Improving large language models for clinical named entity recognition via prompt engineering

Yan Hu, MS[1], Qingyu Chen, PhD[2,3], Jingcheng Du, PhD[1], Xueqing Peng, PhD[2], Vipina Kuttichi Keloth, PhD[2], Xu Zuo, MS[1], Yujia Zhou, MS[1], Zehan Li, MS[1], Xiaoqian Jiang, PhD[1], Zhiyong Lu, PhD[3], Kirk Roberts, PhD[1], Hua Xu, PhD*[,2]

[1]McWilliams School of Biomedical Informatics, Houston, TX, United States, [2]Section of Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, CT, United States, [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States

*Corresponding author: Hua Xu, PhD, Section of Biomedical Informatics and Data Science, School of Medicine, Yale University, 100 College St, New Haven, CT 06510, USA (hua.xu@yale.edu)

Romen Samuel Rodis Wabina
Student, Data Science for Healthcare and Clinical Informatics

Introduction

- **Electronic health records (EHRs) contain a vast quantity of unstructured data (e.g., clinical notes), which can offer valuable insights into patient care and clinical research.**
- **Manually extracting patient's information from clinical notes presents a challenge.**
  - **Labor-intensive and time-consuming**
- **Researchers have developed various natural language processing (NLP) techniques for automating clinical information extraction.**
  - **Clinical Named Entity Recognition (NER)**

Clinical Named Entity Recognition

- **Clinical named entity recognition (NER) focuses on recognizing clinical words or phrases and determining their semantic categories, such as medical condition, treatment, and tests.**

Recent laboratory tests of a **63-year-old male** with **CKD stage 3** showed a **serum creatinine of 2.1mg/dL** and an **eGFR of 35mL/min/1.73 m²**.
He has history of **hypertension** and **T2DM**.

Clinical Named Entity Recognition

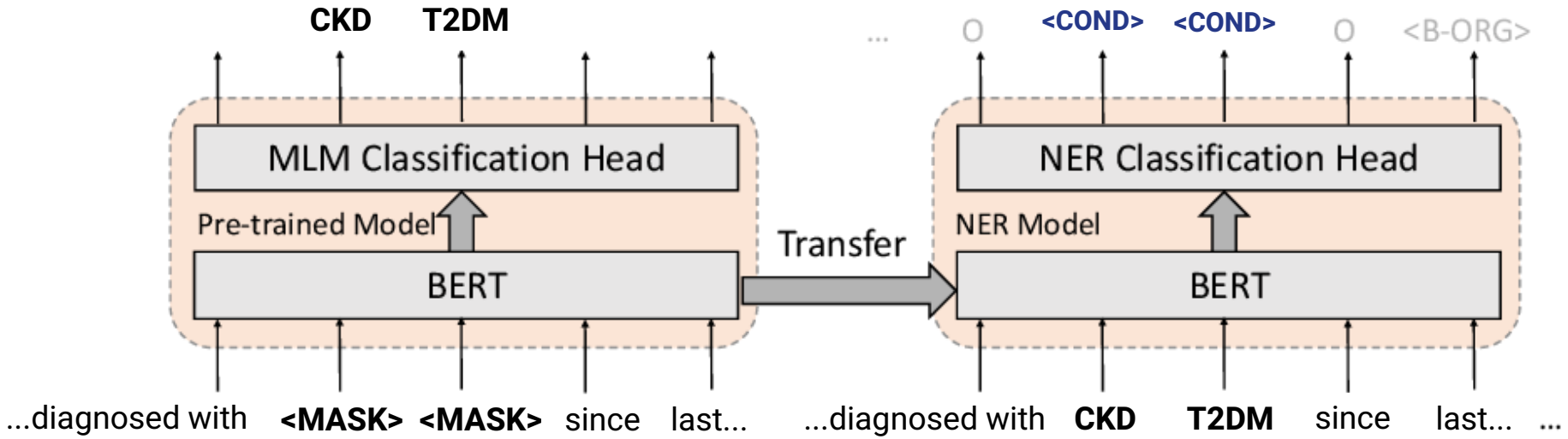
- **Clinical named entity recognition (NER) focuses on recognizing clinical words or phrases and determining their semantic categories, such as age, sex, medical condition, treatment, and tests.**

Recent laboratory tests of a 63-year-old | age male | sex with CKD stage 3 | condition showed a serum creatinine of 2.1mg/dL | test , and an eGFR of 35mL/min/1.73 m² | test He has history of hypertension | condition and T2DM | condition

- **Early clinical NER systems often depend on predefined lexical resources and syntactic/semantic rules derived from extensive manual analysis of text.**
- **Over the past decade, machine learning (ML) approaches have gained popularity in clinical NER research.**

Transformers in Clinical NER

- **Transformer-based models have emerged as the leading method for clinical NLP applications.**
  - **Bidirectional Encoder Representations from Transformers (BERT)**



- **Domain-specific language models:**
  - **BioBERT, PubMedBERT, ClinicalBERT, BioClinicalBERT**
  - **These models have been applied to clinical NER via transfer learning and have shown improved performance with fewer annotated samples.**

Generative Pre-Trained Transformers

- **OpenAI unveiled GPT3.5 (November 2022) & GPT4 (March 2023)**
  - **GPT3.5 has 6B parameters while GPT4 has 175B parameters**
  - **Conversational agent adept at following complex instructions and can generate high-quality responses**
  - **Machine translation and question-answering**
- **Numerous studies are currently exploring the wide range of possibilities offered by GPTs**
  - **GPT3.5 passed the US medical license exam with about 60% accuracy**
  - **GPT4 achieved comparable performance in biomedical question-answering (QA) tasks in comparison to the human expert.**
  - **GPT3.5 and GPT4 exhibited great potential for clinical NER, especially in circumstances where labeled data are not available.**
- **This study aims to investigate the potential of GPT models for clinical NER tasks.**

Prompt Engineering

- **Prompt engineering has emerged as a crucial aspect of utilizing GPT models effectively**
  - **A process of structuring an instruction that can be understood by a generative model (i.e., GPT).**
  - **Guides the model to generate desired outputs.**

- **Studies have explored prompt engineering for GPT models in biomedical and healthcare domain:**
  - **biomedical question-answering**
  - **text classification**

- **No work has been conducted on prompt engineering for GPT models specifically targeting NER tasks in clinical texts.**

Contributions

1. **Proposed a prompt framework for clinical NER by incorporating different prompt strategies and demonstrate its effectiveness on NER task.**

2. **Established a novel benchmark to evaluate the performance of GPT3.5 and GPT4 in clinical NER task.**

Dataset 1: Medical Transcription Samples (MTSamples)

- **MTSamples contain sample transcription reports for many specialties and different work types**
- **163 fully synthetic discharge summaries**
- **No real patient information.**
- **Annotated according to the annotation guidelines from the 2010 i2b2 challenge, which aims at extracting the following entities:**

Medical Problem | Treatment | Test

Dataset 2: Vaccine Adverse Event Reporting System (VAERS)

- **91 safety reports about extracting nervous system disorder-related events.**
- **Anonymized and do not contain personally identifiable information.**

Investigation | Procedure

Nervous adverse event | Other adverse event

Model Settings

- **GPTs were accessed through Application Programming Interface (API)**
  - OpenAI's GPT3.5 (GPT-3.5-turbo-0301) and GPT4 (GPT-4-0314)
    - GPT3.5 per 1,000 tokens = $0.03 for input and $0.06 for output
    - GPT4 per 1,000 tokens = $0.001 for input and $0.002 for output
  - Temperature for GPT models was set to zero
    - controls the randomness in the model's predictions
    - ranging from 0 (completely deterministic) to 1 or higher (increasingly random and diverse outputs)

- **ChatGPT vs task-specific NER model**
  - Compared GPT models to traditional supervised learning approaches
  - GPT models vs BioClinicalBERT vs Conditional Random Field (CRF)
  - Pre-trained BioClinicalBERT using transformers package
  - Hyperparameters:
    - learning rate of 5e–5
    - batch size of 4
    - 20 epochs
    - weight decay of 0.01 using AdamW optimizer.

Prompt Engineering

The study proposed four task-specific prompts for clinical NER tasks:

Zero-Shot Prompting

- Prompt 1 (Baseline Prompt)
- Prompt 2 (Entity definitions and Annotation Guidelines)
- Prompt 3 (Error Analysis)

- Prompt 4 (Few-Shot Prompts)

The study compared the effectiveness of different prompt components by incrementally incorporating annotation guideline-based prompts, error analysis-based instructions, and annotated samples.

Prompt Engineering

- **Characteristics of Prompt 1 (Baseline Prompt)**
  - **Baseline prompt with task description and format specification.**
  - **Provides the LLMs with basic information about the tasks to perform and in what format.**
  - **Highlight the named entities within an HTML file using tags with a class attribute.**

  - **MTSamples dataset has three entities: problem, treatment, and test while VAERS has four entities, including investigation, nervous_AE, other_AE, and procedure.**

Prompt 1: Baseline Prompt (MTSamples)

Your task is to generate an HTML version of an input text, marking up specific entities related to healthcare.

The entities to be identified are: 'medical problems', 'treatments', and 'tests'.
Use HTML tags to highlight these entities. Each should have a class attribute indicating the type of the entity.

Entity Markup Guide

Use to denote a medical problem.
Use to denote a treatment.
Use to denote a test.

Leave the text as it is if no such entities are found.

Prompt 1: Baseline Prompt (MTSamples)

Your task is to generate an HTML version of an input text, marking up specific entities related to healthcare.

The entities to be identified are: 'medical problems', 'treatments', and 'tests'.
Use HTML tags to highlight these entities. Each should have a class attribute indicating the type of the entity.

Entity Markup Guide

Use to denote a medical problem.
Use to denote a treatment.
Use to denote a test.

Leave the text as it is if no such entities are found.

Prompt

+

She has had no polyuria, polydipsia or other problems.

PREDICTION

She has had no
 polyuria ,
 polydipsia , or other problems.

Prompt Engineering

- **Characteristics of Prompt 2 (Baseline Prompt)**
  - **Annotation guideline-based prompts**
  - **This prompt contains entity definitions and annotation guidelines.**

Entity Definitions

Medical Problems are defined as: phrases that contain observations made by patients or clinicians about the patient's body or mind...

Treatments are defined as: phrases that describe procedures, interventions, and substances given to a patient ...

Tests are defined as: phrases that ....

Annotation Guidelines

Only complete noun phrases (NPs) and adjective phrases (APs) should be marked.

Include articles and possessives. Conjunctions and other syntax that denote lists should be included if they occur within the modifiers or are connected by a common set of modifiers.

- **Entity definitions offer comprehensive, unambiguous descriptions of an entity within the context of a given task.**
- **Specification of annotation also offers accurate identification of entities:**
  - **What types of phrases to be included (e.g., noun or adjective phrases)**

Prompt 2: Baseline Prompt + Entity Definitions and Annotation Guidelines

Prompt 1: Baseline Prompt +

Medical Problems are defined as: phrases that contain observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease. They are loosely based on the UMLS semantic types of pathologic functions, disease or syndrome, mental or behavioral dysfunction, cellormolecular dysfunction, congenital abnormality, acquired abnormality, injury or poisoning, anatomic abnormality, neoplastic process, virus/bacterium, sign or symptom, but are not limited by UMLS coverage.

Treatments are defined as: phrases that describe procedures, interventions, and substances given to a patient to resolve a medical problem. They are loosely based on the UMLS semantic types therapeutic or preventive procedure, medical device, steroid, pharmacologic substance, biomedical or dental material, antibiotic, clinical drug, and drug delivery device. Other concepts that are treatments but that may not be found in UMLS are also included. Treatments that a patient had, will have, may have in the future, or are explicitly mentioned that the patient will not have been all marked as treatments.

Tests are defined as: phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample to discover, rule out, or find more information about a medical problem. They are loosely based on the UMLS semantic types laboratory procedure, diagnostic procedure, but also include instances not covered by UMLS.

Only complete noun phrases (NPs) and adjective phrases (APs) should be marked. Terms that fit concept semantic rules, but that are only used as modifiers in a noun phrase should not be marked. Include all modifiers with concepts when they appear in the same phrase except for assertion modifiers. You can include up to one prepositional phrase (PP) following a markable concept if the PP does not contain a markable concept and either indicates an organ/body part or can be rearranged to eliminate the PP (we later call this the PP test).

Include articles and possessives. Conjunctions and other syntax that denote lists should be included if they occur within the modifiers or are connected by a common set of modifiers. If the portions of the list are otherwise independent, they should not be included. Similarly, when concepts are mentioned in more than one way in the same noun phrase (such as the definition of an acronym or where a generic and a brand name of a drug are used together), the concepts should be marked together. Concepts should be mentioned in relation to the patient or someone else in the note. Section headers that provide formatting, but that are not specific to a person are not marked.

Prompt Engineering

- **Characteristics of Prompt 3**
  - **Error analysis-based instructions**
  - **Incorporated additional guidelines following error analysis of GPT models**

  - **The researchers noticed that GPT models often tend to annotate consultation procedures as test entities. To prevent this, we incorporated a specific rule stating,**
    "Consultation procedures should not be annotated as tests."

Prompt 3: Baseline Prompt + Entity Definitions and Annotation Guidelines + Error Analysis

Prompt 1: Baseline Prompt + Prompt 2: Entity Definitions and Annotation Guidelines +

Vital signs or vital signs with abnormal readings should be annotated as tests.

Medical specialists, services, or healthcare facilities should not be annotated, even if they might seem to fit into the categories of 'tests', 'treatments', or 'medical problems'. These entities are part of the healthcare delivery system and do not directly denote a test, treatment, or medical problem.

Consultation procedures should not be considered as tests.

Prompt Engineering

- **Characteristics of Prompt 4**
  - **Few-shot prompts (i.e., annotated samples)**
  - **To assist GPTs in understanding the task that can generate accurate results**
  - **Randomly selected one- or five-annotated samples and formatted the samples according to task description and entity mark-up guide.**

Prompt 4: Baseline Prompt + Entity Definitions and Annotation Guidelines + Error Analysis + Few-Shot

Prompt 1: Baseline Prompt + Prompt 2: Entity Definitions and Annotation Guidelines + Prompt 3: Error Analysis +

Example Input 1: At the time of admission , he denied fever , diaphoresis , nausea , chest pain or other systemic symptoms .

Example Output 1: At the time of admission , he denied
fever ,
diaphoresis ,
nausea ,
chest pain or other systemic symptoms .

Evaluation Criteria

- **The performance of the models was evaluated using Precision (P), Recall (R), and F1 scores.**
- **Evaluation scores were computed based on both exact-match and relaxed-match criteria.**

**Ground truth: type 2 diabetes **

- **In exact match, an extracted entity should have identical boundary and entity type as that in the gold standard.**

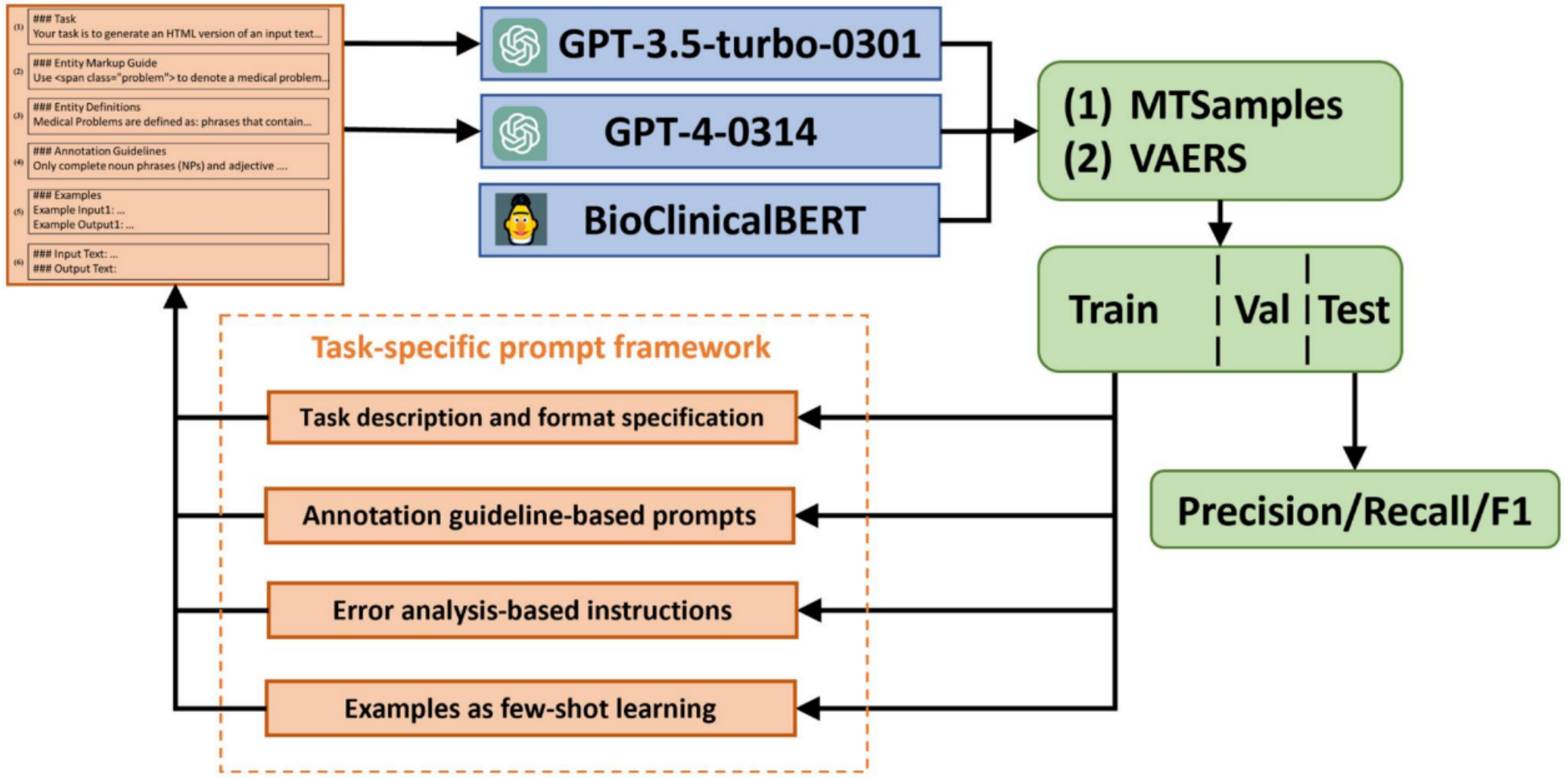
  **Prediction: type 2 diabetes **

- **For relaxed match, an extracted entity that exhibits overlap in text and shares the same entity type with the gold standard is acceptable.**

  **Prediction: diabetes **

  - **Relaxed match, because the extracted entity "diabetes" overlaps with "type 2 diabetes" in the text and shares the same entity type, despite not matching exactly in boundary.**

Overview of the Study Workflow



Zero-Shot Performance of GPT3.5 and GPT4 (using prompts 1-3)

Models	Prompt strategies	MTSamples						VAERS					
		Exact match			Relaxed match			Exact match			Relaxed match		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-3.5	Baseline prompts only (1)	0.492	0.327	0.393	0.794	0.528	0.634	0.510	0.146	0.227	0.626	0.187	0.288
	+ Annotation guideline-based prompts (1)+(2)	0.453	0.405	0.428	0.736	0.680	0.707	0.575	0.200	0.297	0.687	0.243	0.359
	+ Error analysis-based instructions (1)+(2)+(3)	0.462	0.412	0.436	0.755	0.687	0.719	0.569	0.233	0.331	0.730	0.305	0.431
GPT-4	Baseline prompts only (1)	0.486	0.546	0.514	0.762	0.852	0.804	0.420	0.397	0.408	0.599	0.568	0.583
	+ Annotation guideline-based prompts (1)+(2)	0.478	0.577	0.523	0.752	0.919	0.827	0.559	0.444	0.495	0.743	0.593	0.660
	+ Error analysis-based instructions (1)+(2)+(3)	0.488	0.570	0.526	0.777	0.908	0.838	0.536	0.469	0.500	0.727	0.650	0.686

Model-specific effects of prompts

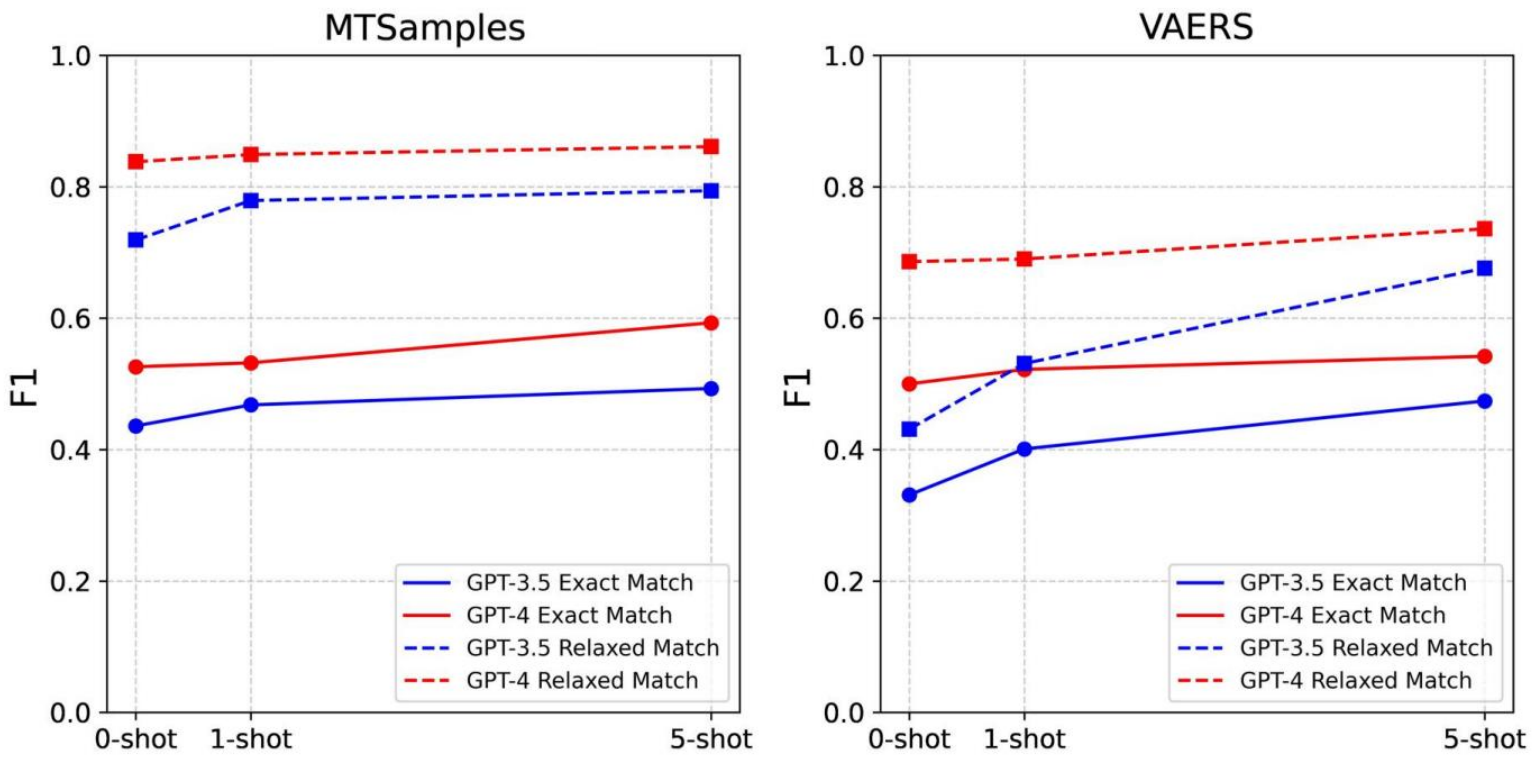
- **The integration of annotation guideline-based and error analysis with Prompt 1 offers an improvement in the performance metrics of both GPT models across each dataset.**

- **Prompts 2 and 3 have a more pronounced effect on GPT3.5 than on GPT4**
  - **GPT3.5 demonstrated an average increase of 0.09 in overall F1 scores**
  - **GPT4 showed a restrained average improvement of 0.06**

Data-specific effects of prompts

- **Both prompts 2 and 3 had more impact on the VAERS dataset compared to the MTSamples.**
- **Average increase of 0.11 for VAERS in overall F1 scores.**

Few-Shot Performance of GPT3.5 and GPT4

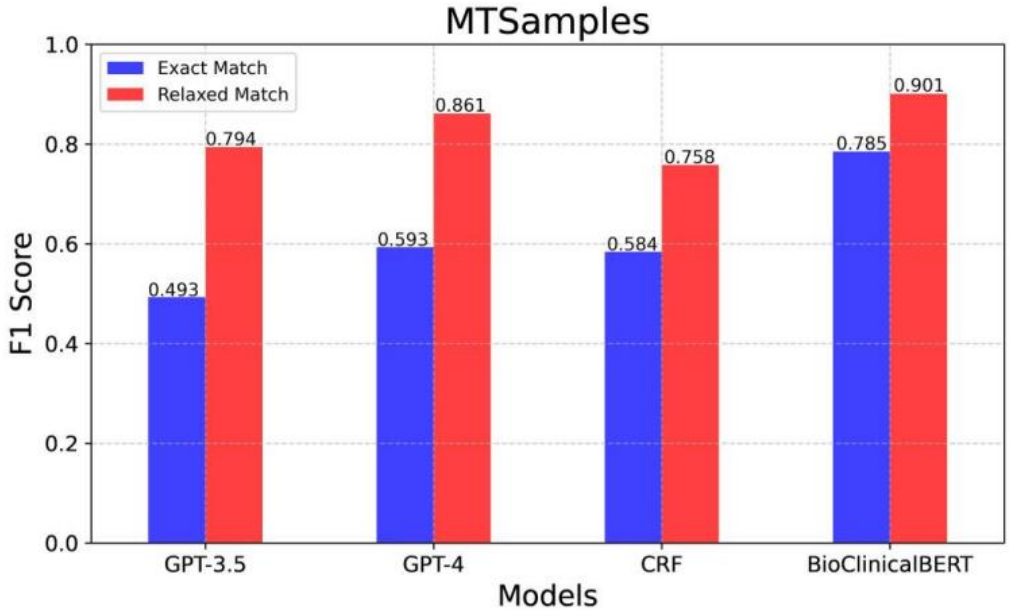


Effect of Few-Shot on Model Performance

- **The inclusion of more examples (shots) leads to better model performance.**
- **A combination of 5-shot and all prompts produced the best results by GPT4.**

Performance Comparison to BioClinicalBERT

	MTSamples					
	Exact match			Relaxed match		
Model	P	R	F1	P	R	F1
GPT-3.5	0.515	0.472	0.493	0.827	0.764	0.794
GPT-4	0.555	0.637	0.593	0.804	0.926	0.861
CRF	0.511	0.681	0.584	0.662	0.887	0.758
BioClinicalBERT	0.785	0.785	0.785	0.915	0.887	0.901

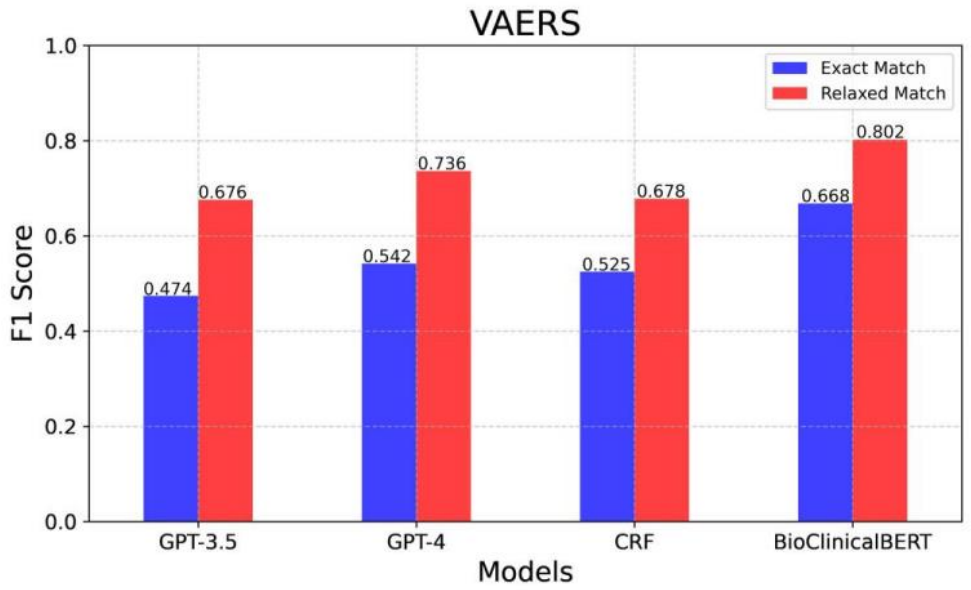


Using MTSamples dataset

- **Among the four models, BioClinicalBERT demonstrated the highest performance in MTSamples dataset.**
- **BioClinicalBERT achieved overall F1 score of 0.785 and 0.901 under exact match and relaxed match, respectively.**
- **GPT3.5 had the lowest performance, yet still demonstrated a decent performance with scores of 0.794 in relaxed criteria.**
- **GPT4 showcased highly competitive performance using the relaxed match criteria, accomplishing F1 score of 0.861.**

Performance Comparison to BioClinicalBERT

	VAERS					
	Exact match			Relaxed match		
Model	P	R	F1	P	R	F1
GPT-3.5	0.526	0.432	0.474	0.735	0.626	0.676
GPT-4	0.513	0.574	0.542	0.701	0.774	0.736
CRF	0.473	0.591	0.525	0.609	0.764	0.678
BioClinicalBERT	0.698	0.640	0.668	0.846	0.761	0.802



Using VAERS dataset

- **BioClinicalBERT also demonstrated the highest performance in VAERS dataset.**
- **BioClinicalBERT achieved overall F1 score of 0.668 and 0.802 under exact match and relaxed match, respectively.**
- **In relaxed match criteria, CRF had comparable performance to GPT3.5 in VAERS dataset.**
- **GPT4 showcased highly competitive performance using the relaxed match criteria with F1 score of 0.736.**

Error Analysis

- **The study classified errors into four types**
  - **Incorrect extraction**
    - **LLM erroneously identifies text as an entity when it should not be recognized as such.**
  - **Missing entity**
    - **LLM fails to identify an entity that should have been recognized.**
  - **Incorrect entity**
    - **LLM correctly identifies an entity but categorizes it under the wrong entity type.**
  - **Boundary error**
    - **LLM does not correctly mark the start and end of an entity.**

Ground Truth

She has had no
 polyuria ,
 polydipsia , or other problems.

Incorrect Extraction

She has had no polyuria, polydipsia, or other problems.

Missing Entity

She has had no polyuria, polydipsia, or other problems.

Incorrect Entity

She has had no
polyuria,
polydipsia, or other problems.

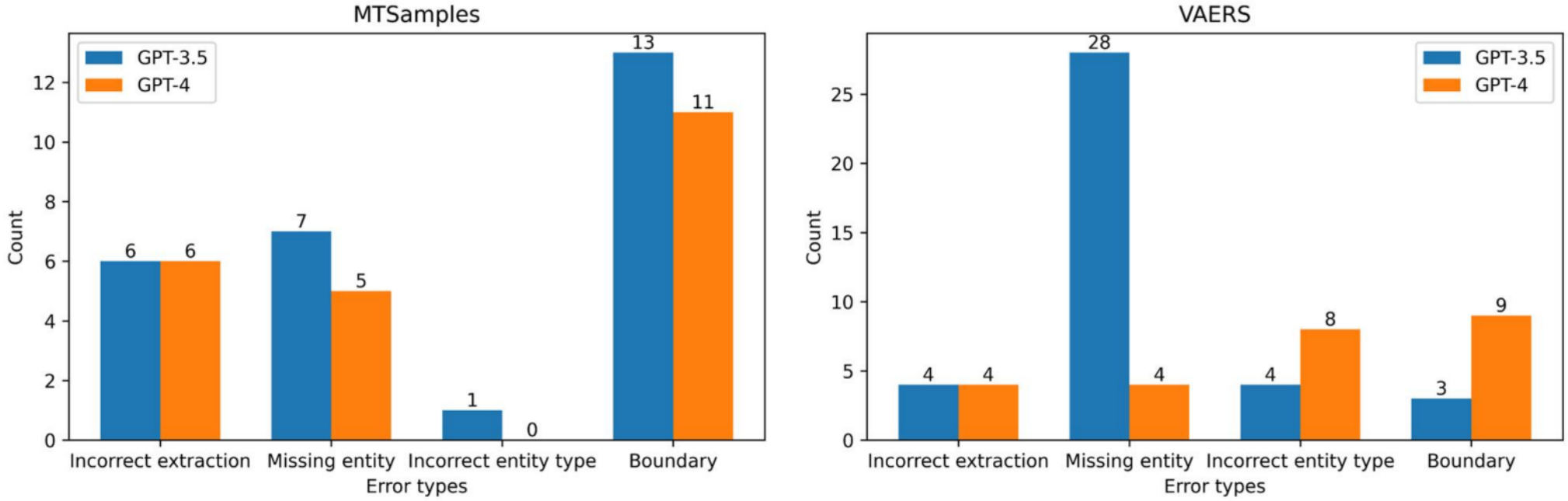
Boundary Error

She has had no
polyuria, polydipsia,
or other problems.

Error Analysis

- **A random sample of 20 sentences was selected from the outputs generated by each GPT model across 2 datasets.**
  - **Selection included sentences with both false positives and false negatives.**
  - **The error analysis was conducted based on exact match.**

- **GPT3.5 and GPT4 exhibited similar error patterns for the MTSamples dataset.**
  - **Both models encountered challenges when it came to identifying correct entity boundaries.**
    - **article words (e.g., 'the' in 'the study drug')**
    - **modifiers (e.g., 'another' in 'another large stroke')**



Discussion

- **LLMs are quick and easy path to build more generalizable NER systems.**
  - Without any model training or fine-tuning, GPTs exhibited exceptional performance in NER
- **This study shows the unrealized potential of LLMs in clinical NER by proposing task-specific prompt framework that incorporates the following:**
  - annotation guidelines
  - error-analysis-based instructions
  - few-shot examples
- **Performance of GPT models improved with task-specific prompts**
  - GPT4 shows a competitive performance as that of BioClinicalBERT
  - GPTs with few-shot prompts can achieve performance that is close to the finetuned models
- **Incorporating annotation guidelines in prompts can improve performance**
  - Medical knowledge are still critical in LLM-based NER system
  - Prompts with annotated examples is effective for improving performance
- **Considering that no training data was used in GPT models, their performance is already impressive, which hints the potential of LLMs in clinical NER tasks.**
  - While the results demonstrate a promising direction, they also underscore the need for further refinement and development before LLMs can consistently outperform established models like BioClinicalBERT

Future Work

- **Future studies can compare OpenAI's GPT3.5 and GPT4 to other LLMs in clinical NER task**
  - Google Gemini
  - Anthropic's Claude
  - Falcon
- **Selection of informative and representative samples has not been investigated in this study.**
  - What type of sample is best suited for few-shot prompts?
- **Other prompt strategies should also be considered.**
  - Chain-of-thought prompting
  - Self-consistency prompts
- **Better evaluation schema to assess LLM performance more accurately.**
  - e.g., GPT models recognized lab tests with abnormal values as medical problems.
- **Other advanced learning algorithms could be explored.**
  - Few-shot algorithms
  - Retriever-Augmented Generation (RAG)

Questions?

Appendix

- **MTSamples was annotated according to the annotation guidelines from the 2010 i2b2 challenge**
  - **This guidelines describes the specific types of information that should be annotated for the clinical NER task**

General Guidelines for Concepts

Only complete noun phrases (NPs) and adjective phrases (APs) should be marked. Terms that fit concept semantic rules, but that are only used as modifiers in a noun phrase should not be marked.

- *The man was obese . The obese man came to the clinic .*
- *She developed diabetes . She takes diabetes medication .*
  - *diabetes medication* is one concept and the full NP should be marked.
- *Patient underwent catheterization . Catheterization report showed …*
- *Patient arrived in the surgery suite . Surgery was performed .*

Include all modifiers with concepts when they appear in the same phrase except for assertion modifiers. (Please see assertion annotation guidelines for a description).

- *bilateral DVT*
- *some recurrent angina*
- *high grade LAD lesion*
- *chronic hepatitis*
- *cataract surgery*
- *diabetes medication*
- *head CT*
- *chest X-ray*
- *no fever*
- *possible tamponade*

Categories of Concepts

Concepts are defined in three general categories that are each annotated separately.

1) *Medical Problems:* phrases that contain observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease. They are loosely based on the UMLS semantic types of pathologic functions, disease or syndrome, mental or behavioral dysfunction, cell or molecular dysfunction, congenital abnormality, acquired abnormality, injury or poisoning, anatomic abnormality, neoplastic process, virus/bacterium, sign or symptom, but are not limited by UMLS coverage.

2) *Treatments:* phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem. They are loosely based on the UMLS semantic types therapeutic or preventive procedure, medical device, steroid, pharmacologic substance, biomedical or dental material, antibiotic, clinical drug, and drug delivery device. Other concepts that are treatments but that may not be found in UMLS are also included. Treatments that a patient had, will have, may have in the future, or are explicitly mentioned that the patient will not have are all marked as treatments.

3) *Tests:* phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem. They are loosely based on the UMLS semantic types laboratory procedure, diagnostic procedure, but also include instances not covered by UMLS.

Appendix: BERT models

- **BioBERT, PubMedBERT, ClinicalBERT, and BioClinicalBERT are all variants of BERT to handle text found in the biomedical and healthcare domains.**
  - **PubMedBERT(PubMed abstracts)**
  - **BioBERT (PubMed abstracts and PMC full-text articles)**
  - **ClinicalBERT (MIMIC-III database)**
  - **BioClinicalBERT (BioBERT pre-trained with MIMIC-III database)**
- **ClinicalBERT or BioClinicalBERT are suitable models compared to BioBERT and PubMedBERT for clinical NER.**
  - **Both are optimized for the nuances and language used in clinical documents rather than purely biomedical literature.**
  - **BioClinicalBERT offer advantages for documents contain a mix of clinical narratives and references to biomedical research.**

Appendix: GPT API Parameters

- **Temperature** adds randomness to the responses
  - Influences the randomness of the generated responses
  - Higher temperature makes the answers more diverse; lower value makes output more deterministic
- **Top-p Sampling** (nucleus sampling) controls the diversity and quality of the responses (default value = 1).
  - Top-p sampling selects tokens whose probability is higher than a given threshold.
  - If the generated text includes irrelevant words, consider decreasing the probability threshold (p).
- **Maximum Tokens** allows users to limit the length of the generated response **(default = 4096 tokens)**.
- **Frequency Penalty** reduces the chance of repeatedly sampling the same sequences of tokens.
- **Presence Penalty** can be used to encourage the model to use a diverse range of tokens in the generated text.
  - It instructs the language model to utilize different words, promoting variety in the outputs
- **Multiple Responses** allows users to generate multiple alternative responses for a given conversation.
- **Logit Bias** allows users to specify a custom condition for the completion.

Appendix: Conditional Random Field

- **Conditional Random Fields (CRFs) are valuable for labeling and segmenting sequential data**
  - **A traditional statistical modeling method often used in pattern recognition**
  - **Assigning labels to each element in a sequence of observation**

x	She	has	had	no	polyuria	polydipsia	or	other	problems
y	O	O	O	O	B-PROB	B-PROB	O	O	O

- **Extract the features f_k for each word based on feature definitions:**
  - k: **word (bag-of-words), capitalization, prefixes and suffixes, position**
  - **e.g., bag-of-words: 'polyuria', capitalization: False, prefix: 'poly', suffix: 'sia', position: <COMMA>**

- **Calculate the scores (numerator) based on feature definitions f_k and its weight λ_k**

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) = \frac{\exp(\text{total score for sequence})}{Z(x)}$$

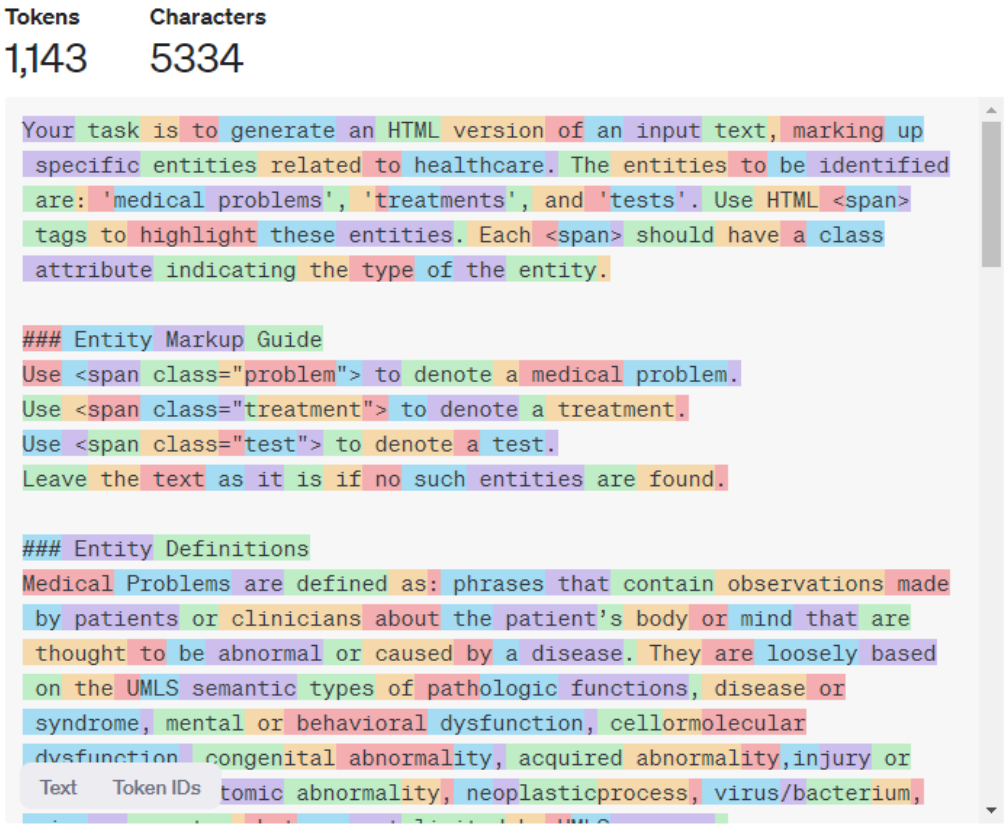
- $Z(x)$ **is the normalization factor**

OpenAI's large language models (sometimes referred to as GPT's) process text using tokens, which are common sequences of characters found in a set of text.

- **The models learn to understand the statistical relationships between these tokens and excel at producing the next token in a sequence of tokens.**

- **Prompt 1 = 124 tokens (549 characters)**
- **Prompt 2 = 668 tokens (3,312 characters)**
- **Prompt 3 = 753 tokens (3,755 characters)**

- **Prompt 4 (one-shot) = 853 tokens (4,410 characters)**
- **Prompt 4 (five-shot) = 1,143 tokens (5,334 characters)**



Prompt Engineering

- **Characteristics of Prompt 1 (Baseline Prompt)**
  - **Baseline prompt with task description and format specification.**
  - **Provides the LLMs with basic information about the tasks to perform and in what format.**
  - **Highlight the named entities within an HTML file using tags with a class attribute.**

  - **VAERS dataset has four entities: investigation, nervous_AE, other_AE, and procedure.**

Prompt 1: Baseline Prompt (VAERS)

Your task is to generate an HTML version of an input text, marking up specific entities related to healthcare. The entities to be identified are: 'investigations', 'nervous adverse events', 'other adverse events', and 'procedures'. Use HTML tags to highlight these entities. Each should have a class attribute indicating the type of the entity.

Entity Markup Guide

Use to denote an investigation.
Use to denote a nervous adverse event.
Use to denote another adverse event.
Use to denote a procedure.

If no entity found, leave the text as it is.

fluzone qiv hd administered to a minor patient with no reported adverse event; patient was supposed to receive the Flumist Nasal Spray but she grabbed the FLUZONE QIV HD and inadvertently gave it to intranasally with no reported adverse event; patient was supposed to receive the Flumist Nasal Spray but she grabbed the FLUZONE QIV HD and inadvertently gave it to intranasally with no reported adverse event; Initial information received from Regulatory Authority on 18-Dec-2023 regarding an unsolicited valid non-serious case received from a nurse. This case involves a 14 years old male patient to whom influenza quadrival A-B high dose HV vaccine [Fluzone High-Dose Quadrivalent] was administered who was supposed to receive the Influenza Vaccine Live Reassort 3v (Flumist) nasal Spary but she grabbed the Fluzone QIV HD and inadvertently gave it to intranasally with no reported adverse event. The patient's past medical history, medical treatment(s), vaccination(s) and family history were not provided. On an unknown date, the minor patient received (dose 1) of 0.7 ml of suspect influenza quadrival A-B high dose HV vaccine, (lot 370679: formulation, strength and expiry date; unknown) via nasal route in unknown administration site for Immunization with no reported adverse event (product administered to patient of inappropriate age) (unknown latency). On an unknown date the was supposed to receive the flumist nasal spray but she grabbed the fluzone qiv hd and inadvertently gave it to intranasally with no reported adverse event (wrong product administered) (incorrect route of product administration) (unknown latency). This suspected adverse reaction report is submitted and classified as a medication error solely and exclusively to ensure the marketing authorization holder's compliance with the requirements set out in the Directive 2001/83/EC and Module VI of the Good Pharmacovigilance Practices. The classification as a medical error is in no way intended, nor should it be interpreted or construed as an allegation or claim made by the marketing authorization holder that any third party has contributed to or is to be held liable for the occurrence of this medication error.

Datasets	Entities	Train	Valid	Test	Total
MTSamples	Medical problem	538	203	199	940
	Treatment	149	43	35	227
	Test	120	39	50	209
VAERS	Investigation	148	29	59	236
	Nervous adverse event	406	83	162	651
	Other adverse event	301	62	167	530
	Procedure	338	57	126	521