







## DATASETS, BENCHMARKS, AND PROTOCOLS

# CORAL: Expert-Curated Oncology Reports to Advance Language Model Inference

Madhumita Sushil , M.Sc., Ph.D.,<sup>1</sup> Vanessa E. Kennedy , M.D.,<sup>2</sup> Divneet Mandair , M.D.,<sup>2</sup> Brenda Y. Miao , B.A.,<sup>1</sup> Travis Zack , M.D., Ph.D.,<sup>1,2</sup> and Atul J. Butte , M.D., Ph.D.<sup>1,2,3,4</sup>

Received: September 1, 2023; Revised: December 27, 2023; Accepted: January 21, 2024; Published: March 13, 2024

## Abstract

**BACKGROUND** Both medical care and observational studies in oncology require a thorough understanding of a patient's disease progression and treatment history, often elaborately documented within clinical notes. As large language models (LLMs) are being considered for use within medical workflows, it becomes important to evaluate their potential in oncology. However, no current information representation schema fully encapsulates the diversity of oncology information within clinical notes, and no comprehensively annotated oncology notes exist publicly, thereby limiting a thorough evaluation.

**METHODS** We curated a new fine-grained, expert-labeled dataset of 40 deidentified breast and pancreatic cancer progress notes at the University of California, San Francisco, and assessed the abilities of three recent LLMs (GPT-4, GPT-3.5-turbo, and FLAN-UL2) in *zero-shot* extraction of detailed oncological information from two narrative sections of clinical progress notes. Model performance was quantified with BLEU-4, ROUGE-1, and exact-match (EM) F1 score metrics.

**RESULTS** Our team annotated 9028 entities, 9986 modifiers, and 5312 relationships. The GPT-4 model exhibited overall best performance, with an average BLEU score of 0.73, an average ROUGE score of 0.72, an average EM F1 score of 0.51, and an average accuracy of 68% (expert manual evaluation on subset). Notably, GPT-4 was proficient in tumor characteristic and medication extraction and demonstrated superior performance in advanced reasoning tasks of inferring symptoms due to cancer and considerations of future medications. Common errors included partial responses with missing information and hallucinations with note-specific information.

**CONCLUSIONS** By developing a comprehensive schema and benchmark of oncology-specific information in oncology notes, we uncovered both the strengths and the limitations of LLMs. Our evaluation showed variable zero-shot extraction capability among the GPT-3.5-turbo, GPT-4, and FLAN-UL2 models and highlighted a need for further improvements,

*Drs. Zack and Butte contributed equally to this article.*

*The author affiliations are listed at the end of the article.*

*Dr. Sushil can be contacted at [Madhumita.Sushil@ucsf.edu](mailto:Madhumita.Sushil@ucsf.edu) or at Bakar Computational Health Sciences Institute, 490 Illinois Street, Cubicle 2215, 2nd FL., North Tower, San Francisco, CA 94143.*

particularly in complex medical reasoning, before performing reliable information extraction for clinical research and complex population management and documenting quality patient care. (Funded by the National Institute of Health, Food and Drug Administration and others.)

## Introduction

Cancer care is complex, often involving multiple treatments across different institutions. Most of this complexity is only captured within the textual format of an oncologist's clinical note. Optimal clinical decision-making, as well as research studies based on real-world data, requires a nuanced and detailed understanding of this complexity, naturally leading to widespread interest in oncology information extraction research.<sup>1</sup> Recently, large language models (LLMs) have shown impressive performance on several natural language processing (NLP) tasks in medicine, including obtaining high scores on United States Medical Licensing Examination questions,<sup>2,3</sup> medical question answering,<sup>4</sup> promising performance for medical consultation, diagnosis, and education,<sup>5</sup> identifying key findings from synthetic radiology reports,<sup>6</sup> biomedical evidence and medication extraction,<sup>7</sup> and breast cancer recommendations.<sup>8</sup> However, due to the lack of publicly available and comprehensively annotated oncology datasets, the analysis of these LLMs for information extraction and reasoning in real-world oncology data remains fragmented and understudied.

To date, prior studies on oncology information extraction have focused either on elements represented within ICD-O3 codes or cancer registries<sup>9,10</sup> or on a subset of cancer- or problem-specific information.<sup>11-15</sup> No existing information representation and annotation schema is adept enough to encompass comprehensive textual oncology information in a manner that is agnostic as to note type, problem, and disease. Although similar frameworks are being created for tabular oncology data,<sup>16</sup> efforts for textual data sources have been limited to pilot studies,<sup>17</sup> surveys of oncology elements studied across different research contributions,<sup>18,19</sup> and domain-specific schemas.<sup>20</sup>

In this research, we aimed to develop an expert-labeled oncology note dataset to enable the evaluation of LLMs in extracting clinically meaningful, complex concepts and relations. We do this by developing a schema and guidelines for

comprehensively representing and annotating textual oncology information, creating a dataset of 40 oncology progress notes labeled according to this schema, and benchmarking the baseline performance of the recent LLMs for zero-shot extraction of oncology information (that is, extraction without any previous training on the task). Sample workflow is demonstrated in Figure S1 in the Supplementary Appendix.

## Methods

### ONCOLOGY-SPECIFIC INFORMATION REPRESENTATION SCHEMA

To holistically represent oncology information within clinical notes, we developed a detailed schema based on a hierarchical, conceptual structure of oncology information (also called frame semantics),<sup>17,18,20</sup> agnostic to cancer types and note types under consideration. It comprises the following broad concepts: patient characteristics, temporal information, location-related information, test-related information, test results-related information, tumor-related information, treatment-related information, procedure-related information, clinical trial, and disease state.

Broad concepts further encompassed expert-determined fine-grained concepts. For example, radiology test, genomic test, and diagnostic laboratory test were represented within the "tumor test" category. The schema was implemented through three annotation modalities: entities or phrases of specific types, attributes or modifiers of entities, and relations between entity pairs. These relations could be descriptive (for example, relating a biomarker name to its results), temporal (for example, indicating when a test was conducted), or advanced (for example, relating a treatment to resulting adverse events). Together, the schema comprised 59 unique entities, 23 attributes, and 26 relations (Table S1 in the Supplementary Appendix).

The concepts and relationships annotated within this new schema incorporate nuanced details such as symptom history attributed to the diagnosed cancer, clinical trials considered for patient enrollment, genomic findings, reasons for switching treatments, and detailed social history of the patient not otherwise represented in cancer registries or a structured medical record. These concepts and relationships are significantly more inclusive and specific than those extracted by existing clinical NLP pipelines such as cTakes<sup>21</sup> and DeepPhe.<sup>22</sup> An openly available clinical NLP model, Stanza,<sup>23,24</sup> was used to prehighlight problems,

treatments, and tests within text to aid the annotators. Elaborate annotation guidelines are provided in the supplementary materials, and the annotation schema in the format of an open-source annotation software, BRAT (<https://brat.nlplab.org/>), is shared along with the source code at <https://github.com/MadhumitaSushil/OncLLMExtraction>.

## DATA

We collected data for 20 breast cancer patients and 20 pancreatic cancer patients from the University of California, San Francisco (UCSF) Information Commons, containing patient data between 2012 and 2022, deidentified as previously described.<sup>25</sup> All dates within notes were shifted by a random patient-level offset to maintain anonymity.<sup>25</sup> Only patients with corresponding tabular staging data, documented disease progression, and an associated medical oncology note were considered for document sampling. Some gene symbols, clinical trial names, and cancer stages were inappropriately redacted in our automated handling, and these were manually added back to the clinical notes under the UCSF Institutional Review Board numbers 18-25163 and 21-35084.

These two diseases were chosen for their dissimilarity. Breast cancer is frequently curable and heavily reliant on biomarker and genetic testing and treatment plans integrating radiation, surgical, and medical oncology. Pancreatic cancer has high mortality rates, and its treatment involves highly toxic traditional chemotherapy regimens.

All narrative sections except direct copy-forward of radiology and pathology reports were annotated, using the knowledge schema described above, by one of two oncology fellows and/or a medicine student. This process produced a final corpus of 40 expert-annotated clinical notes, which is available freely through the controlled-access repository *PhysioNet*. (Note: Users will need to create an account on <https://physionet.org/> and sign a data use agreement before downloading the dataset, which is available at <https://physionet.org/content/curated-oncology-reports/1.0/>. A valid Collaborative Institutional Training Initiative [CITI] certificate will need to be uploaded before signing the data use agreement.)

An additional 100 notes each for breast and pancreatic cancer were automatically labeled by GPT-4 using the same prompts as the benchmarking tests in this study.

## ZERO-SHOT LLM EXTRACTION BASELINE

To establish the baseline capability of LLMs in extracting detailed oncological history, we evaluated three recent LLMs without any task-specific training (i.e., “zero-shot” extraction): the GPT-4 model,<sup>26</sup> the GPT-3.5-turbo model (base model for the ChatGPT interface<sup>27</sup>), and the openly available foundation model FLAN-UL2<sup>28</sup> on the following tasks derived from two narrative sections of clinical progress notes for breast and pancreatic cancer: *History of Present Illness* (HPI) and *Assessment and Plan* (A&P):

1. **Symptom presentation:** Identify all symptoms experienced by the patient, symptoms present at the time of cancer diagnosis, and symptoms experienced due to the diagnosed cancer, all further related to the date or time (datetime) of their occurrence.
2. **Radiology tests:** List radiology tests conducted for the patient paired with their datetime, site of the test, medical indication for the test, and the test result.
3. **Genomic tests:** List genetic and genomic tests conducted for the patient paired with the corresponding datetime and the test result.
4. **First cancer diagnosis date:** Infer the datetime for the first diagnosis of cancer for the patient.
5. **Tumor characteristics:** Extract tumor characteristics in the following groups: biomarkers, histology, stage (using either of two common staging systems, numeric or tumor-node-metastasis, commonly called TNM), grade, and metastasis (along with the site of metastasis and the procedure that diagnosed metastasis), all paired with their datetime.
6. **Administered procedures:** Identify all interventional procedures conducted for the patient paired with their datetime, site, medical indication, and outcome.
7. **Prescribed medications:** List medications prescribed to the patient, linked to the beginning datetime, end datetime, reason for prescription, continuity status (continuing, finished, or discontinued early), and any hypothetical or confirmed adverse events attributed to the medication.
8. **Future medications:** Infer medications that are either planned for administration or discussed as a potential option, paired with their consideration (planned or hypothetical) and potential adverse events discussed in the note.

We used the generative pretrained transformer (GPT) models via the Health Insurance Portability and Accountability

Act (HIPAA)-compliant Microsoft Azure OpenAI studio and application programming interface, so that no data were permanently transferred to or stored by Microsoft or OpenAI. Separately, we implemented the openly available FLAN-UL2 model on the internal computing environment. Model inputs were provided in the format `{system role description} {note section text, prompt}`. GPT model settings and task-specific prompts are provided in the Supplementary Appendix. Examples of structured output format were provided with prompts to enable automated evaluation.

## EVALUATION

An automatic quantitative evaluation compared the manually annotated, related entity pairs to the corresponding model output. Because LLMs generate free-text outputs to represent entity mentions and their relations, model performance was quantified using two evaluation metrics for comparing pairs of text: BLEU-4 with smoothing<sup>29</sup> and ROUGE-1.<sup>30</sup> BLEU and ROUGE metrics quantify the overlap between sequences of  $n$  words (or  $n$ -grams) in the generated output and the reference annotations (see the Supplementary Appendix). The BLEU score quantifies the precision of  $n$ -grams between the model output and reference annotations while also penalizing very short outputs compared with references, and the ROUGE score similarly quantifies the recall of  $n$ -grams by comparing annotated snippets to model-generated answers to penalize the model when annotations are not included in outputs.

The exact-match F1 score between model outputs and annotated phrases was quantified to evaluate the model's ability to generate lexically identical outputs. Of note is that the exact-match metric is overly strict; for example, *ER+ : 2020* and *ER positive: 2020* are considered separate answers for exact-match scores. Additionally, the accuracy of the best-performing model was quantified for 11 entity extraction tasks and 20 relation extraction tasks on a random subset of half of the notes — 10 notes from breast cancer and 10 from pancreatic cancer — through review by an independent oncologist. Model outputs were presented to the oncologist as tables of the requested information by first parsing the GPT-4 model output automatically to populate the table. For example, the table for radiology test included columns for test name, date-time of the test, site of the test, reason for the test, and the test results. The expert assessment divided each output into three broad categories, correct, partially correct, and incorrect, which were aggregated to compute model

accuracy. Partially correct or incorrect outputs were further subcategorized based on types of errors (see the Supplementary Appendix).

## Results

### BENCHMARKING DATASET CREATION

Across 40 breast and pancreatic cancer progress notes, 9028 entities, 9986 entity attributes, and 5312 relationships were annotated, demonstrating the high density of clinically relevant information in the complex medical oncology narratives. Patient demographics are presented in [Table 1](#), and a sample of the annotated documentation is presented in [Figure 1](#). Manual annotation was time-consuming; it took the oncology fellows 88 hours to annotate 27 documents, the medical student 50 hours to annotate 17 documents (4 of which were also annotated by a fellow), and the independent reviewer 116 hours to review potential errors in annotating all 40 documents. The mean interannotator agreement of entities — computed as the mean F1 score of overlap between their entity spans<sup>31</sup> — was 0.81, indicating high agreement.

Description of initial cancer diagnosis and disease and treatment progression were elaborately represented within the annotations ([Fig. 2](#)). As anticipated, pancreatic cancer notes presented more palliative and supportive treatment entities and more symptoms attributable to cancer. Conversely, breast cancer notes contained more diagnostic and staging tests, as well as laterality, lymph node involvement, and biomarkers, reflecting the more complex diagnostic and staging workup required for this disease. Temporal relations were common (1200 relations), as were indications of findings from a test or procedure (566 relations) and relations attributing adverse events to pre-existing conditions or treatments (232 relations).

### INFORMATION EXTRACTION

#### *GPT-4 Outperforms the GPT-3.5 Model and the FLAN-UL2 Model*

The GPT-4 model performed better than the GPT-3.5-turbo and FLAN-UL2 models ([Fig. 3](#)), demonstrating an average BLEU score of 0.73, an average ROUGE score of 0.72, and an average exact-match F1 (EM-F1) score of 0.51 compared with the average BLEU, ROUGE, and EM-F1 scores of 0.61, 0.58, and 0.29 respectively, for the

Table 1. Demographic Distribution for the Annotated Data Cohort Comprising 40 Patients, Additionally Stratified by Disease Group.			
Demographic Property and Category	Breast Cancer	Pancreatic Cancer	All
Race/ethnicity			
Native Hawaiian or Other Pacific Islander	2	0	2
Latinx	3	3	6
Native American or Alaska Native	1	0	1
Southwest Asian and North African	1	0	1
Black or African American	4	4	8
Asian	4	4	8
Multirace/ethnicity	0	3	3
Other	1	2	3
White	4	4	8
Unknown/declined	0	0	0
Sex			
Male	0	10	10
Female	20	10	30
Age group, yr			
Under 30	0	0	0
31–40	6	0	6
41–50	3	1	4
51–60	6	4	10
61–70	2	8	10
71–80	2	6	8
81–89	1	1	2
89+	0	0	0

GPT-3.5-turbo model and 0.53, 0.27, and 0.06, respectively, for the FLAN-UL2 model. We additionally experimented with the clinical-T5-large model<sup>32</sup> and LLaMA 7B, LLaMA 13B, and LLaMA-2 13B models.<sup>33</sup> However, we did not obtain reasonable outputs from these models for our task, presumably because they are not tuned on task descriptions, which is known to improve the instruction-following abilities of LLMs.<sup>34</sup>

Performance differences were notable for tasks requiring advanced reasoning, for example, inferring symptoms present at the time of cancer diagnosis and hypothetical discussions of future medications.

### *Best Synthesis of Tumor Characteristics and Medication History*

High scores — 0.95 BLEU, 0.93 ROUGE, 0.91 EM-F1 — were obtained by the GPT-4 model when extracting tumor grade paired with temporal information. High performances were also obtained in extracting cancer summary stage (0.85 BLEU, 0.80 ROUGE, 0.69 EM-F1), TNM

stage (0.82 BLEU, 0.78 ROUGE, 0.71 EM-F1), and future medication (0.88 mean BLEU, 0.84 mean ROUGE, 0.70 mean EM-F1), suggesting promising capabilities in the automated extraction of these parameters.

The GPT-4 model demonstrated good performance with potential for further improvements in extracting genomics datetime and results (0.81 mean BLEU, 0.80 mean ROUGE, 0.68 mean EM-F1), radiology tests with their datetime (0.80 BLEU and ROUGE, 0.52 EM-F1), the prescribed medications with their start datetime, end datetime, current continuity status, potential adverse events, and adverse events experienced due to the medication (0.80 mean BLEU, 0.76 mean ROUGE, 0.57 EM-F1), symptoms with their datetime (0.71 BLEU, 0.74 ROUGE, 0.45 EM-F1), metastasis extraction (0.64 mean BLEU, 0.65 mean ROUGE, 0.44 EM-F1), and for identifying symptoms that occurred due to cancer (0.67 BLEU, 0.75 ROUGE, 0.5 EM-F1). Lexical differences between annotated information and model outputs were common when extracting longer phrases, such as reasons for tests or test



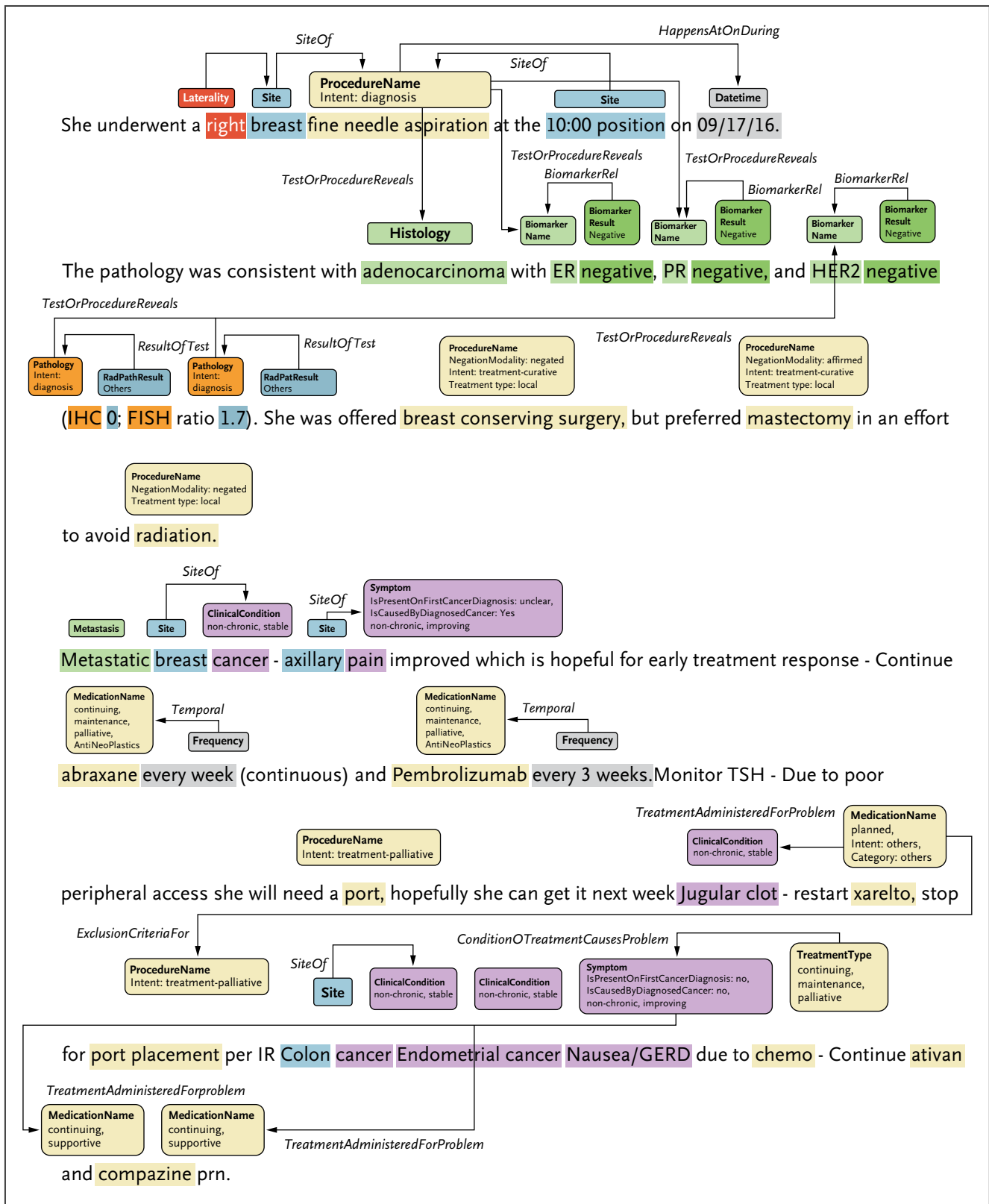


Figure 1. A Sample of the Annotated, Deidentified Medical Oncology Progress Notes.

The colored highlights refer to different types of entity spans within text. The arrows indicate the relations between the pair of entities linked. Within the box next to entity types, the corresponding modifier values for those entities are listed.

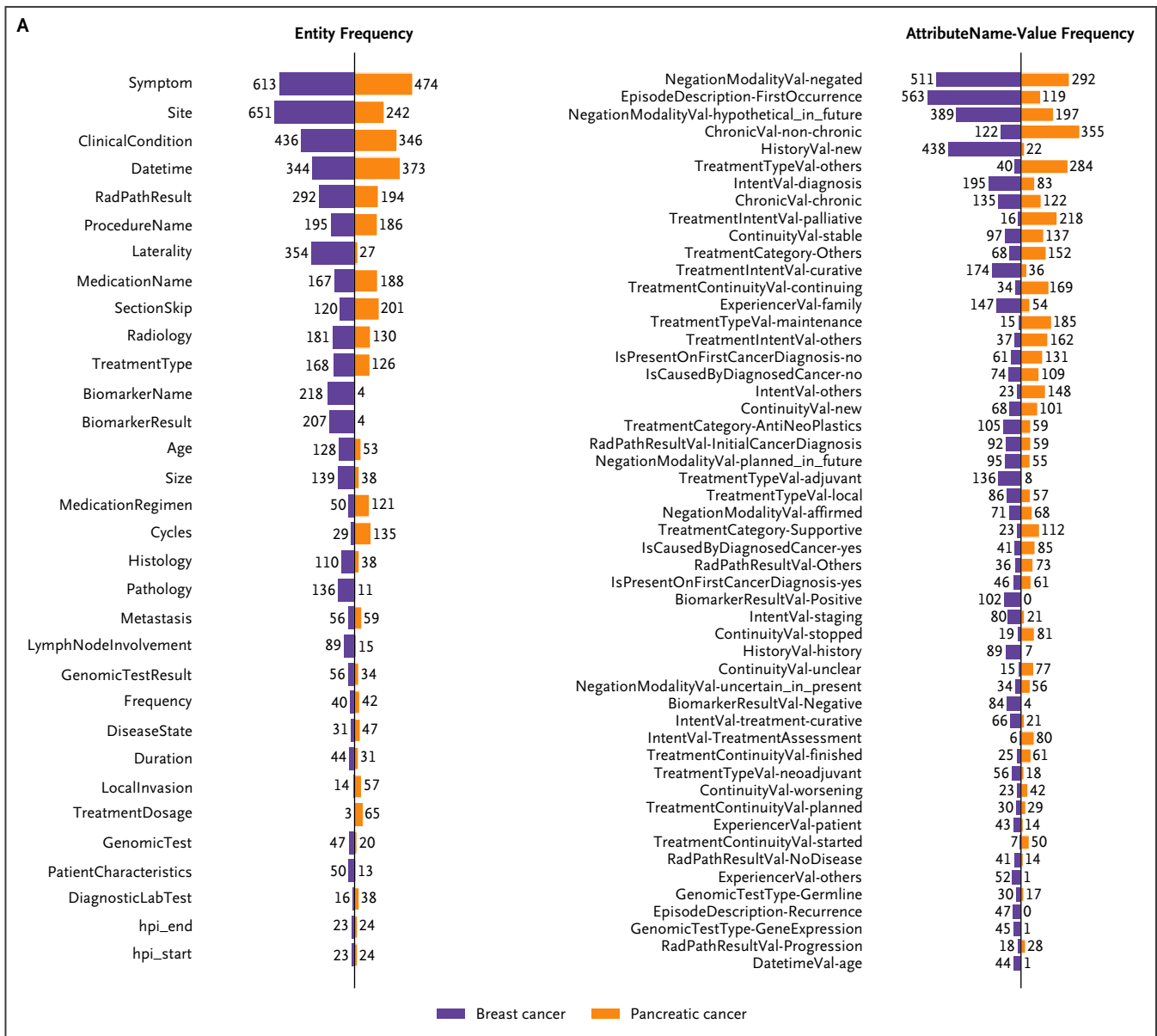


Figure 2. Distribution of (Panel A) the Annotated Entity Mentions and the Attribute Values for These Entities and (Panel B) Temporal, Descriptive, and Advanced Relations in the Annotated Corpus of Breast and Pancreatic Cancer Medical Oncology Progress Notes.

results, compared with short responses such as cancer grade or stage. The open-source FLAN-UL2 model and the GPT-3.5 model demonstrated higher lexical differences than the GPT-4 model, as evident from significantly lower EM-F1 scores. When extracting histological subtypes and treatment-relevant tumor biomarkers, the models frequently provided more information than necessary,

for example providing grade, stage, and biomarkers of a tumor in addition to the requested histological subtype. The poorest quantitative performance was obtained for procedure extraction (0.58 mean BLEU, 0.57 mean ROUGE, 0.33 EM-F1). GPT-4 model performance was comparable across the two cancer types, although the model demonstrated marginally better performance in

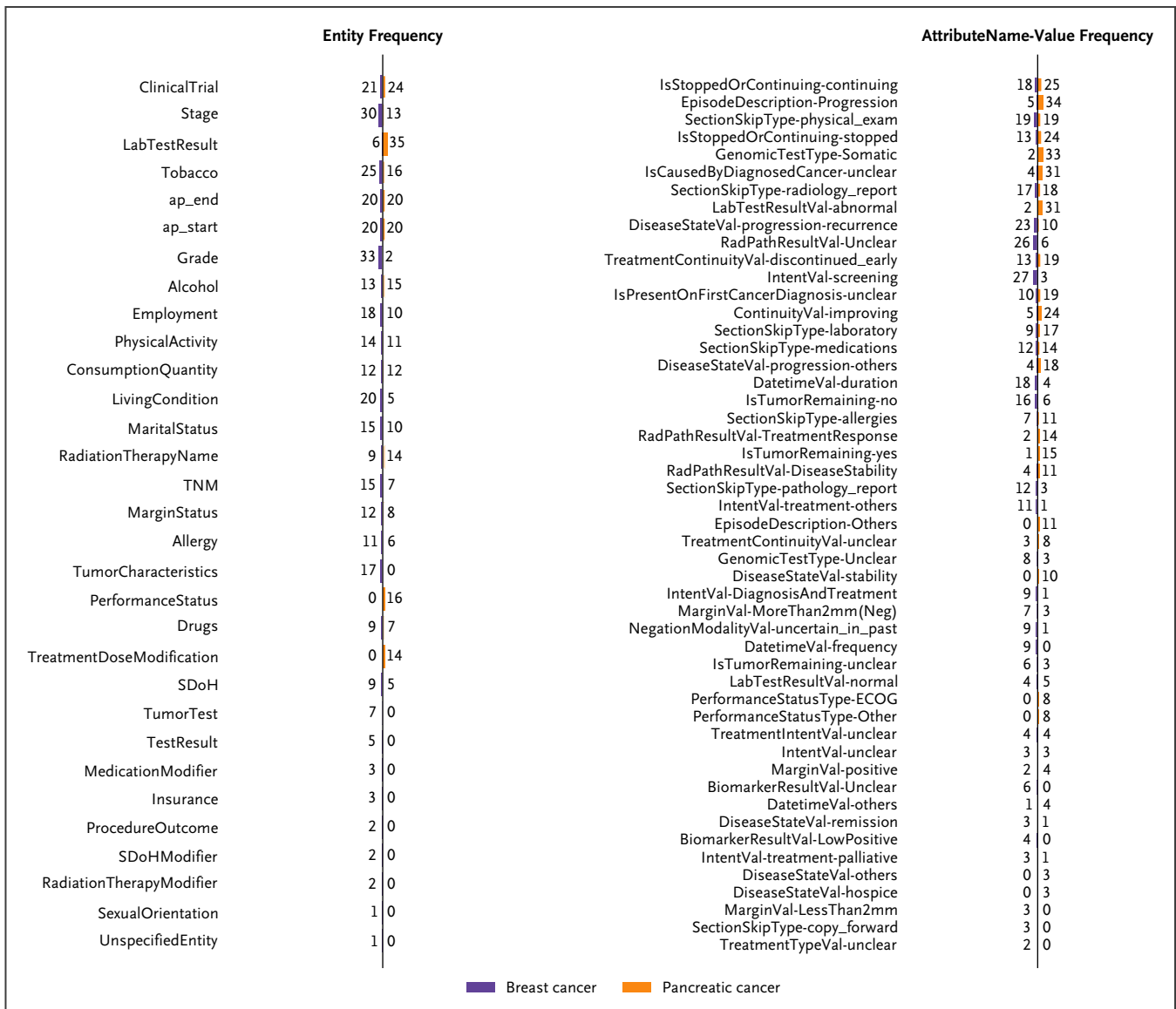


Figure 2. Continued.

medication and biomarker extraction for breast cancer and in genomic and procedure extraction for pancreatic cancer (Fig. S2).

We did not find any significant differences in model performance across either male and female genders or different races and ethnicities for any of the three models for any of the three metrics, as computed with the Kruskal-Wallis test with false-discovery-rate correction by the Benjamini-Hochberg method (Tables S2 and S3).

### Expert Evaluation Confirms Superior Oncologic Information Extraction Ability

A medical oncologist additionally evaluated GPT-4 model outputs on a subset of 10 breast cancer and 10 pancreatic cancer notes for the first cancer diagnosis date, symptoms, radiology tests, procedures, histology, metastasis, and future medications (Table 2). It took the oncologist nearly 90 hours to quantify GPT-4 model accuracy across 31 categories for the 20 notes. The expert evaluations showed that the GPT-4 model outputs were overall 68% accurate.



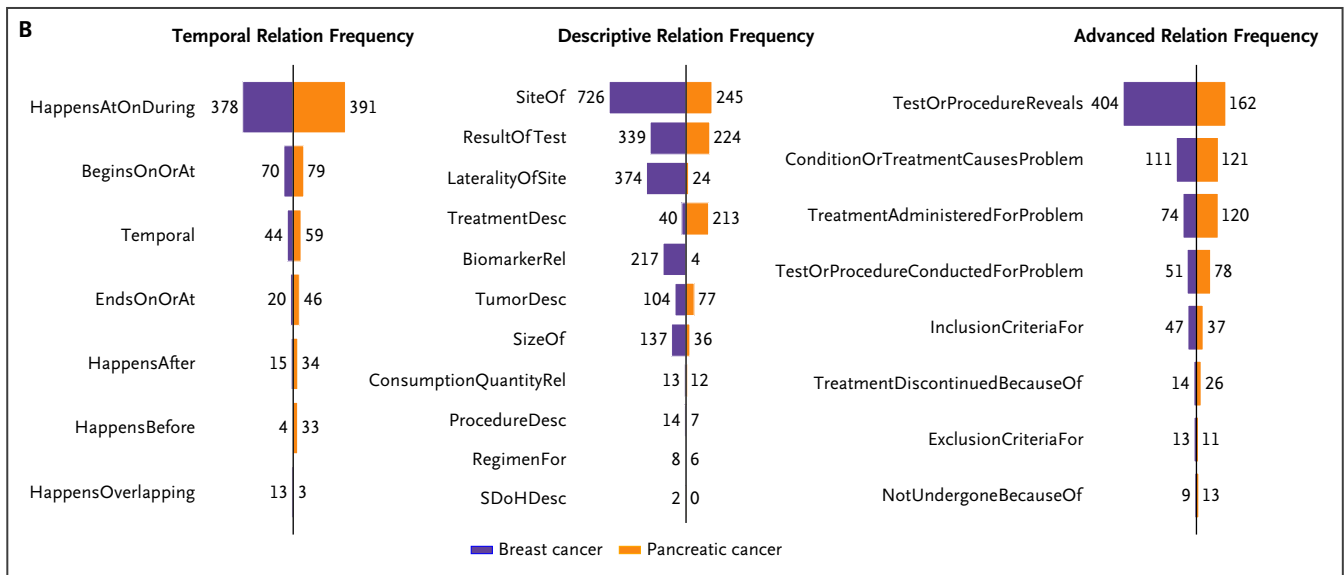


Figure 2. Continued.

An additional 3% of the cases were correct but missing some desired information. Another 1% of the cases were deemed to be uncertain due to linguistic ambiguity. These findings support the automated quantitative evaluations, highlighting the excellent oncologic information extraction ability of the GPT-4 model. The most common error (22%) occurred in cases in which the model produced output from note text that did not correspond to the requested information (hallucinations 1; [Table 2](#), error category 3b). Of the remaining errors, in 6% of the cases, the model returned “Unknown” instead of correct answers, and in 1% of the cases the model fabricated information (hallucinations 2; [Table 2](#), error category 3c).

Partial correct answers with some missing information were most frequent for tumor histology and radiology test results. Hallucinations from note text included incorrectly categorizing biomarkers, genomic tests, procedures, and radiology tests and incorrect inferences of information such as symptoms present at the time of cancer diagnosis or symptoms caused due to the diagnosed cancer. The model produced information without direct references in text most frequently when inferring whether a medication was considered hypothetically or was planned for administration and in identifying the reasons for tests and procedures, for example specifying that positron emission tomography/computed tomography was conducted to evaluate metabolic activity.

Finally, the most common cases in which the model produced “Unknown,” despite the correct information being present in the note, were mentions of cancer histology, future medications, and symptoms due to cancer.

## Discussion

The schema presented in this article, coupled with its associated annotated dataset, offers a robust benchmark for assessing LLM performance against human specialist curators in extracting complex details from medical oncology notes. Across 40 patient consultation notes, we manually annotated 9028 relevant entities, 9986 attributes, and 5312 relations, highlighting the information-dense nature of these notes. Our schema facilitated capturing the nuanced rhetoric in medical oncology narratives, spanning information such as family history, disease-relevant objective and temporal data, social determinants of health factors, causality between diagnoses, treatments, and symptoms, and treatment intent and response, including potential and current adverse events. This new, rich dataset of real-world oncology progress notes will enable several follow-up advances in language models for oncology.

The dataset facilitated the benchmarking of the zero-shot capability of LLMs in oncologic history summarization. It demonstrated the surprising zero-shot capability of the

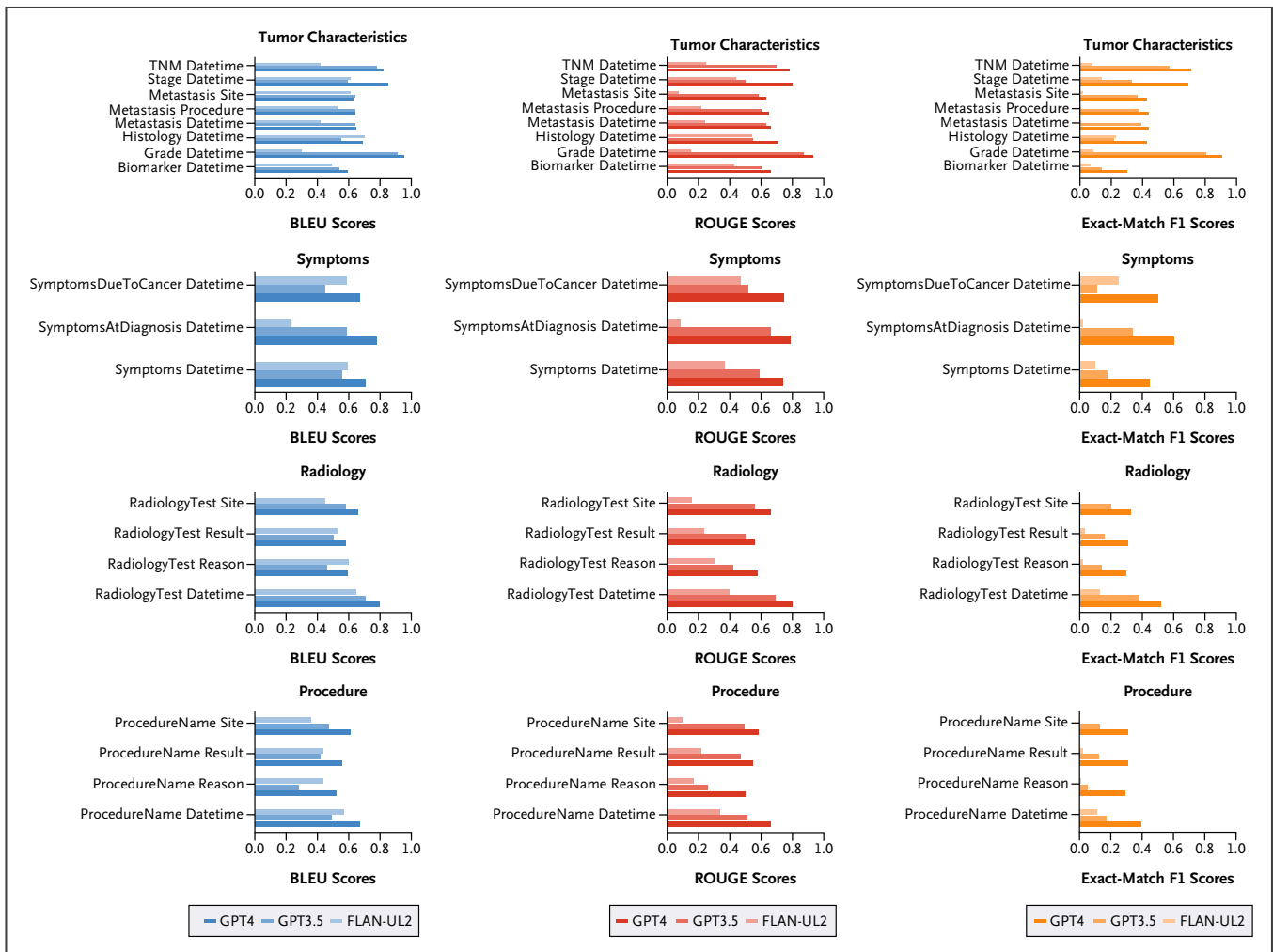


Figure 3. Mean BLEU, ROUGE, and Exact-Match Scores.

BLEU scores (precision-focused; first column), ROUGE scores (recall-focused; second column), and exact-match F1 score (third column) for entity-relation extraction in different oncology tasks, grouped by the category of inference. *Entity1 Entity2* on the y-axis represents the performance for extracting the mention of entities *Entity1* and *Entity2* and thereby correctly inferring their relationship.

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 and the reason for their prescription. The model, however, also showed room for improvement in causal inference, such as inferring whether a symptom was caused by cancer. An open-source counterpart, the FLAN-UL2 model, demonstrated high precision but low recall, suggesting that it may be a promising alternative to the proprietary GPT-4 model if fine-tuned further on in-domain data.

Although current zero-shot performances are impressive because no task-specific fine-tuning was performed, the

obtained accuracy may not be directly usable in clinical settings. Meanwhile, manual information extraction from notes is time-consuming, which contributes to an underutilization of NLP in electronic health record (EHR)-based observational research.<sup>35</sup> To obtain research-usable capability for oncology information extraction with minimal manual involvement, it is promising to explore strategies such as few-shot learning, where a few annotated examples are provided to the model to learn better,<sup>36</sup> advanced prompt designs such as chain-of-thought prompting<sup>37</sup> to benefit reasoning and selection-inference prompting<sup>38</sup> to first select the relevant entities before inferring more advanced relations and entity attributes, and in-domain fine-tuning for adapting open-source models to clinical

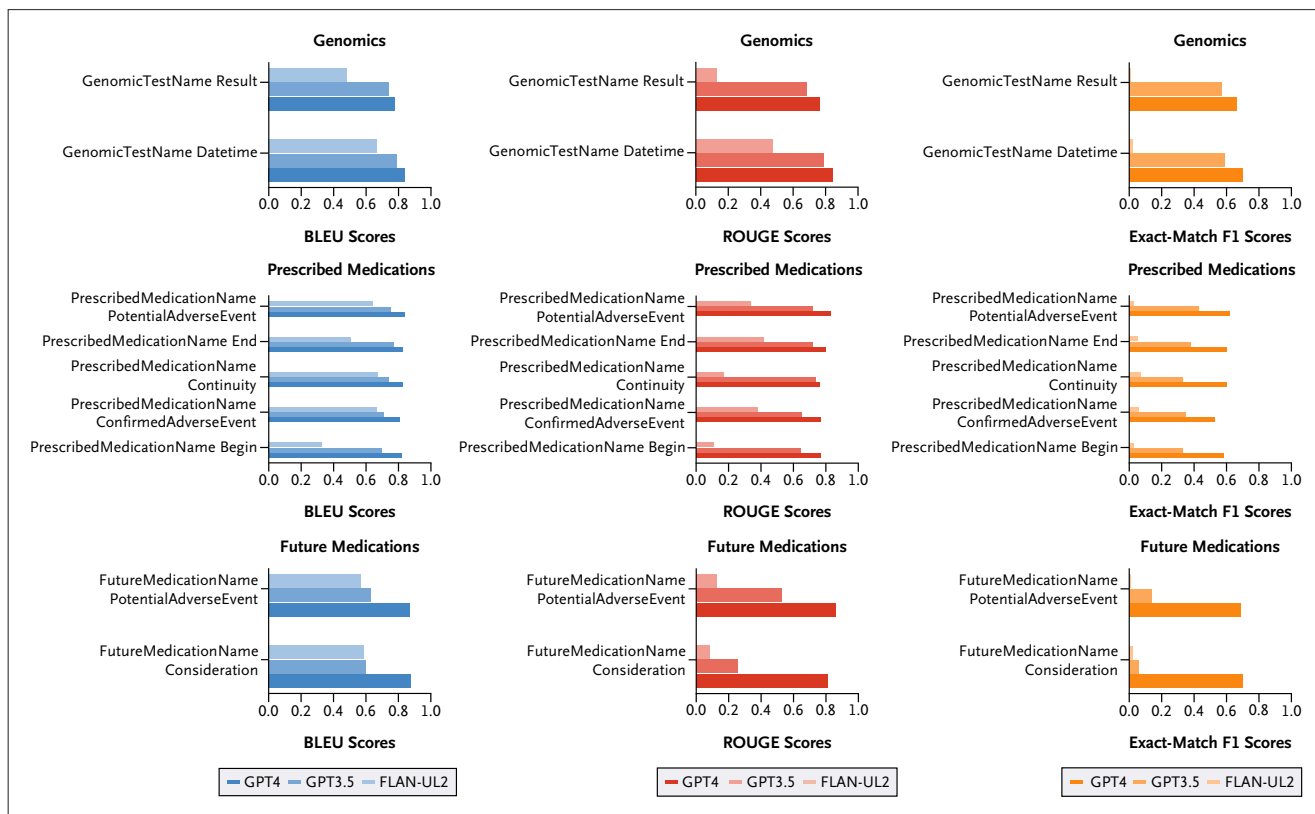


Figure 3. Continued.

domains. This dataset would be a critical resource for follow-up benchmarking studies, reducing the need for prior extensive domain-specific annotations for validating the performance of novel clinical LLMs and prompting strategies. Although the annotations in the dataset may be less verbose than model outputs, and annotator fatigue may contribute to additional errors, expert manual evaluations corroborated the findings of automated evaluations, demonstrating the reliability of the findings.

Although our study did not uncover any statistically significant disparities in model performance across sex and race/ethnicity, sample sizes in this study may be insufficient for this analysis, and there may be systematic biases in model performance that need to be studied further on larger datasets. Finally, small changes in prompts can result in a big impact on model performance. Further studies are needed to establish detailed guidelines with regard to prompt design and to quantify the impact of prompt engineering on model performance.

The current capability of LLMs in extracting tumor characteristics, medication, and adverse drug events demonstrated promise for enhanced postapproval real-world drug and device safety monitoring from unstructured data, automatically populating population-wide cancer registries, text-based cohort selection for EHR-based research studies, and speeding up the matching of patients to clinical trial criteria. Easier access to text data, potentially facilitated by LLMs in a human-in-the-loop setting, will improve research on patients' outcomes and public health by providing evidence for better data-driven guidelines and incorporating text-based variables in previously unutilized ways.

Although this dataset represents a small number of patients, the number of annotated sentences and oncologic concepts is large, making it comparable in sample size to existing benchmarking clinical NLP datasets, but much larger in breadth. An additional set of 200 GPT-4-labeled notes would facilitate larger benchmarking

**Table 2. Expert Manual Evaluation of GPT-4 Outputs for a Subset of Inference Categories on a Subset of 10 Breast Cancer Notes and 10 Pancreatic Cancer Notes.\***

Entity/Relation and Inference Category	Sample Size	Correct	Partially Correct			Incorrect			
		1	2a	2b	2c	3a	3b	3c	3d
<b>Entity</b>									
FirstCancerDiagnosis	40	0.88	0.00	0.05	0.00	0.00	0.08	0.00	0.00
Symptoms	140	0.84	0.04	0.01	0.00	0.00	0.06	0.00	0.04
SymptomsDueToCancer	129	0.83	0.00	0.01	0.01	0.02	0.08	0.00	0.05
FutureMedication	91	0.77	0.04	0.02	0.00	0.00	0.05	0.00	0.11
SymptomsAtDiagnosis	63	0.76	0.00	0.00	0.00	0.00	0.24	0.00	0.00
Metastasis	51	0.63	0.14	0.00	0.00	0.00	0.20	0.00	0.04
RadiologyTest	116	0.53	0.27	0.00	0.00	0.00	0.13	0.00	0.07
Procedure	107	0.51	0.15	0.01	0.01	0.01	0.27	0.01	0.03
Biomarker	154	0.51	0.05	0.01	0.01	0.00	0.40	0.00	0.03
GenomicTestName	84	0.50	0.00	0.00	0.00	0.00	0.46	0.00	0.04
Histology	73	0.25	0.01	0.58	0.00	0.00	0.03	0.00	0.14
<b>Relation</b>									
Metastasis - Datetime	51	0.90	0.00	0.00	0.00	0.00	0.06	0.00	0.04
SymptomsAtDiagnosis - Datetime	63	0.87	0.00	0.00	0.00	0.00	0.11	0.00	0.02
Metastasis - Site	51	0.86	0.02	0.00	0.00	0.00	0.08	0.00	0.04
Histology - Datetime	72	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.14
Metastasis - Procedure	51	0.80	0.06	0.04	0.00	0.00	0.06	0.00	0.04
Symptoms Datetime	140	0.77	0.00	0.06	0.00	0.01	0.09	0.00	0.06
RadiologyTest - Datetime	116	0.76	0.00	0.00	0.00	0.03	0.14	0.00	0.07
FutureMedication - PotentialAdverseEvent	91	0.74	0.00	0.02	0.00	0.04	0.09	0.00	0.11
SymptomsDueToCancer - Datetime	129	0.73	0.01	0.07	0.00	0.02	0.09	0.00	0.09
FutureMedication - Consideration	91	0.73	0.00	0.00	0.00	0.00	0.09	0.08	0.11
Procedure - Datetime	107	0.67	0.00	0.07	0.00	0.02	0.19	0.00	0.05
Biomarker - Datetime	141	0.65	0.00	0.00	0.00	0.00	0.32	0.00	0.03
Procedure - Site	107	0.61	0.00	0.01	0.00	0.01	0.34	0.01	0.03
RadiologyTest - Reason	116	0.52	0.02	0.06	0.00	0.03	0.29	0.01	0.08
RadiologyTest - Result	116	0.52	0.02	0.23	0.00	0.00	0.16	0.00	0.07
GenomicTestName - Datetime	84	0.49	0.00	0.00	0.00	0.00	0.46	0.00	0.05
Procedure - Result	107	0.49	0.00	0.15	0.00	0.00	0.30	0.04	0.03
GenomicTestName - Result	84	0.46	0.00	0.00	0.00	0.00	0.46	0.04	0.04
Procedure - Reason	107	0.42	0.04	0.07	0.00	0.03	0.42	0.00	0.03
RadiologyTest - Site	116	0.40	0.01	0.01	0.00	0.03	0.47	0.01	0.09
Total/mean	2872	0.66	0.03	0.05	0.00	0.01	0.19	0.01	0.05

\* Each note was first divided into the *History of Present Illness* and *Assessment and Plan* sections, thereby resulting in inference over 40 note snippets. Each cell represents the model outputs as a fraction of GPT-4 outputs in that category. When only one entity is mentioned, for example in *Procedure*, the scores represent the extraction of that entity by itself. When two entities are separated with a hyphen, for example in *Procedure - Datetime*, the scores represent an evaluation of correctly linking the entities together, for example pairing the *procedure* with the right *date or time of the procedure*. Partially correct answer categories are defined as follows: 2a, The output contains more information than necessary, and the extra information is correct; 2b, The output is correct, but some information is missing from the output; 2c, The output contains more information than necessary, but the extra information is incorrect. Incorrect answer categories are defined as follows: 3a, Independent expert reviewer determines that the note text is ambiguous, where either manual annotation or model answer could be considered correct; 3b, Hallucinations 1: the model answers from the information mentioned in the note, but the answer is incorrect for the question asked; 3c, Hallucinations 2: the model fabricates information not discussed in the text; 3d, Correct output is present in the input text, but the model returns unknown.

studies through further expert analysis and would also enable fine-tuning of open-source models with weak labels (model distillation). Finally, although we used the data from only two cancers within a single academic institution, the information representation and annotation schema were designed to be both cancer and institution agnostic, which will facilitate an extension of the analysis to large multicenter, multicancer studies to obtain generalizable conclusions.

---

## Conclusions

We successfully created a benchmarking dataset of 40 expert-annotated breast and pancreatic cancer medical oncology notes by validating a new detailed schema for representing in-depth textual oncology-specific information, which is shared openly for further research. This dataset served as a testbed to benchmark the zero-shot extraction capability of three LLMs: GPT-3.5-turbo, GPT-4, and FLAN-UL2. We found that the GPT-4 model showed the best performance with an average of 0.73 BLEU, 0.72 ROUGE, and 0.51 EM-F1 scores, highlighting the promising capability of language models to summarize oncologic patient history and plan without substantial supervision and aid future research and practice by reducing manual efforts. With further prompt engineering and model fine-tuning, combined with minimal manual supervision, LLMs will potentially be usable to extract important facts from cancer progress notes needed for clinical research, complex population management, and documenting quality patient care.

## Disclosures

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This work was funded through the National Cancer Institute of the National Institutes of Health award number P30CA082103, Food and Drug Administration grant U01FD005978 to the UCSF-Stanford Center of Excellence in Regulatory Sciences and Innovation (CERSI), National Institutes of Health grant UL1 TR001872 to UCSF CTSI, and a philanthropic gift from Priscilla Chan and Mark Zuckerberg.

Author disclosures and other supplementary materials are available at [ai.nejm.org](https://ai.nejm.org).

This research would not have been possible without immense support from several people. We thank the University of California, San Francisco (UCSF) AI Tiger Team, Academic Research Services, Research Information Technology, and the Chancellor's Task Force for

Generative AI for their software development, analytical and technical support related to the use of Versa API gateway (the UCSF secure implementation of large language models and generative AI via API gateway), Versa chat (the chat user interface), and related data asset and services. We thank Boris Oskotsky and the Wynton high-performance computing platform team for supporting high-performance computing platforms that enable the use of large language models with deidentified patient data. We thank Michelle Wang for help with annotator recruitment and Binh Cao for data annotations. We thank Jennifer Creasman, Alysa Gonzales, Dalia Martinez, and Lakshmi Radhakrishnan for help with correcting clinical trial and gene name redactions. We thank Prof. Kirk Roberts for helpful discussions regarding frame semantics-based annotations of cancer notes, Prof. Artuur Leeuwenberg for discussions about temporal relation annotation, and all members of the Butte laboratory for helpful discussions in the internal presentations.

## Author Affiliations

- <sup>1</sup> Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco
- <sup>2</sup> Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco
- <sup>3</sup> Center for Data-driven Insights and Innovation, University of California Office of the President, Oakland
- <sup>4</sup> Department of Pediatrics, University of California, San Francisco, San Francisco

## References

1. Savova GK, Danciu I, Alamudun F, et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer Res* 2019;79:5463-5470. DOI: [10.1158/0008-5472.CAN-19-0579](https://doi.org/10.1158/0008-5472.CAN-19-0579).
2. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. March 20, 2023 (<https://arxiv.org/abs/2303.13375>). Preprint.
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198. DOI: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198).
4. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172-180. DOI: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2).
5. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;388:1233-1239. DOI: [10.1056/NEJMs2214184](https://doi.org/10.1056/NEJMs2214184).
6. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* 2023;307:e230725. DOI: [10.1148/radiol.230725](https://doi.org/10.1148/radiol.230725).
7. Agrawal M, Hagselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Goldberg Y, Kozareva Z, Zhang Y, eds. *Proceedings of the 2022 Conference on*



- Empirical Methods in Natural Language Processing. Kerrville, TX: Association for Computational Linguistics, 2022:1998-2022.
8. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* 2023;307:e230424. DOI: [10.1148/radiol.230424](https://doi.org/10.1148/radiol.230424).
  9. Alawad M, Yoon H-J, Tourassi GD. Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. In: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). Piscataway, NJ: IEEE, 2018:218-221.
  10. Breitenstein MK, Liu H, Maxwell KN, Pathak J, Zhang R. Electronic health record phenotypes for precision medicine: perspectives and caveats from treatment of breast cancer at a single institution. *Clin Transl Sci* 2018;11:85-92. DOI: [10.1111/cts.12514](https://doi.org/10.1111/cts.12514).
  11. Yala A, Barzilay R, Salama L, et al. Using machine learning to parse breast pathology reports. *Breast Cancer Res Treat* 2017;161:203-211. DOI: [10.1007/s10549-016-4035-1](https://doi.org/10.1007/s10549-016-4035-1).
  12. Odisho A, Park B, Altieri N, et al. Pd58-09 Extracting structured information from pathology reports using natural language processing and machine learning. *J Urol* 2019;201(Suppl 4):e1031-1032. DOI: [10.1097/01.JU.0000557177.97226.63](https://doi.org/10.1097/01.JU.0000557177.97226.63).
  13. Li Y, Luo Y-H, Wampfler JA, et al. Efficient and accurate extracting of unstructured EHRs on cancer therapy responses for the development of RECIST natural language processing tools: part I, the corpus. *JCO Clin Cancer Inform* 2020;4:383-391. DOI: [10.1200/CCI.19.00147](https://doi.org/10.1200/CCI.19.00147).
  14. Altieri N, Park B, Olson M, DeNero J, Odisho AY, Yu B. Supervised line attention for tumor attribute classification from pathology reports: higher performance with less data. *J Biomed Inform* 2021;122:103872. DOI: [10.1016/j.jbi.2021.103872](https://doi.org/10.1016/j.jbi.2021.103872).
  15. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *J Am Med Inform Assoc* 2022;29:1208-1216. DOI: [10.1093/jamia/ocac040](https://doi.org/10.1093/jamia/ocac040).
  16. Belenkaya R, Gurley MJ, Golozar A, et al. Extending the OMOP common data model and standardized vocabularies to support observational cancer research. *JCO Clin Cancer Inform* 2021;5:12-20. DOI: [10.1200/CCI.20.00079](https://doi.org/10.1200/CCI.20.00079).
  17. Roberts K, Si Y, Gandhi A, Bernstam E. A FrameNet for cancer information in clinical narratives: schema and annotation. In: Calzolari N, Choukri K, Cieri C, et al., eds. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Paris: European Language Resources Association, 2018:272-279.
  18. Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J Biomed Inform* 2019;100:103301. DOI: [10.1016/j.jbi.2019.103301](https://doi.org/10.1016/j.jbi.2019.103301).
  19. Mirbagheri E, Ahmadi M, Salmanian S. Common data elements of breast cancer for research databases: a systematic review. *J Family Med Prim Care* 2020;9:1296-1301. DOI: [10.4103/jfmmpc.jfmmpc\\_931\\_19](https://doi.org/10.4103/jfmmpc.jfmmpc_931_19).
  20. Datta S, Ulinski M, Godfrey-Stovall J, Khanpara S, Riascos-Castaneda RF, Roberts K. Rad-SpatialNet: a frame-based resource for fine-grained spatial relations in radiology reports. In: Calzolari N, Béchet F, Blache P, et al., eds. Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2020: 2251-2260.
  21. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17: 507-513. DOI: [10.1136/jamia.2009.001560](https://doi.org/10.1136/jamia.2009.001560).
  22. Savova GK, Tseytlin E, Finan S, et al. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Res* 2017;77:e115-e118. DOI: [10.1158/0008-5472.CAN-17-0615](https://doi.org/10.1158/0008-5472.CAN-17-0615).
  23. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a Python natural language processing toolkit for many human languages. March 16, 2020 (<http://arxiv.org/abs/2003.07082>). Preprint.
  24. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the Stanza Python NLP library. *J Am Med Inform Assoc* 2021;28:1892-1899. DOI: [10.1093/jamia/ocab090](https://doi.org/10.1093/jamia/ocab090).
  25. Radhakrishnan L, Schenk G, Muenzen K, et al. A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open* 2023;6:ooad045. DOI: [10.1093/jamiaopen/ooad045](https://doi.org/10.1093/jamiaopen/ooad045).
  26. Open AI. GPT-4 technical report. March 15, 2023 (<http://arxiv.org/abs/2303.08774>). Preprint.
  27. Open AI. ChatGPT. (<https://chat.openai.com>).
  28. Tay Y. A new open source Flan 20B with UL2. March 3, 2023 (<https://www.yitay.net/blog/flan-ul2-20b>).
  29. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D, eds. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Kerrville, TX: Association for Computational Linguistics, 2002:311-318.
  30. Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: Moens M-F, Szpakowicz S, eds. Text summarization branches out. Kerrville, TX: Association for Computational Linguistics, 2004:74-81.
  31. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;12:296-298. DOI: [10.1197/jamia.M1733](https://doi.org/10.1197/jamia.M1733).
  32. Lehman E, Hernandez E, Mahajan D, et al. Do we still need clinical language models? In: Mortazavi BJ, Sarker T, Beam A, Ho JC, eds. Proceedings of Machine Learning Research. Vol. 209. Machine Learning Research Press, 2023:578-597.
  33. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. February 27, 2023 (<https://arxiv.org/abs/2302.13971>). Preprint.
  34. Weller O, Lourie N, Gardner M, Peters ME. Learning from task descriptions. In: Webber B, Cohn T, He Y, Liu Y, eds. Proceedings of the 2020 Conference on Empirical Methods in Natural

- Language Processing (EMNLP). Kerrville, TX: Association for Computational Linguistics, 2020:1361-1375.
35. Fu S, Wang L, Moon S, et al. Recommended practices and ethical considerations for natural language processing-assisted observational research: a scoping review. *Clin Transl Sci* 2023;16:398-411. DOI: [10.1111/cts.13463](https://doi.org/10.1111/cts.13463).
36. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Cohn T, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in neural information processing systems* 33. Red Hook, NY: Curran Associates, 2020:1877-1901.
37. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022 ([https://openreview.net/pdf?id=\\_VjQlMeSB\\_J](https://openreview.net/pdf?id=_VjQlMeSB_J)).
38. Creswell A, Shanahan M, Higgins I. Selection-inference: exploiting large language models for interpretable logical reasoning. 2023 (<https://openreview.net/forum?id=3Pf3Wg6o-A4>).