontinuedBecauseOf
  - TestOrProcedureReveals X
  - ConditionOrTreatmentCausesProblem
  - TreatmentAdministeredForProblem
  - X NotUndergoneBecauseOf Y
  - InclusionCriteria
  - ExclusionCriteria

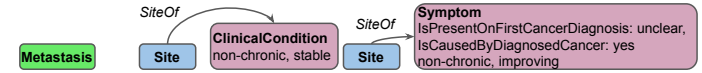
Creating a detailed, expert-labeled dataset

- First progress notes of a diverse set of 20 breast and 20 pancreatic cancer patients from the University of California, San Francisco were selected for annotations
  - Patients with an encounter-specific breast/pancreatic cancer diagnosis and corresponding structured staging data were included.
  - First note is assumed to contain detailed diagnosis and treatment history from the first consultation
- Incorrect redactions, e.g. clinical trial names, genomics, stage manually corrected first
  - Was important for robust benchmarking
- All narrative sections of the reports were annotated
- Inter-annotator agreement was established on a subset of 4 notes. Remaining reports each singly-annotated thereafter. All annotations are reviewed by a third person for missing entities/relations.
- Annotations took a total of 254 person-hours.

Creating a detailed, expert-labeled dataset



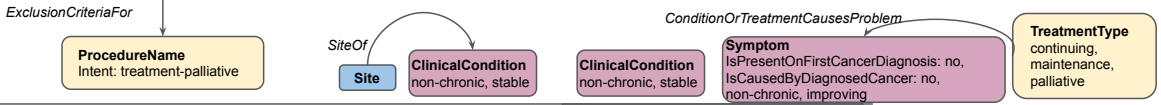
Creating a detailed, expert-labeled dataset



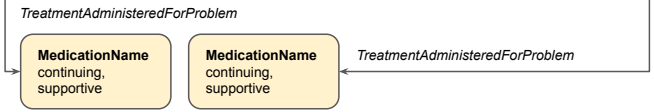
Metastatic breast cancer - axillary pain improved which is hopeful for early treatment response - Continue

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

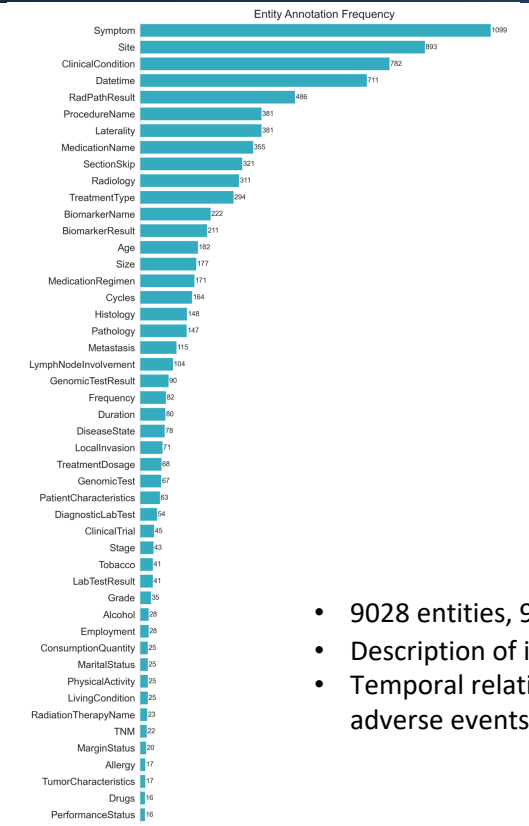


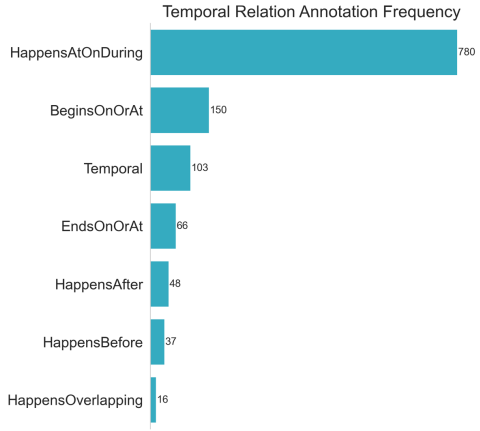
for port placement per IR Colon cancer Endometrial cancer Nausea/GERD due to chemo - Continue ativan

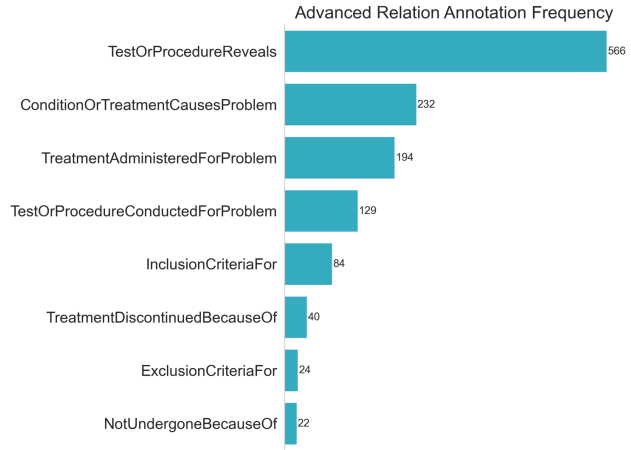


and compazine prn.

Inference-based relation representation across 40 oncology progress reports







- 9028 entities, 9986 entity attributes, and 5312 relationships were annotated
- Description of initial cancer diagnosis, and disease and treatment progression were elaborately documented
- Temporal relations were common, as were indications of findings from a test or procedure, and relations attributing adverse events to pre-existing conditions or treatments

Oncology Relation Extraction Tasks

- Symptom presentation
  - Symptoms that occurred: datetime
  - Symptoms at the time of cancer diagnosis: datetime
  - Symptoms due to diagnosed cancer: datetime
- Tumor characteristics
  - Biomarkers + datetime
  - Histology + datetime
  - Stage (TNM/non-TNM) + datetime
  - Grade + datetime
  - Metastasis + site of metastasis + procedure/test that revealed metastasis + datetime of the procedure/test
- Genetic tests with datetime, results
- Radiology tests and procedures with datetime, medical indication for test, site of test, results
- Medications
  - Cancer-specific prescribed medications, start date, end date, reason for prescription, whether it is continuing, finished, or discontinued early, potential adverse events (listed), adverse events that occurred
  - Cancer-specific medications discussed for future use, whether they are planned or discussed hypothetically, and potential adverse events (mentioned)

Evaluation

BLEU-4 (with smoothing)

- Precision-focused metric for text generation
- Score based on an overlap between phrases (ngrams) between model output and the reference annotations
- Also penalizes very short outputs compared to references
- High score if most of the generated words are present in the annotated text samples

ROUGE-1

- Recall-focused metric for text generation
- Score based on an overlap between phrases (ngrams) between the model output and the reference annotations
- High score if most of the annotated words are present in the generated text

Exact Match F1

Expert human evaluations

Example of radiology scoring:

Model output:

chest MRI: december 2015

Reference annotations:

MRI: 12th december 2015

BLEU score: 0.45

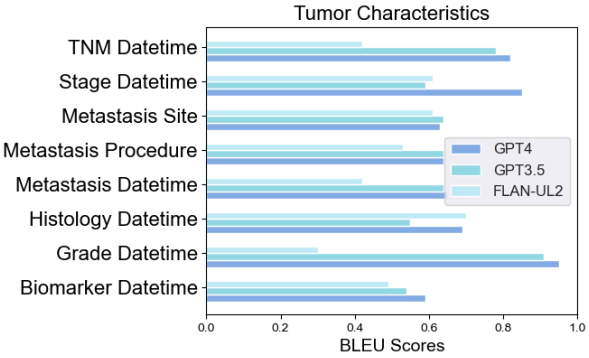
(the model provided the site chest in addition to the annotated info)

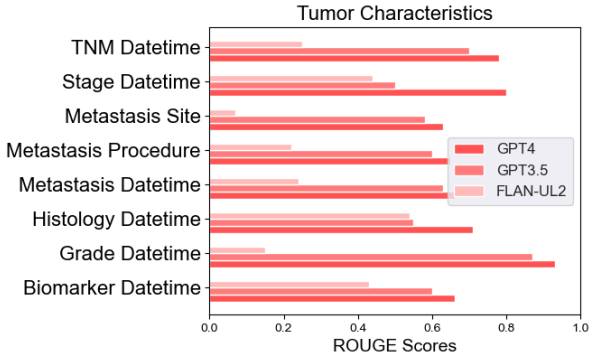
ROUGE score: 0.75

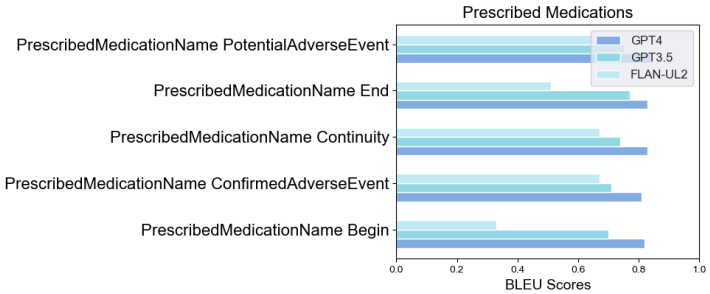
(the date 12th missing from output)

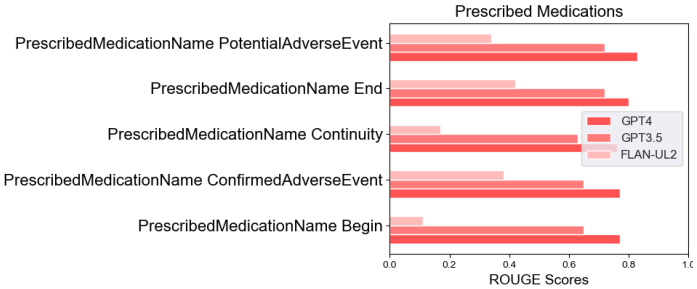
Results: Zero-shot phenotyping with LLMs

- GPT-3.5 and FLAN-UL2 underperform compared to the GPT-4 model, but FLAN-UL2 can be a promising open-source alternative.
- GPT-4 is excellent at extracting cancer stage and grade with their timelines, and details of prescribed medication timeline, associated adverse events.
- It can also extract timelines of tumor histology, metastasis, and biomarkers reasonably well, albeit with high verbosity.



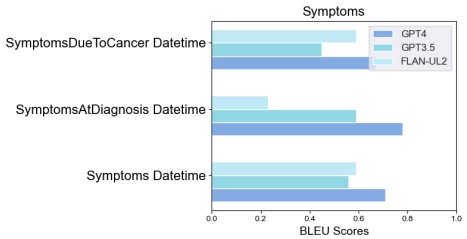


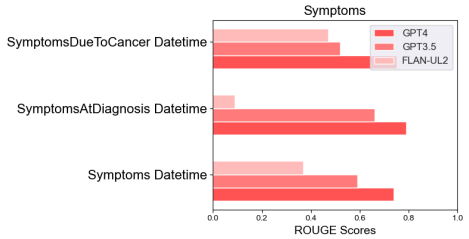


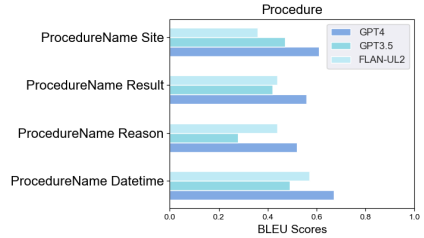


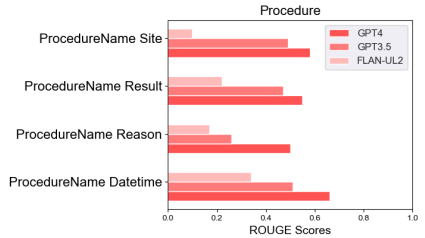
Results: Zero-shot phenotyping with GPT-4

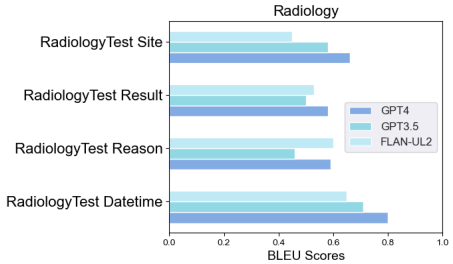
- GPT-4 model is good at extracting symptoms at the time of cancer diagnosis, and genetic test timeline and results.
- However, it can be improved in inference of symptoms that occurred due to cancer, radiology test results, and procedures.

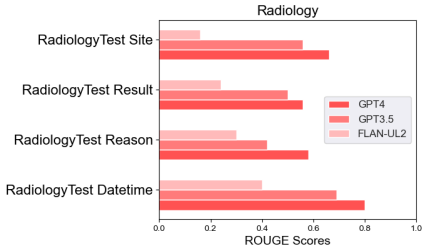


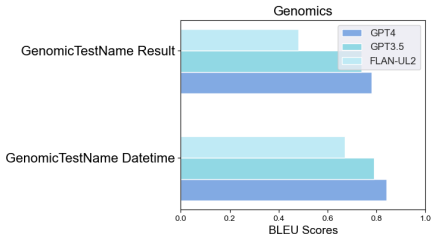


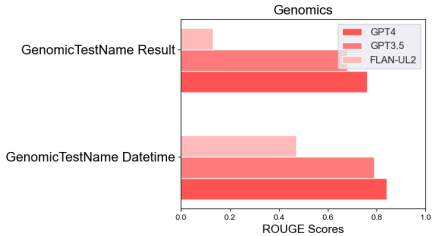












Conclusions

- GPT-4 has an impressive ability to extract oncology-relevant information from clinical notes despite no known task-specific pre-training.
- However, further research is needed to improve inference before the model can be used for clinical research, complex population management, and documenting quality patient care.
- The benchmarking dataset with detailed annotation guidelines and the annotation schema are being shared to build upon zero-shot performance and advance open-source inference.

UCSF Bakar Computational Health Sciences Institute

Acknowledgements

Atul Butte, MD, PhD
Travis Zack, MD, PhD

Vanessa Kennedy, MD
Divneet Mandair, MD
Binh Cao
Brenda Miao

UCSF AI Tiger Team
Academic Research Services
Research Information Technology
The Chancellor's Task Force for Generative AI
Wynton HPC team

Reach out at:

Madhumita.Sushil@ucsf.edu

@madhumitasushil

Sample prompt

System role:

Pretend you are an oncologist. Answer based on the given clinical note for a patient.

User prompt template:

{task-specific-prompt}
Answer as concisely as possible.
Use '\" for special quotation characters.
Do not return any information not present in the note.
Do not return any explanation.

For this note, please return all cancer-directed medications that were prescribed to the patient.
Pair these medications with the date they were prescribed, and the date they were stopped as accurately as possible.
If the medication name has been identified, add the details of the symptom or clinical finding that it was prescribed for.
Also add details about the medication's continuity status among the following options: 'continuing', 'finished', 'discontinued early', or 'unknown'.

Additionally include any problems that were caused due to the medication, and any potential problems that the medication can cause, only if it is mentioned in text.

If any information is not present within the note, please return 'unknown'.

Please return as namedtuples separated by newlines in the following format:

PrescribedMedEnt(MedicationName='Medication identified', Begin={'Medication start date or time'}, End={'Medication end date or time'}, Reason={'symptom or clincal finding that the known medication was prescribed for'}, Continuity='continuity status of the medication',

ConfirmedAdvEvent={'problems that were certainly caused due to the medication'}, PotentialAdvEvent={'problems that could potentially be caused due to the medication, but did not certainly happen.'})

Example:
PrescribedMedEnt(MedicationName='Anastrozole', Begin={'01/01/2019'}, End={'01/01/2020'}, Reason={'unknown'}, Continuity='finished', ConfirmedAdvEvent={'swelling'}, PotentialAdvEvent={'unknown'})

PrescribedMedEnt(MedicationName='Abraxane', Begin={'01/03/2022'}, End={'unknown'}, Reason={'cancer'}, Continuity='started', ConfirmedAdvEvent={'unknown'}, PotentialAdvEvent={'swelling'})

Please do not provide an answer if the MedicationName itself is not available.
DO NOT return medications that are planned or may be given in future.
Do not skip any fields in the given format.

Queries through Versa GPT-4 API on RAE