

UCSF

UC San Francisco Previously Published Works

Title

A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports

Permalink

<https://escholarship.org/uc/item/06p015n5>

Journal

Journal of the American Medical Informatics Association, 31(10)

ISSN

1067-5027

Authors

Sushil, Madhumita

Zack, Travis

Mandair, Divneet

et al.

Publication Date

2024-10-01



DOI

10.1093/jamia/ocae146

Peer reviewed

Research and Applications

A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports

Madhumita Sushil , PhD^{*1}, Travis Zack, MD, PhD^{1,2}, Divneet Mandair, MD^{1,2}, Zhiwei Zheng, MEng³, Ahmed Wali, MEng³, Yan-Ning Yu, MEng³, Yuwei Quan, MEng³, Dmytro Lituiev , PhD¹, Atul J. Butte, MD, PhD^{1,2,4,5}

¹Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA 94158, United States, ²Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA 94158, United States, ³University of California, Berkeley, Berkeley, CA 94720, United States, ⁴Center for Data-driven Insights and Innovation, University of California, Office of the President, Oakland, CA 94607, United States, ⁵Department of Pediatrics, University of California, San Francisco, San Francisco, CA 94158, United States

*Corresponding author: Madhumita Sushil, PhD, Bakar Computational Health Sciences Institute, 490 Illinois Street, Cubicle 2215, 2nd Fl, North Tower, San Francisco, CA 94148, United States (Madhumita.Sushil@ucsf.edu)

Drs. Madhumita Sushil, Travis Zack, and Divneet Mandair share co-first authorship.

Zhiwei Zheng, Ahmed Wali, Yan-Ning Yu, and Yuwei Quan share equal contribution as a single team for Master of Engineering Capstone project at UC Berkeley.

Abstract

Objective: Although supervised machine learning is popular for information extraction from clinical notes, creating large annotated datasets requires extensive domain expertise and is time-consuming. Meanwhile, large language models (LLMs) have demonstrated promising transfer learning capability. In this study, we explored whether recent LLMs could reduce the need for large-scale data annotations.

Materials and Methods: We curated a dataset of 769 breast cancer pathology reports, manually labeled with 12 categories, to compare zero-shot classification capability of the following LLMs: GPT-4, GPT-3.5, Starling, and ClinicalCamel, with task-specific supervised classification performance of 3 models: random forests, long short-term memory networks with attention (LSTM-Att), and the UCSF-BERT model.

Results: Across all 12 tasks, the GPT-4 model performed either significantly better than or as well as the best supervised model, LSTM-Att (average macro F1-score of 0.86 vs 0.75), with advantage on tasks with high label imbalance. Other LLMs demonstrated poor performance. Frequent GPT-4 error categories included incorrect inferences from multiple samples and from history, and complex task design, and several LSTM-Att errors were related to poor generalization to the test set.

Discussion: On tasks where large annotated datasets cannot be easily collected, LLMs can reduce the burden of data labeling. However, if the use of LLMs is prohibitive, the use of simpler models with large annotated datasets can provide comparable results.

Conclusions: GPT-4 demonstrated the potential to speed up the execution of clinical NLP studies by reducing the need for large annotated datasets. This may increase the utilization of NLP-based variables and outcomes in clinical studies.

Key words: electronic health records; large language models; breast cancer; pathology; natural language processing.

Introduction

Over the past decade, supervised machine learning methods have been the most popular technique for information extraction from clinical notes.¹ However, supervised learning for clinical text is arduous, requiring curation of large domain-specific datasets, interdisciplinary collaborations to design and execute standardized annotation schema, and significant time from multiple domain experts for the meticulous task of data annotation. Supervised modeling can often require subsequent iterative development driven by advanced technical expertise, which can be limiting for certain practitioners. The entire process thus takes a significant amount of time between problem conception and obtaining final results.

These challenges, combined with the limited availability of clinical notes corpora, have contributed to an under-utilization of Natural Language Processing (NLP) in observational studies from Electronic Health Records (EHRs).²

Recently, language models have demonstrated promising ability for transfer learning, ie, the ability to use knowledge from pre-trained models to improve performance on a related task. This is encouraging for information extraction from clinical text without extensive task-specific model training.^{3–5} Prompt-based querying is popular with generative language models, where practitioners can query the model in natural language to obtain the desired information, sometimes by presenting a few examples of the task they may be trying to

Received: February 7, 2024; Revised: May 27, 2024; Editorial Decision: May 30, 2024; Accepted: June 3, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

solve. Querying large language models (LLMs) like the GPT-4 model have demonstrated varying levels of proficiency in medical inference tasks, such as diagnosing complex clinical cases,^{6–8} answering the United States medical licensing exam questions,^{9,10} radiology report interpretation,^{11,12} clinical notes-based patient phenotyping,^{13–16} automated clinical trial matching,^{17,18} clinical concept extraction,¹⁹ drafting replies to inbox messages,²⁰ recommending treatments,²¹ and improving patient interaction with health systems.^{22,23} However, to understand whether LLMs may be able to perform well in clinical settings without curating large training datasets, few studies have investigated whether zero-shot inference with LLMs can perform as well as task-specific supervised learning in low-resource settings. In this study, utilizing a large corpus of breast cancer pathology notes, we investigate this hypothesis. To this end, we have a 3-fold contribution:

- 1) We developed an annotation schema and detailed guidelines to create an expert-annotated dataset of 769 breast cancer pathology reports with document-level, treatment-relevant information. We further analyzed the curation process to identify frequent modes of disagreements in data annotation, which we additionally present here.
- 2) To establish a baseline of automated breast cancer pathology classification against that of expert clinicians, using the newly curated dataset, we benchmarked the performance of supervised machine learning models of varied levels of complexity, which include a random forest classifier, a long short-term memory network (LSTM) classifier, and a transformers-based BERT classifier trained on UCSF EHR data.
- 3) We finally queried proprietary and open source LLMs to obtain *zero-shot* classification results, ie, results without using any task-specific labeled dataset from UCSF, which we compared to the supervised learning performance obtained earlier. We additionally analyzed the errors made by the best LLM and the best supervised model to understand their limitations.

Materials and methods

Data

Breast cancer pathology reports between January 1, 2012 and March 31, 2021 were retrieved from the University of California, San Francisco (UCSF) clinical data warehouse, deidentified and date-shifted with the Philter algorithm as previously described.²⁴ Access to this de-identified dataset qualifies as non-human subjects research, and no further Institutional Review Board approval was necessary for this study. Patients with breast cancer were identified by querying for encounters with the ICD-9 codes 174, 175, 233.0, or V10.3, or the ICD-10 codes C50, D05, or Z85.3. The cohort was restricted to pathology reports by selecting the note type “Pathology and Cytology.” Notes shorter than 300 characters in length, and those unrelated to breast cancer, for example, those about regular cervical cancer screening through pap smears, were removed through keyword match for “cervix,” “cervical,” and “vaginal.” A flow diagram for the inclusion and exclusion criteria is presented in [Figure 1](#). Among the final set of notes, 769 pathology reports were randomly selected for manual labeling with treatment-relevant breast cancer pathology.

Annotation schema and guidelines were designed in collaboration with oncology experts, who reviewed breast cancer diagnostic and treatment guidelines to determine the most relevant features to infer from pathology reports, along with the categories of these features. To align with the clinical decision-making process, if multiple features of the same category were present, annotators were asked to focus on the one portending poorest prognosis for document level annotations. For example, if there were 2 independent tumors within the report corresponding to grade 2 and grade 3, respectively, the annotator was asked to record grade 3. To analyze categories relevant for prognostic inference, categories such as final tumor margins and lymphovascular invasion were added in addition to commonly investigated categories of biomarkers, histopathology, and grade. The final cohort of 769 breast cancer pathology reports was annotated through 12 key tasks, including 9 single-label tasks and 3 multi-label tasks ([Figure 2](#)). Each report mentioned metadata such as the report date and patient ID, along with the pathologist’s comments and the complete clinical diagnosis. Text spans corresponding to 4 labels (cancer pTNM stage, number of examined and involved lymph nodes, tumor size, and tumor type) were pre-highlighted within text with an internal convolutional neural network (CNN)-based model that had been previously trained for named entity recognition in 5 active learning rounds. The training of this internal tool encompassed approximately 2500 pathology notes across colon, lung, kidney, brain, breast, and prostate cancers based on initial annotations developed earlier.^{25,26} The open-source software LabelStudio²⁷ was used to further add document-level labels. To establish a good inter-annotator agreement, a group of 2 independent oncology fellows annotated the documents jointly in the first phase. After achieving high inter-annotator agreement, the fellows further labeled the documents in the training subset (570 documents) independently. Furthermore, the test subset (100 reports) was established with documents that were annotated by both oncology fellows, and any disagreements between discordant labels were manually adjudicated by a third reviewer. Similarly, the validation subset (99 reports) was annotated in parallel by 3 medical students and any disagreements were independently adjudicated by the same reviewer. The complete annotation guidelines are provided in the [Supplementary Materials, Section S1](#).

Supervised modeling

Supervised machine learning classifiers were trained independently for each of the 12 breast cancer pathology classification tasks on the training subset of 570 pathology notes. Three models of varied complexity were included in the analysis—a random forests classifier,¹⁹ a Long Short Term Memory networks (LSTM) classifier with attention,^{20,21} and a fine-tuned UCSF-BERT (base) model.^{22,23} An analysis of these models accounts for different levels of resources that may be available at different institutions as well as different capabilities of the model architecture itself. While the random forests classifier is a bag-of-ngrams model that does not consider the order of phrases within a document, the LSTM model and the UCSF-BERT model provide a sequential architecture that accounts for word order in input documents. However, while the UCSF-BERT model is powerful due to self-supervised pretraining on clinical data, it is limited in the length of sequences it can process (512 tokens), thereby

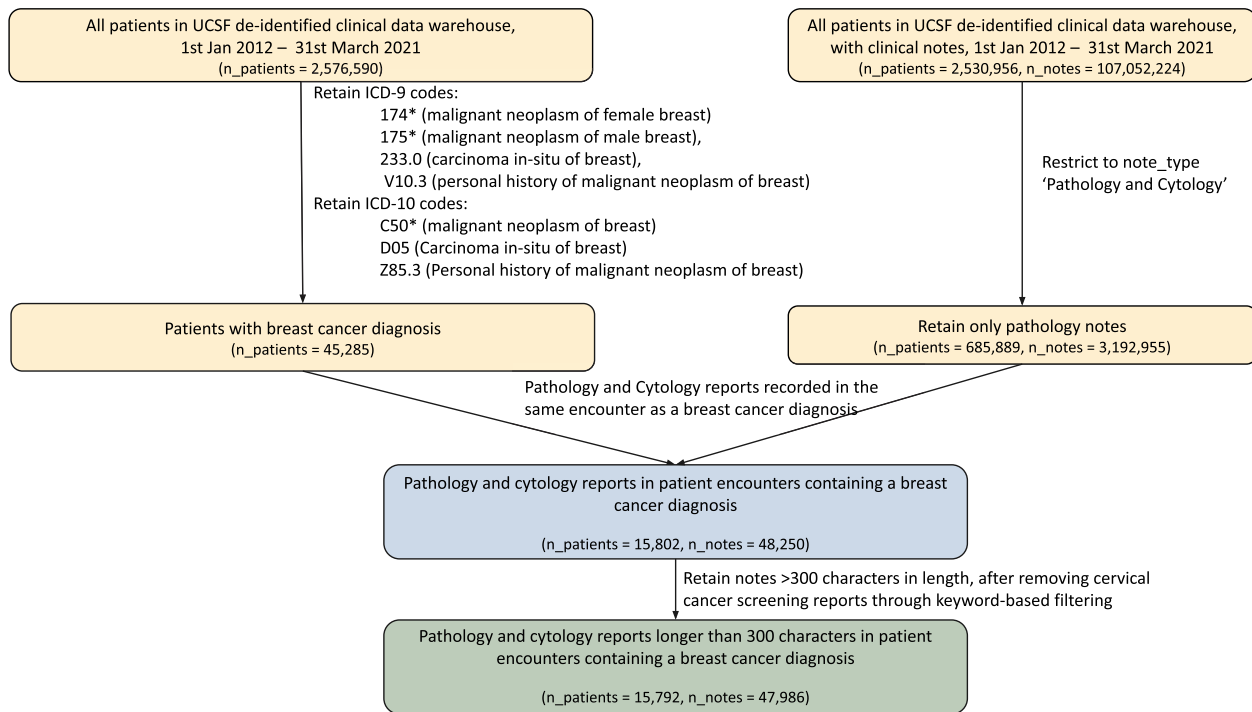


Figure 1. Flow diagram representing inclusion and exclusion criteria for breast cancer pathology report selection before data annotation. Number of patients and number of clinical notes is represented at each stage. The final annotated subset represents a random sample of the final representative dataset obtained in this manner.

The test for **estrogen receptors** is positive. There is variable (ranging from weak to strong) nuclear staining in ~70% of tumor cells. Internal positive control is present.

The test for **progesterone receptors** is positive. There is moderate to strong nuclear staining in ~80% of tumor cells. Internal positive control is present.

Result of ***** test: This carcinoma is negative for ***** oncoprotein over-expression.

An immunohistochemical assay was performed by manual morphometry on block ***** using the ***** monoclonal antibody to ***** oncoprotein. The staining intensity of this carcinoma was 1 on a scale of 0-3. Carcinomas with staining intensity scores of 0 or 1 are considered

FINAL PATHOLOGIC DIAGNOSIS
 ***** lymph node, left axillary, biopsy:
 One lymph node with no tumor identified (0/1). See comment.
 B. Left breast, partial mastectomy:
 1. Invasive ductal carcinoma, ***** grade 2, margins negative for tumor.
 2. Ductal carcinoma in situ, intermediate grade.
 3. Non-proliferative fibrocystic change.
 See comment.
 C. Sentinel lymph node, left axillary, biopsy:
 One lymph node with no tumor identified (0/1).

Sites examined

Left Breast^{nl} Left LN^{nl} Right Breast^{nl} Right LN^{nl} Other tissues^{nl}
 Unknown^{nl}

Sites of disease

Left Breast^{nl} Left LN^{nl} Right Breast^{nl} Right LN^{nl} Other tissues^{nl} None^{nl}
 Unknown^{nl}

Histology

No malignancy^{nl} LCIS^{nl} DCIS Invasive ductal Invasive lobular Medullary
 Mucinous Tubular Papillary Metaplastic BC Mixed Carcinoma NOS
 Unknown

LN involvement

0 involved 1-3 involved 4-9 involved 10+ involved Unknown

Biopsy type

Biopsy Lumpectomy Mastectomy Unknown

ER

Low positive Positive Negative Unknown

PR

Positive Negative Unknown

HER2

Positive Negative Equivocal Equivocal Positive Equivocal Negative
 Unknown

Max grade

0 1 (Low) 2 (Intermediate) 3 (High) Unknown

LVI

Present Absent Unknown

Margins

Positive margin Less than 2mm More than/eq to 2mm Unknown

DCIS Margins

Positive margin Less than 2mm More than/eq to 2mm Unknown

Figure 2. Sample of an annotated pathology report, along with the corresponding document-level annotation schema. The *Unknown* labels refer to the cases where a label could not be inferred based on the information provided in the pathology report.

potentially limiting its performance on document-level tasks unlike the LSTM model.

Model hyperparameters for all supervised models were fine-tuned on the validation set consisting of 99 pathology notes, and the final classification performance was reported on the held-out test set of 100 pathology notes. Details of hyperparameter tuning and the final model settings are available in the [Supplementary Materials, Section S2.1](#). To obtain a reliable estimate of minority class performance, model

performance was evaluated on a held-out test set with the metric of macro-averaged F1-score instead of accuracy.

Random forests classifier

The random forests model was initialized with a TF-IDF vector of n-grams within pathology notes. Pathology reports were pre-processed to remove punctuations and symbols and were converted to lowercase before vectorization. For single-label tasks, training data samples of the minority classes were

up-sampled to reflect a uniform distribution and address data imbalance. Validation and test data were not modified and reflected the real-world distribution. To find the best parameters, a random grid search was performed, using 3-fold cross-validation on the training data and 15 iterations.

LSTM networks

The word embeddings in the LSTM model were initialized with fasttext²⁴ embeddings of 250 dimensions, trained on a corpus of 110 million clinical notes at UCSF. The choice of attention with the LSTM model was additionally made to ensure that long sequences of words can be processed effectively by allowing the model to dynamically focus on different parts of the input. Pathology reports were pre-processed in the similar manner to the use of random forests classifier. To address the data imbalance in multi-label tasks, asymmetric loss²⁵ was used, while the categorical cross-entropy loss was used for single-label tasks.

UCSF-BERT model

The UCSF-BERT model, which was already pretrained from scratch on 75 million clinical notes at UCSF, was fine-tuned further on pathology classification-specific tasks. Similar pre-processing settings as the random forests and the LSTM model were used for both single label and multi-label tasks. Cross-entropy loss was used for all single-label tasks, and asymmetric loss was used for all multi-label tasks to address class imbalance.

Zero-shot inference with LLMs

Proprietary models

Two large language models, the GPT-3.5 model and the GPT-4 model,²⁸ were queried via the HIPAA-compliant Azure OpenAI Studio to provide the requested category of breast cancer pathology information from a given pathology report. (The AI framework to safely use OpenAI application programming interfaces is called Versa at UCSF.) Data were not permanently transferred to or stored by either OpenAI or Microsoft for any purposes. Model inputs were provided in the format `{system role description} {note section text, prompt}`. The specific prompt, model version, and the model hyperparameters are provided in the [Supplementary Materials, section S2.2](#). All classification labels were requested through a single prompt, as one call to the model for each pathology report. Prompt development was performed on the development set, and the final results were reported on a held-out test set. Model outputs were requested in the JSON format, which were post-processed into python dictionaries to automatically evaluate model outputs.

Open source models

We additionally compared 2 open source models, the Starling-7B-beta model²⁹ and the ClinicalCamel-70B model,³⁰ using the same prompts and model settings as GPT-4 and GPT-3.5 models. Prompts were formatted into chat templates specific to the individual models via the HuggingFace transformers library.³¹ We also analyzed the Llama2-7B-chat model but did not obtain any relevant results for this task and chose to exclude it for further comparisons.

Significance testing

Approximate randomized testing³² was used to test for significance between the performance of the best LLM and the best

supervised model. To estimate the *P*-value, model outputs were permuted 100 000 times, counting the number of times that the resulting difference between their macro F1-scores is as or more extreme than that observed with the data. The significance level of 0.01 was chosen to assert significance.

Results

Breast cancer pathology information extraction dataset

769 breast cancer pathology reports were annotated with detailed breast cancer pathology information across 12 key tasks ([Figure 2](#)). Minimum, maximum, mean, and median document length of the dataset were 36, 4430, 723.4, and 560 words, respectively, and the interquartile range was 508 words. The dataset included a population diverse across demographics and age, with nearly 1% of cases being male breast cancer, which reflects the relative incidence of this disease ([Table 1](#)). Median patient age was 55 years. To encourage reproducibility and further research, upon manuscript publication, the dataset will be freely shared through the controlled-access repository PhysioNet. Average inter-annotator agreement, as quantified with Krippendorff's alpha,³³ was 0.85, which varied across tasks ([Supplementary Materials, Table S3](#)). Classification of *DCIS margins* and the multi-label category of *sites examined* showed the lowest inter-annotator concordance, while *lympho-vascular invasion* and *invasive carcinoma margin status* showed the highest concordance.

Sources of disagreements between annotators in the development and the test sets were analyzed by an independent adjudicator. Common sources of disagreements included differences in inferring the most aggressive (“worst”) sample when multiple samples were analyzed, incorrectly including information from patient history into labels for the current report, linguistic or clinical ambiguity in the pathology

Table 1. Socio-demographic distribution of patients in the annotated dataset.

Sample characteristic	Count (percentage) (<i>n</i> = 769)	
Gender		
Male	7	(0.91%)
Female	762	(99.09%)
Age		
Median [IQR]	55.0	[19.0]
Race/ethnicity		
White	505	(65.67%)
Asian	101	(13.13%)
Latinx	42	(5.46%)
Black or African-American	36	(4.68%)
Native Hawaiian or Other Pacific Islander	7	(0.91%)
Other	25	(3.25%)
Multi-Race/Ethnicity	15	(1.95%)
Unknown/Declined	38	(4.94%)
Language		
English	702	(91.29%)
Russian	18	(2.34%)
Unknown/Declined	17	(2.21%)
Chinese—Cantonese	9	(1.17%)
Spanish	9	(1.17%)
Vietnamese	4	(0.52%)
Other	10	(1.30%)

IQR, interquartile range.

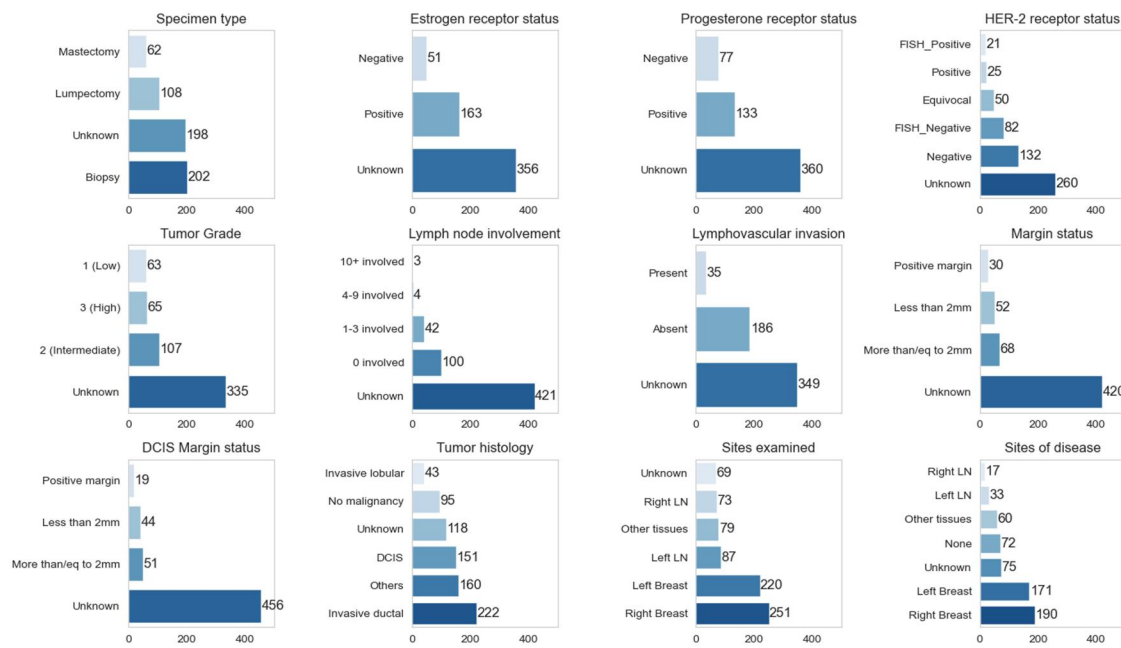


Figure 3. Class distribution for all tasks in the training data for supervised classification.

report, discordant interpretation of procedures involving excisions when differentiating between a histopathology report and a cytology report, inconsistencies in categorizing metastatic disease sites as “other tissues” and histology as “others,” and inconsistent execution of the annotation guidelines for annotating molecular pathology reports and grade information.

The class distribution across the annotated data was highly skewed, resulting in a highly imbalanced dataset. Certain low frequency categories such as groups of histology codes or the *low positive* and *positive* categories of Estrogen Receptor status were combined before further automated classification, resulting in the final distribution presented in Figure 3. The *Unknown* class, which corresponded to the case where the requested information could not be inferred from the given note, was the majority class across 8 of 12 tasks. Among the remaining classes, high imbalances were observed in tasks of inferring the category of the number of lymph nodes involved, lymphovascular invasion, tumor margins, and HER-2 receptor status.

Comparison of model performance

Despite no task-specific training, the GPT-4 model either outperformed or performed as well as our task-specific supervised models trained on task-specific breast cancer pathology data (Figure 4 and Table S4, Supplementary Materials). For both the GPT-4 model and the GPT-3.5 model, all model responses were automatically parsed as JSON without any errors. However, 2 responses of the Starling model and 24 responses of the ClinicalCamel model could not be parsed automatically and were considered “Unknown” for evaluation. The average macro F1 score of the GPT-4 model across all tasks was 0.86, of the LSTM model with attention was 0.74, of the random forests model was 0.59, of the UCSF-BERT model was 0.56, of the GPT-3.5-turbo model (zero-shot) was 0.55, of the ClinicalCamel-70B was 0.34, and that of the Starling model was 0.36. The GPT-4 model was significantly better than the LSTM model (the best supervised

model) for the task of margin status ($P < .01$). This task encompassed a large training data imbalance resulting in a sparsity of class-specific training instances. For all other tasks, no significant differences were obtained between the zero-shot GPT-4 model and the supervised LSTM model.

The GPT-3.5-turbo model and the open source LLMs performed significantly worse than the GPT-4 model for all tasks. Similarly, the UCSF-BERT model, which is a transformer model pre-trained on the corpus of UCSF clinical notes,³⁴ did not outperform the simpler LSTM-Att model for several tasks, although it did match the performance of the GPT-3.5-turbo model. The random forests classifier performed well on keyword-oriented tasks, like pathology type classification and biomarker status classification, but underperformed on tasks requiring more advanced reasoning, like grade and margins inference. Oversampling training data for mitigating label imbalance in single-label classification tasks demonstrated mixed benefit across tasks and models (Figure S1a), although the choice of asymmetric loss showed consistent improvements compared to the use of binary cross-entropy loss (Figure S1b).

Error analysis

The confusion matrix of the GPT-4 model revealed that it had difficulties in differentiating the unknown class from the class that indicated no lymph node involvement and no lympho-vascular invasion (Supplementary Materials, Figure S2). Furthermore, margin status inference was challenging for the model, where *more than 2 mm margins* (negative margins) were confused with *less than 2 mm margins*. Confusion between classes was more prevalent in multi-label tasks than single-label tasks. Further errors from the GPT-4 model were prevalent when the task design was ambiguous in model prompts, such as the grouping of sparse histology into an “others” category, the assignment of metastatic sites for breast cancer as “other tissues than breast or lymph nodes,” or the inference of pathology reports unrelated to breast

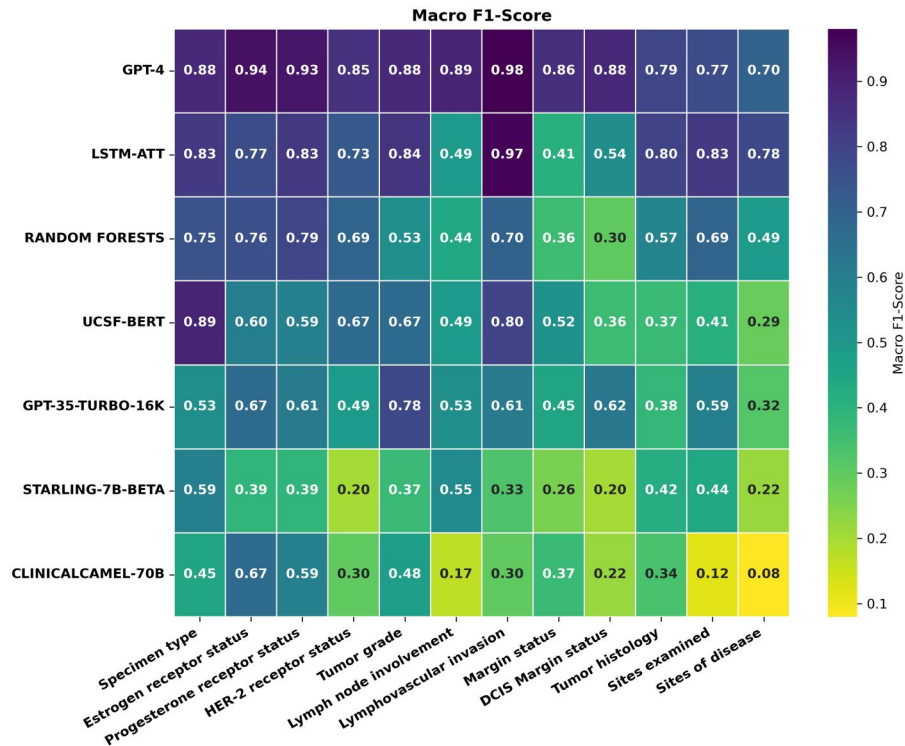


Figure 4. Classification performance, as measured by % Macro F1 score, for different models for each classification task. The LSTM model, the UCSF-BERT model, and the Random Forests model were trained in a supervised setup on task-specific training data. All other models (GPT-3.5, GPT-4, Starling-7B-beta, and ClinicalCamel-70B) were queried and evaluated in a zero-shot setup, ie, without any further task-specific training.

cancer. The latter set of errors correspond to common sources of disagreements identified during the data annotation process.

Manual analysis of the GPT-4 model errors revealed several consistent sources of errors, described in Table 2. Common sources of errors in biomarker reporting included the reporting of results from clinical history or tests conducted at other clinical sites that were not confirmed in the current report. Furthermore, the GPT-4 model incorrectly reported nuclear grade as the overall tumor grade when the overall grade was not discussed in the note. Moreover, common errors in reporting tumor margins were concerned with mathematical inferences over multiple margin thicknesses (for example, anterior, posterior, medial, etc), where the value representing the thinnest margin was to be provided. Manual analysis additionally uncovered several error sources for multi-label tasks. The model performed inconsistently when inferring sites of benign findings; while the model frequently missed reporting the site of benign findings as a site examined for tumors, it also sometimes included sites of benign findings as a site of cancer. Furthermore, sentinel and axillary lymph nodes were frequently reported as tissues other than breast or lymph nodes, although they were annotated as lymph node sites. Some errors related to complex cases were also found, for example, 1% staining results for progesterone receptors were provided as negative by the model, whereas they were annotated as positive. Finally, errors related to task setup were reflected in histology-related errors, where the model could not reliably abstain from providing histology from reports unrelated to breast cancer and from molecular pathology reports for ERBB2 despite being instructed as such, and errors due to the grouping of histologies like LCIS into an “others” category.

Manual error analysis of the LSTM-Att model, which was the best supervised classification model, revealed a few similarities with errors from the GPT-4 model (nuclear grade provided instead of total grade, numerical inference error from multiple margins samples, difficulty in learning the “others” category for tumor histology, difficulty in answering from reports unrelated to breast cancer or from molecular pathology reports for ERBB2). However, additional errors due to poor model generalization were identified, where linguistic variability compared to the training set resulted in incorrect responses (Table 2). Several errors occurred even when correct answers were directly mentioned within text, which could potentially indicate that the model did not use broader context for providing outputs, but rather overfit on specific keyword patterns in the training set. Finally, independent training of classifiers for multi-label classification posed challenges in learning interaction between multiple labels (Table 2).

Discussion

Task-specific supervised learning models trained on manually annotated data have been the standard approach in clinical NLP for over a decade.¹ Using a manually annotated dataset of 769 breast cancer pathology reports focused on the most clinically relevant report features, our study compared the performance of supervised learning models, including random forests classifier, LSTM models, and the UCSF-BERT model, with a zero-shot classification performance of 2 proprietary LLMs, the GPT-4 model, and the GPT-3.5-turbo model, and 2 open-source LLMs, the Starling-7B-beta model, and the ClinicalCamel-70B model. We found that even in zero-shot setups, the GPT-4 model performs as well as or

Table 2. Common categories of errors for the GPT-4 model and the LSTM model along with corresponding examples.

Task	Error category	Example
GPT-4 model Biomarkers: ER, PR	Reporting from history	<i>Clinical Diagnosis/History:</i> ... This cancer was 6 cm in greatest dimension by imaging and was ER, PR, and *****/neu positive. <u>GPT-4 output:</u> Positive <u>Annotation:</u> Unknown
	Complex case	<i>The cells stain strongly for the estrogen receptor (>95% at 3+ staining; on a scale of 0–3+) and only rare cells stained for the progesterone receptor (~1% at 3+ staining; on a scale of 0–3+).</i> <u>GPT-4 output:</u> Negative <u>Annotation:</u> Positive
Biomarker: HER2	Incorrect inference due to redaction	<i>This carcinoma is positive for ***** oncoprotein over-expression. The staining intensity of this carcinoma is 3+ on a scale of 0–3.</i> <u>GPT-4 output:</u> Unknown <u>Annotation:</u> Positive
Tumor grade	Nuclear grade reported instead of final total grade	- Invasive tumor grade (modified *****.****): - Nuclear grade: 2, 2 points. - Mitotic count: <10 mitotic figures/10 HPF, 1 point. - Tubule/papilla formation: >75%, 1 point. - Total grade/points: 1. <u>GPT-4 output:</u> 2 (Intermediate) <u>Annotation:</u> 1 (Low)
	Grade reported from history	<i>Clinical history:</i> <i>The patient underwent needle core biopsy of a left breast mass at an outside institution 08/27/2013, which revealed infiltrating carcinoma interpreted as high-grade ductal carcinoma.</i> <u>GPT-4 output:</u> 3 (High) <u>Annotation:</u> Unknown
Margin status	Error in numerical inference from multiple margins	- Margins for invasive tumor: Negative. - Deep margin: Negative; (tumor is 0.1 cm away, on slide B15). - Medial margin: Negative; (tumor is >1 cm away). - Lateral margin: Negative; (tumor is >1 cm away). - Anterior/superior margin: Negative; (tumor is 0.5 cm away, on slide B11). - Anterior/inferior margin: Negative; (tumor is 0.8 cm away, on slide B15). <u>GPT-4 output:</u> More than or equal to 2 mm <u>Annotation:</u> Less than 2 mm
	Complex case: Margins before final resection reported	<i>Resection margins for invasive tumor: The initial lumpectomy (****_.****_.****) demonstrated extension of tumor to the green-inked margin. No residual tumor is identified in the select slides submitted for review from the left modified radical mastectomy (****_.****_.****).</i> <u>GPT-4 output:</u> Positive <u>Annotation:</u> More than or equal to 2 mm
DCIS margin status	Error in numerical inference from multiple margins	<i>Resection margins for DCIS:</i> <i>Posterior margin: Negative (tumor is 0.4 cm away, on slide A8-1).</i> <i>Anterior nipple/areolar base: A cauterized duct suspicious for DCIS is present at the inked margin, although cauterization artifact precludes definitive evaluation. Evaluable DCIS is present immediately adjacent to this cauterized focus, <0.1 cm from the margin (on slide A3-1).</i> <i>Medial margin: Negative (by report, tumor is >1 cm away).</i> <i>Lateral margin: Negative (by report, tumor is >1 cm away).</i> <i>***** margin: Negative (by report, tumor is >1 cm away).</i> <i>Inferior margin: Negative (by report, tumor is >1 cm away).</i> <u>GPT-4 output:</u> Less than 2 mm <u>Annotation:</u> Positive
Lymph node involvement	Different reporting for benign findings from sites that did not include lymph nodes	<i>FINAL CYTOLOGIC DIAGNOSIS:</i> <i>Soft tissue, right upper chest, fine needle aspiration: Benign fibroadipose tissue and skeletal muscle; see comment.</i> <u>GPT-4 output:</u> 0 involved <u>Annotation:</u> Unknown

(continued)

Table 2. (continued)

Task	Error category	Example
Lympho-vascular invasion	No invasive carcinoma was identified but lympho-vascular invasion not mentioned explicitly	<p>COMMENTS: <i>The upper outer quadrant and the area deep to the nipple were extensive sampled. No invasive carcinoma or ductal carcinoma in situ was identified.</i></p> <p><u>GPT-4 output:</u> Absent <u>Annotation:</u> Unknown</p>
Sites examined	All sites not reported, particularly when no tumors are found	<p><i>Final Diagnosis:</i> A. <i>Lymph node, right axilla sentinel #1, biopsy: No tumor in one lymph node (0/1).</i> B. <i>Right breast, total skin-sparing mastectomy:</i> 1. <i>No invasive or in situ carcinoma identified; see comment.</i> 2. <i>Hematoma with adjacent surgical site changes.</i> 3. <i>Nonproliferative fibrocystic change (stromal fibrosis, microcysts and apocrine metaplasia).</i> C. <i>Right breast, nipple, biopsy: No tumor.</i> D. <i>Skin, left chest/breast, biopsy: No tumor.</i></p> <p><u>GPT-4 output:</u> Right LN, Right Breast <u>Annotation:</u> Left Breast, Right LN, Right Breast</p>
Sites of disease	Incorrect reporting of benign finding as a site of disease	<p><i>DIAGNOSIS:</i> Cerebrospinal Fluid BENIGN.</p> <p><i>CLINICAL DATA:</i> 73-year-old female with history of breast cancer, now with bone metastasis and focus of leptomeningeal metastasis.</p> <p><u>GPT-4 output:</u> Other tissues <u>Annotation:</u> None</p>
Tumor histology	Other histology is not reported	<p><i>FINAL PATHOLOGIC DIAGNOSIS</i> ... A. <i>Left breast (slides A1, A3, A8-10 only), modified radical mastectomy:</i> 1. <i>Invasive ductal carcinoma with focal micropapillary features, multifocal by report, largest focus 2.2 cm, ***** grade 2, margins negative; see comment.</i> 2. <i>Ductal carcinoma in situ, intermediate and low nuclear grades, cribriform and solid patterns with necrosis, cauterized duct at nipple base margin; see comment.</i> 3. <i>Atypical ductal hyperplasia.</i> 4. <i>Flat epithelial atypia.</i> 5. <i>Usual ductal hyperplasia, apocrine metaplasia and dilated ducts.</i> 6. <i>Detached calcifications.</i> B. <i>Left axillary lymph nodes, dissection: Metastatic carcinoma in 5 of 19 lymph nodes (July 03), largest focus 0.5 cm; see comment.</i> C. <i>Left axillary sentinel lymph nodes, biopsy: Metastatic carcinoma in 1 of 2 lymph nodes (February 15), largest focus 1.1 cm, with extranodal extension; see comment.</i> D. <i>Right breast, total skin-sparing mastectomy: Benign breast tissue with cystic dilatation of ducts.</i> E. <i>Right breast, new inferior margin, excision: Benign fibroadipose tissue with no significant pathologic abnormality.</i></p> <p><u>GPT-4 output:</u> Invasive ductal, DCIS, Others <u>Annotation:</u> Invasive ductal, DCIS</p>
Tumor histology	Task setup-related error: Molecular pathology reports	<p><i>Molecular Diagnostics Report</i> ... <i>Formalin-fixed, paraffin-embedded tissue on glass slides.</i> ... <i>Gastric: Adenocarcinoma.</i></p> <p><u>GPT-4 output:</u> Other tissues <u>Annotation:</u> Unknown</p>
LSTM model Biomarkers: ER, PR, HER2	Insufficient generalization to test set	<p>These demonstrate that the cells are negative for *****/6 (on blocks ...), diffusely positive for ER (on blocks ...), and positive for synaptophysin and chromogranin.</p> <p>LSTM output: ER Positive Annotation: ER Unknown</p> <p>Estrogen and progesterone receptor immunoperoxidase studies are performed on block *****. Strong nuclear staining (3+/3+) for both ER and PR is present in 90% of invasive tumor cells.</p> <p>LSTM output: PR Positive Annotation: PR Negative</p>

(continued)

Table 2. (continued)

Task	Error category	Example
		<ul style="list-style-type: none"> - Estrogen receptor: Positive in most tumor cells. - Progesterone receptor: Positive in most tumor cells. - *****/neu: Borderline (immunohistochemical staining score of 2+) LSTM output: HER-2 positive Annotation: HER-2 Equivocal
Biomarker: HER-2	Results before FISH testing provided by the model when results both before and after FISH testing are discussed	Result of **** test: This carcinoma is equivocal for **** oncoprotein over-expression. The staining intensity of this carcinoma was 2+ on a scale of 0–3. By report, **** FISH was performed and was negative for **** LSTM output: Equivocal Annotation: FISH Negative
Tumor grade	Nuclear grade reported instead of final total grade	Invasive tumor grade (modified ****_****): <ul style="list-style-type: none"> - Nuclear grade: 2, 2 points. ... - Total grade/points: 1. LSTM output: 2 (Intermediate) Annotation: 1 (Low)
Margin status	Insufficient generalization to test set	Carcinoma is located 0.6 cm from the anterior margin (slide ...) and greater than 1 cm from all other margins. LSTM output: Unknown Annotation: More than/eq to 2 mm
	Inference error from multiple margins	Margins for invasive tumor: <ul style="list-style-type: none"> - Posterior margin: Negative (tumor is <0.1 cm away, on slide ...) - Medial margin: Negative (tumor is >0.5 cm away). - Lateral margin: **** but negative (tumor is <0.05 cm away, on slide D28). - Anterior/superior margin: Positive, focal (tumor is focally on ink at margin, on slide D22 keratin immunostain). - Anterior/inferior margin: **** but negative (tumor is <0.05 cm away, on slides D17 and D21). LSTM output: More than/eq to 2 mm Annotation: Positive
DCIS margin status	Insufficient generalization to test set	The immunostains support the presence of cauterized DCIS at ****-inked inferior margin (slide 2E) and the ****-inked anterior margin (slide 2K). LSTM output: Unknown Annotation: Positive
	Inference error	Status of resection margins for ductal carcinoma in situ: In main lumpectomy specimen: DCIS within much less than 0.01 cm in multiple foci; areas of cauterized tissue suspicious for DCIS present at margin LSTM output: Less than 2 mm Annotation: Positive
Lymph node involvement	Insufficient generalization to test set	<ul style="list-style-type: none"> - Lymph node status: - Number of positive lymph nodes: 18. LSTM output: 1-3 involved Annotation: 10+ involved No tumor in 10 lymph nodes (0/10) LSTM output: Unknown Annotation: 0 involved
Lympho-vascular invasion	Insufficient generalization to test set	Lymphovascular space invasion: No definite invasion. LSTM output: Unknown Annotation: Absent
Sites examined	Insufficient generalization to test set	A. Left breast, needle core biopsy ... Invasive carcinoma consistent with breast primary, infiltrating fibroadipose tissue; see comment. B. Right adrenal gland, needle core biopsy ... Large cell-rich B-cell lymphoma; see comment. C. Right adrenal gland, needle core biopsy ... Lymphoid tissue consistent with large cell-rich B-cell LSTM output: Left Breast Annotation: Left Breast, Other tissues

(continued)

Table 2. (continued)

Task	Error category	Example
	Ignoring context; overfitting on keywords	Brain, left cerebellum, resection: Metastatic adenocarcinoma; see comment. LSTM output: Left Breast Annotation: Other tissues
Sites of disease	Ignoring context to provide all sites regardless of whether tumor was identified, potentially worsened due to independent multi-label training	DIAGNOSIS: A. Sentinel lymph node, left axilla, excision: One lymph node with no tumor identified (0/1); see comment. B. Sentinel lymph node #2, left axilla, excision: One lymph node with no tumor identified (0/1); see comment. C. Sentinel lymph node #3, left axilla, excision: One lymph node with no tumor identified (0/1); see comment. D. Left breast, surgical scar, excision: Scar with foreign body reaction to suture material; no tumor identified. E. Left breast, partial mastectomy: Scar with foreign body reaction to suture material; no tumor identified. LSTM output: Other tissues, Left LN, None Annotation: None
Tumor histology	Multi-sample inference error due to independent multi-label training	Final Diagnosis: A. Sentinel lymph node, biopsy: No evidence of carcinoma in one lymph node (0/1). B. Breast, right, partial mastectomy: 1. Invasive lobular carcinoma with associated lobular carcinoma in situ, 1.1 cm, ***** grade 2, margins negative for tumor; see comment. 2. Atypical lobular hyperplasia. 3. Stromal fibrosis and apocrine metaplasia. C. Breast, right, anterior superior margin, excision: Atypical lobular hyperplasia. D. Breast, right, fascia and muscle deep margin, excision: Fibroadipose tissue and skeletal muscle, no evidence of carcinoma. E. Breast, right, new inferior lateral margin, excision: Cyst formation, stromal fibrosis, and apocrine metaplasia. F. Lymph node, biopsy: No evidence of carcinoma in one lymph node (0/1). LSTM output: Others, Invasive lobular, No malignancy Annotation: Others, Invasive lobular

significantly better than simpler, task-specific supervised counterparts on all classification tasks, although other LLMs performed significantly worse. Previous studies have demonstrated similar results, showing that in zero-shot setups, LLMs consistently perform the same as or outperform fine-tuned models on biomedical NLP datasets with small training data sizes (fewer than 1000 training examples).^{35,36} Similar small datasets are common in medical informatics studies since domain expertise is frequently required for reliably annotating clinical notes, making the process time-consuming and difficult to scale.³⁷ This study enhances previous findings on a new real-world clinical dataset, reinforcing that the GPT-4 model is promising for use in classification tasks in low-resource clinical settings. Open source models need further developments before they are conducive for use in similar settings.

Tasks where the training data contained high label imbalance were particularly conducive for using GPT-4 model over task-specific supervised models, including compared to pre-trained models like the UCSF-BERT model. Given that the GPT-4 model is already trained on internet-scale corpora and the specifics of model training are not available publicly, the model may already encode a fundamental understanding of breast cancer pathology, which may explain its surprising zero-shot capability on these tasks, including that on complex and imbalanced tasks like margins inference. However, the reasons behind the striking performance difference between the GPT-3.5 and GPT-4 models remain unclear due to the

closed nature of these models, although similar trends have been observed in previous medical NLP studies.^{38,39,15} However, if access to models like the GPT-4 model is prohibitive due to either privacy or computational constraints, comparable performance on EHR-based NLP tasks such as pathology classification can be obtained with simpler deep learning classifiers trained on task-specific datasets, particularly if annotated sample sizes are sufficiently large and class imbalance can be controlled through targeted annotations of minority classes for model training.

An analysis of the GPT-4 model errors indicated several errors due to insufficient understanding of idiosyncratic task-design choices, for example differentiating between “Unknown” and “no lymph node involvement” categories. When histopathological samples were complex and could not be characterized entirely within one of the pre-defined histologic categories, the GPT-4 model still provided the imperfect option rather than using the “Other” category for ambiguous cases. This demonstrates how models may be susceptible to information loss due to the artificial nature of many medical classifications schemas that, in reality, exist on a continuum. It is possible that these errors can be mitigated with strategies, such as few-shot learning to demonstrate a better understanding of annotation-specific choices, or chain-of-thought-prompting to elucidate reasoning and avoid answering from incomplete or old information within text report. However, it has been demonstrated earlier that the GPT-4 model cannot process long input contexts efficiently,⁴⁰ and we leave this

question for future research. Finally, analysis of the LSTM model errors identified those stemming from training the model on an insufficiently diverse dataset, for example, incorrect responses when results were discussed in free-form patterns rather than following the structure of a standard breast cancer pathology note. Although these findings are not surprising, they highlight the challenge of supervised model generalization in low-resource settings, which may add to clinical deployment-related challenges.

Although we found promising performance of the GPT-4 model compared to task-specific supervised models, several design choices may have impacted the findings. The dataset was curated from a single health system, and further validation of the findings on pathology reports from other health systems may improve the reliability of the results. Although potential de-identification errors may have impacted the capability of LLMs, the data reflects real-world setups for retrospective observational studies in a privacy-preserving manner. Furthermore, although it may be possible to further improve model performance with more hyperparameter and prompt tuning, the findings of this study will inform future studies on the development of more advanced prompting and few-shot strategies for LLMs to obtain even better performance, the development of effective annotated datasets for simpler supervised classification setups, the evaluation of newer LLMs for clinical information extraction, and the analysis of output sensitivity to input prompts and model settings. Moreover, the studied classifiers may exhibit biases against specific demographics, and caution must be exercised when deploying them in clinical workflows. These biases need to be investigated further in the future to establish concrete guidelines for their use. Finally, we note that the label *Unknown* in the dataset covers 2 distinct scenarios: (1) although the information may be present within a patient's EHR record in some form, it is not present or identifiable within the specific pathology note, or (2) the feature is not relevant or cannot be obtained from this context. For instance, while HER-2 status may not be identifiable from a specific note (and thus is *Unknown* at the time of annotation), we would expect that this would be an identifiable piece of information in the patient's EHR record at some point. This subtle difference should be noted in future utilization of this dataset.

Pathology reports represent a foundational source of clinical information, both for diagnosis and medical decision making and for cohort development for research. Given the importance of this information for clinical oncology, along with the challenges and time required for accurate interpretation, the ability to accurately extract salient features from pathology reports could improve physician workflow as well as facilitate cohort development for large scale research analyses. Accurate zero-shot methods for inferring treatment-relevant pathology will enable swift identification and categorization of complex pathology features, thus holding the potential to expedite the development of research cohorts, enabling rapid hypothesis testing for retrospective research and quicker clinical trial enrolments within the clinic. For instance, these methods can be utilized to quickly screen large volumes of pathology reports for identifying similar patients suffering from a rare cancer subtype to facilitate personalized treatments and tumor board discussions for new patients, while also freeing up specialists to focus on more nuanced clinical decision-making tasks. This utility depends on the accuracy of automatic extraction of relevant information

from notes, as any errors in this workflow may cause further harm to patients in clinical settings. Based on our analyses, zero-shot inference with the GPT-4 model shows strong promise for cohort identification and labeling for clinical research, but may not yet be sufficiently robust for direct integration into clinical workflows.

Despite widespread studies in oncology information extraction from textual clinical records,^{41,42} annotated datasets of breast cancer pathology reports are not publicly available. To make the findings of this study replicable and promote further research on breast cancer pathology extraction, the dataset curated in this study along with corresponding source code will be shared publicly through a controlled-access repository PhysioNet, accessible via a data use agreement.

Conclusions

The study compared breast cancer pathology classification abilities of 7 models of varying sizes and architecture, finding that the GPT-4 model, even in zero-shot setups requiring no further model training, performed similarly to or better than the LSTM model with attention trained on nearly 570 pathology report examples. The GPT-4 model outperformed simpler baselines for classification tasks with high label imbalance. However, when large training datasets were available, no significant differences were observed between the performance of simpler models like the LSTM model with attention compared to the GPT-4 model. The results of this study demonstrated that while LLMs may relieve the need for resource-intensive data annotations for creating large training datasets in medicine, if there are privacy, computational, or cost-related concerns regarding the use of LLMs with patient data, it may be possible to obtain reliable performance with simpler models by developing large annotated datasets, with particular focus on minority class labeling potentially in an active-learning setup.

Acknowledgments

This research would not have been possible without support from several people. The authors thank the 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, the UCSF Information Commons team, and the Wynton high-performance computing platform team at UCSF for supporting high-performance computing platforms that enable the use of language models with de-identified patient data. We further thank Debajoyti Datta and Michelle Turski for feedback on breast cancer pathology annotation schema, all members of the Butte lab for helpful discussions in the internal presentations, and Gundolf Schenk and Lakshmi Radhakrishnan for discussions related to clinical note de-identification.

Author contributions

Madhumita Sushil ideated and led the study, formed the study team, set up the technical pipelines, developed and

executed data extraction, modeling, and inference pipelines, advised on supervised model development, and oversaw the data annotation process, and analyzed model errors. MS additionally wrote the manuscript and incorporated co-author feedback.

Travis Zack and Divneet Mandair led the data extraction and annotation processes, developing the annotation schema and guidelines, annotated pathology reports, and analyzed model errors. They led the clinical aspects of the study and additionally provided critical feedback to the manuscript.

Zhiwei Zheng, Ahmed Wali, and Yuwei Quan developed supervised classification pipelines and additionally provided critical feedback to the manuscript.

Yan-Ning Yu adjudicated annotation disagreements, reviewed the dataset for potential errors, clarified annotation guidelines, and analyzed errors in model outputs and additionally provided critical feedback to the manuscript.

Dmytro Lituiev developed the internal CNN model used to label tumor type, cancer stage, and lymph node involvement within notes before further annotation. He further provided critical feedback for data annotation and to the manuscript, thereby improving research dissemination.

Atul J. Butte supervised the entire study, setting its critical direction including data extraction, annotation, modeling, and interpretation of results, acquired partial funding for the study, and provided critical feedback to different stages of the manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA082103, the FDA grant U01FD005978 to the UCSF–Stanford Center of Excellence in Regulatory Sciences and Innovation (CERSI), and the NIH UL1 TR001872 grant to UCSF CTSI. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of interests

M.S. reports no financial associations or conflicts of interest. T.Z. is a medical advisor and minor shareholder at OpenEvidence.com. D.M. is a consultant to Third Rock Ventures. Z. Z. reports no financial associations or conflicts of interest. A. W. is currently an employee of Abbott. Y.-N.Y. is currently an employee of City of Hope. Y.Q. is currently an employee of X-camp Academy. D.L. is currently an employee and minor shareholder of Johnson and Johnson and a co-founder and major shareholder of Synthes AI Corp. A.J.B. is a co-founder and consultant to Personalis and NuMedii; consultant to Mango Tree Corporation, and in the recent past, Samsung, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in Personalis and NuMedii; is a minor shareholder in Apple, Meta (Facebook),

Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna Health, Assay Depot, and Vet24seven, and several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. A.J.B. receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. A.J.B.'s research has been funded by NIH, Peraton (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervall Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity. None of these entities had any bearing on this research or interpretation of the findings.

Data availability

The dataset of 769 pathology reports curated in this study will be made available through the controlled-access repository PhysioNet after signing a data use agreement. The source code for the study, including those for supervised modeling and large language model inference, are available through the GitHub repository: <https://github.com/MadhumitaSushil/BreastCaPathClassification>.

References

1. Wu H, Wang M, Wu J, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *NPJ Digit Med.* 2022;5(1):186.
2. 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(3):398-411.
3. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, et al., eds. *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc.; 2020:1877-1901.
4. Kojima T, Gu S, Shane R, Matsuo M, Iwasawa Y. Y. Large language models are zero-shot reasoners. In: Koyejo S, Mohamed S, Agarwal A, et al., eds. *Adv Neural Inform Process Syst.* 2022;35:22199-22213.
5. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing 1998–2022* (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).
6. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI.* 2023;1(1):AIP2300031.
7. Wang Z, Li R, Dong B, et al. 2023. Can LLMs like GPT-4 outperform traditional AI tools in dementia diagnosis? Maybe, but not today. *arXiv:2306.01499*, preprint: not peer reviewed.
8. Barile J, Margolis A, Cason G, et al. Diagnostic accuracy of a large language model in pediatric case studies. *JAMA Pediatr.* 2024;178(3):313-315. <https://doi.org/10.1001/jamapediatrics.2023.5750>

9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv*, arXiv:2303.13375, preprint: not peer reviewed.
10. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023;13(1):16492.
11. Liu Q, Hyland S, Bannur S, et al. Exploring the boundaries of GPT-4 in radiology. In: Bouamor H, Pino J and Bali K, eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics; 2023:14414-14445. <https://doi.org/10.18653/v1/2023.emnlp-main.891>
12. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology*. 2023;308(3):e231362.
13. Alsentzer E, Rasmussen MJ, Fontoura R, et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *NPJ Digit Med*. 2023;6(1):1-10.
14. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med*. 2024;7(1):1-14.
15. Sushil M, Kennedy VE, Mandair D, et al. CORAL: expert-curated oncology reports to advance language model inference. *NEJM AI*. 2024;1(4):AIdbp2300110. <https://doi.org/10.1056/AIdbp2300110>
16. Truhn D, Loeffler CM, Müller-Franzes G, et al. Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). *J Pathol*. 2024;262(3):310-319.
17. Wong C, Zhang S, Gu Y, et al. Scaling clinical trial matching using large language models: a case study in oncology. In: *Proceedings of the 8th Machine Learning for Healthcare Conference*. PMLR; 2023:846-862.
18. Datta S, Lee K, Paek H, et al. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *J Am Med Inform Assoc*. 2024;31(2):375-385.
19. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. 2024:ocad259. <https://doi.org/10.1093/jamia/ocad259>
20. Garcia P, Ma SP, Shah S, et al. Artificial intelligence—generated draft replies to patient inbox messages. *JAMA Netw Open*. 2024;7(3):e243201.
21. Iqbal U, Lee LT-J, Rahmanti AR, Celi LA, Li Y-CJ. Can large language models provide secondary reliable opinion on treatment options for dermatological diseases? *J Am Med Inform Assoc*. 2024;31(6):1341-1347. <https://doi.org/10.1093/jamia/ocae067>
22. Mirza FN, Tang OY, Connolly ID, et al. Using ChatGPT to facilitate truly informed medical consent. *NEJM AI*. 2024;1(2):Alcs2300145.
23. Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Netw Open*. 2024;7(3):e240357.
24. 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(3):ooad045.
25. Odisho AY, Park B, Altieri N, et al. Natural language processing systems for pathology parsing in limited data environments with uncertainty estimation. *JAMIA Open*. 2020;3(3):431-438.
26. Trivedi HM, Panahiazar M, Liang A, et al. Large scale semi-automated labeling of routine free-text clinical records for deep learning. *J Digit Imaging*. 2019;32(1):30-37.
27. HumanSignal/label-studio. 2024. Accessed June 11, 2024. <https://labelstud.io/>
28. OpenAI. 2023. GPT-4 technical report. *arXiv*, arXiv:2303.08774, preprint: not peer reviewed.
29. Hugging Face. 2024. Nexusflow/Starling-LM-7B-beta. Accessed June 11, 2024. <https://huggingface.co/Nexusflow/Starling-LM-7B-beta>
30. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. 2023. Clinical Camel: an open expert-level medical language model with dialogue-based knowledge encoding. *arXiv*, arXiv:2305.12031, preprint: not peer reviewed.
31. Wolf T, Debut L, Sanh V, et al. 2020. HuggingFace's transformers: state-of-the-art natural language processing. In: Liu Q, Schlangen D, eds. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics; 2020:38-45. <https://aclanthology.org/2020.emnlp-demos.6>
32. Edgington ES. Approximate randomization tests. *J Psychol*. 1969;72(2):143-149.
33. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications; 2018.
34. Sushil M, Ludwig D, Butte AJ, Rudrapatna VA. 2022. Developing a general-purpose clinical language inference model from a large corpus of clinical notes. *arXiv*, arXiv:2210.06566, preprint: not peer reviewed.
35. Jahan I, Laskar MTR, Peng C, Huang JX. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Comput Biol Med*. 2024;171(1527-974X):108189.
36. Chen Q, Du J, Hu Y, et al. 2024. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv*, arXiv:2305.16326, preprint: not peer reviewed.
37. Gao Y, Dligach D, Christensen L, et al. A scoping review of publicly available language tasks in clinical natural language processing. *J Am Med Inform Assoc*. 2022;29(10):1797-1806.
38. Taloni A, Borselli M, Scarsi V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci Rep*. 2023;13(1):18562.
39. Nori H, Lee YT, Zhang S, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv*, arXiv:2311.16452, preprint: not peer reviewed.
40. Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. *Trans Assoc Computat Linguist*. 2024;12(2307-387X):157-173.
41. Wang L, Fu S, Wen A, et al. Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing. *JCO Clin Cancer Inform*. 2022;6:e2200006. <https://doi.org/10.1200/CCI.22.00006>
42. Gholipour M, Khajouei R, Amiri P, Hajesmaeel Gohari S, Ahmadian L. Extracting cancer concepts from clinical notes using natural language processing: a systematic review. *BMC Bioinformatics*. 2023;24(1):405.