From Zero-Shot to Bedside: A Practical Playbook for Adapting Open-Source Large Language Models to Clinical Symptom Extraction

Li-Ching Chen[1,2,3] LICHING_CHEN@BERKELEY.EDU
Travis Zack[2,3,4] TRAVIS.ZACK@UCSF.EDU
Divneet Mandair[2] DIVNEET.MANDAIR2@UCSF.EDU
Aditya Mahadevan[2] ADITYA.MAHADEVAN@UCSF.EDU
Arvind Suresh[2] ARVIND.SURESH@UCSF.EDU
Yuta Ishiyama[2] YUTA.ISHIYAMA@UCSF.EDU
Yiping Li[4] LIYIPINGPURPLE@GMAIL.COM
Julian C. Hong[2,3] JULIAN.HONG@UCSF.EDU
Autl J. Butte[2] ATUL.BUTTE@UCSF.EDU
[1] *UC Berkeley* [2] *UCSF* [3] *Weill Cancer Hub West* [4] *OpenEvidence*

Abstract

Large language models (LLMs) are increasingly applied to clinical notes, but guidance on how to adapt open-source models to specific tasks and manage annotation quality at scale is limited. We present a playbook for fine-tuning LLMs on de-identified clinical notes from patients with pancreatic cancer, spanning both pre-diagnosis and on-treatment settings. We evaluate prompting strategies, contrast open-source models with GPT-4o, and explore disease-level versus task-specific adaptation. A key contribution is an LLM-assisted adjudication workflow in which models flag notes where predictions consistently conflict with initial human labels. This approach concentrated expert review on a small fraction of cases while identifying many true annotation errors, ultimately improving downstream model performance. We further examine the use of machine-generated annotations to augment limited expert labels, showing that balanced mixtures of synthetic and human data can enhance fine-tuned models. Our findings provide practical guidance for deploying open-source LLMs in clinical contexts, offering strategies to improve accuracy, reduce annotation burden, and enable privacy-preserving, site-adapted clinical natural language processing (NLP).

Keywords: Clinical NLP, Large language models, Domain adaptation, Data adjudication, Data augmentation

Data and Code Availability Electronic health record (EHR) data for this study was extracted from the University of California, San Francisco (UCSF) Deidentified Clinical Data Warehouse, a deidentified relational data repository reflecting UCSF Health's EHR (University of California, San Francisco, Academic Research Systems, 2024). The data is accessible to researchers through PhysioNet and listed with the code available here[1].

Institutional Review Board (IRB) Since the UCSF clinical database comprises wholly de-identified medical records and no contact with human subjects was involved, the study does not qualify as human subjects research and is therefore exempt from IRB approval.

1. Introduction

LLMs have progressed from academic curiosities to tools that permeate nearly every facet of the clinical workflow, including triage, documentation, and guideline retrieval, promising practical diagnostic decision support on the horizon (Korom et al., 2025; Mehandru et al., 2024; Chen et al., 2024). The number of peer-reviewed studies assessing LLMs for direct patient-care tasks has climbed steeply in the past two years, mirroring a broader enthusiasm for foundation models across medicine (Khan et al., 2025). However,

1. https://github.com/lichingcat/From-Zero-Shot-to-Bedside-A-Practical-Playbook

© 2025 L.-C. Chen, T. Zack, D. Mandair, A. Mahadevan, A. Suresh, Y. Ishiyama, Y. Li, J.C. Hong & A.J. Butte.

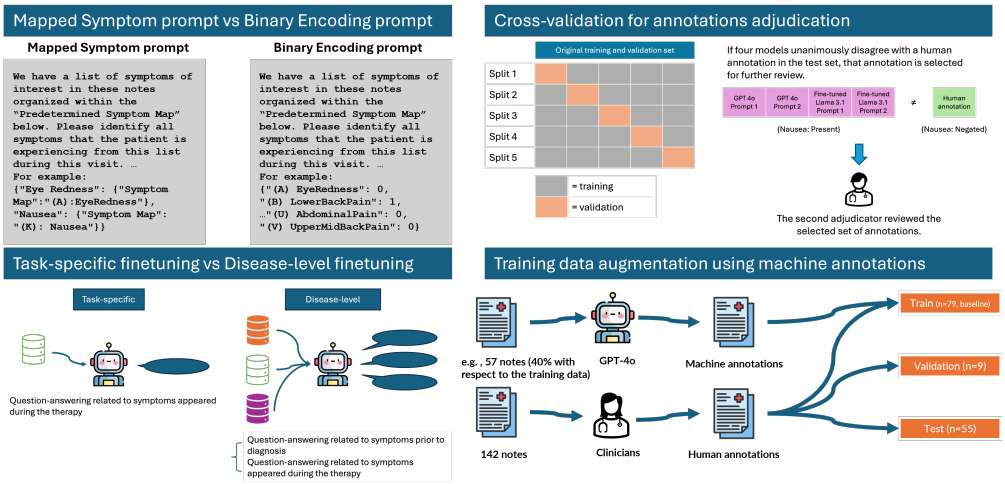


Figure 1: The overall framework of the proposed fine-tuning playbook.

despite this momentum, clinicians and informaticians still lack clear guidance on how best to tailor an LLM to a narrowly defined clinical objective.

A fundamental early choice concerns whether to rely on commercial, Application Programming Interface (API)-hosted systems such as Generative Pretrained Transformer (GPT)-4o or to deploy an open-source alternative, such as Meta Llama or the Mistral suite of models. Proprietary models typically deliver strong zero-shot performance, but their pay-per-token pricing and cloud-only hosting can be prohibitive at scale, and they raise non-trivial questions about routing protected health information (PHI) to third-party endpoints that may not satisfy institutional business-associate-agreement requirements. Open-source models, by contrast, incur only compute costs, can be fine-tuned behind the firewall, and permit unrestricted architectural modification, making them attractive for health-system internal training and deployment.

Even with an open-source model in hand, overall performance depends critically on three intertwined decisions: how to curate reliable task-specific data, which fine-tuning strategy to apply, and how best to balance human- and machine-generated annotations for cost-effective scale-up (Ding et al., 2024). The first of these in particular has posed a challenge for data from clinical documentation, as manual annotation for many of these tasks is arduous and prone to human error or judgment (Sylolypavan et al., 2023; Schroll et al., 2012). Especially for oncology toxicity encoding and grading, annotation discrepancy arises when the guideline is highly complex and involves subjectivity (Fairchild et al., 2020). For the other challenges, the machine learning literature offers general recipes for these steps, but few studies interrogate them in depth for clinical text where documentation is noisy, vocabularies are diverse, and tasks can be quite specific and poorly covered within pretraining data.

In order to explore effective strategies in fine-tuning open-source LLMs for clinical care while preserving the data privacy, we selected pancreatic cancer as a demonstrating example. Although only the 12th most common malignancy worldwide, pancreatic cancer ranks seventh in cancer-related mortality because its early, often nonspecific symptoms delay diagnosis and curative treatment (Bray et al., 2024; Schwingel et al., 2023). Population-level screening remains untenable, yet timely detection markedly improves outcomes, with surgical resection benefiting those identified at an early stage (Stoffel et al., 2023). Recent studies show that distinctive constellations of prodromal symptoms can presage disease onset, but machine learning models have so far drawn mainly on structured codes and questionnaires, overlooking the richer narratives in unstructured clinical text (Kenner et al., 2021; Placido et al., 2023; Malhotra et al., 2021; Appelbaum et al., 2021; Chen et al., 2021). Harnessing LLMs to extract and synthesize these narrative symptom patterns therefore offers a potential route to

earlier pancreatic cancer detection and, ideally, earlier intervention and improved outcomes.

To address these gaps, we systematically evaluate symptom extraction from de-identified pancreatic cancer notes by benchmarking open-source LLMs against GPT-4o, comparing two prompt designs, contrasting task-specific with disease-level adaptation, and quantifying mixes of high-confidence synthetic and expert labels; we also introduce a human-machine adjudication loop to target likely annotation errors. These experiments test the following hypotheses: (i) prompt design influences end-to-end extraction performance; (ii) a human-machine adjudication loop can concentrate expert review on likely label errors and improve downstream accuracy; and (iii) augmenting expert labels with high-confidence machine-generated annotations improves fine-tuned model performance under annotation resource constraints. Across pre-diagnosis and on-treatment settings, our experiments show that prompt framing shifts precision-recall trade-offs, adjudication concentrates expert effort while improving downstream accuracy, and balanced synthetic-human label mixtures strengthen fine-tuned models under annotation constraints. Taken together, these choices form a practical, privacy-preserving playbook for adapting open-source LLMs to specialized clinical-extraction tasks (see Figure 1).

2. Related Work

Recently, a significant interest in smaller, open-source LLMs has flourished in healthcare domain since it can be hosted locally and fine-tuned for specific institutional needs. While general-purpose models are powerful and surpass multiple benchmarks (Hendrycks et al., 2021; Wang et al., 2020; Srivastava et al., 2023), their closed-source nature, reliance on external APIs, and the inherent risks of transmitting privacy-sensitive data make them challenging to deploy within clinical settings (Wiest et al., 2025; Minssen et al., 2023; Wiest et al., 2024).

Medical text is replete with specialized jargon, acronyms, and note templates not well-represented in general pre-training data (Wang et al., 2024). Studies have repeatedly shown that open-source LLMs fine-tuned on medical or biomedical data can match or even outperform proprietary models (Chen et al., 2023; Singhal et al., 2025; Tran et al., 2024). This line of research, which adapts models for specific medical tasks using techniques like Low-Rank Adaptation (LoRA) (Hu et al., 2021), has become a cornerstone of clinical NLP (Christophe et al., 2024; Van Veen et al., 2024; Kim et al., 2025). Specifically, a wide array of NLP tasks can aid in the delivery of oncology care (Benson et al., 2025, 2024), such as reports generation (Rajaganapathy et al., 2025; Bhayana et al., 2024), disease progression prediction (Zhu et al., 2025), and information extraction (Hirakawa et al., 2025).

While optimizing LLMs through model-centric approaches is effective, there is growing attention on data-centric adaptation (Xu et al., 2024; Zhang et al., 2024; Hegselmann et al., 2024). This paradigm posits that the selection, curation, and quality of data are as critical as the model architecture itself, especially in data-constrained clinical environments. This area focuses on overcoming two primary challenges: data scarcity and label quality.

To tackle data scarcity, various methods have emerged for generating large-scale, low-cost training data. One approach is data augmentation using LLMs themselves to generate new instructions given human written instructions or generate labels to massive unlabeled data (Ding et al., 2024; Honovich et al., 2022; Schick and Schütze, 2021; Wang et al., 2022). Studies have used LLMs for data augmentation in different formats; collaborative annotation and data labeling can particularly aid in medical domain (Goel et al., 2023; Hegselmann et al., 2024; Cho et al., 2025).

The second challenge is label quality. Annotation efforts, even done by experts, are prone to errors, which introduce noise during finet-tuning. LLMs have emerged as an useful tools to identify potential label errors (Chong et al., 2022) by flagging inconsistencies between model prediction and human errors. This creates annotation cycles, where human experts review model-flagged errors, verify their correctness, and reuse the corrected data for future training and evaluation (Kiela et al., 2021; Goel et al., 2023). This process concentrates expert efforts on the most challenging or erroneous instances, thereby maximizing the limited annotation resource. In this work, we specifically examine and integrate these data-centric approaches for fine-tuning LLMs on information extraction from clinical notes, demonstrating their effectiveness in optimizing model performance and in allocating expert annotation resources efficiently.

3. Methods

This retrospective study examined patients from UCSF Information Commons between 2012 and 2024

(University of California, San Francisco, Academic Research Systems, 2022). The de-identified EHR were accessed through a HIPAA-compliant computing cluster. We focus on two distinct cohorts: (1) patients who were later diagnosed with pancreatic ductal adenocarcinoma (PDAC), analyzing their clinical notes recorded at least six months prior to their initial diagnosis; and (2) patients who were undergoing chemotherapy for treating PDAC.

Cohort 1: Pre-diagnosis patients and pre-diagnosis symptoms

We identified patients who were diagnosed with PDAC based on the presence of at least four documented International Classification of Diseases (ICD) codes associated with PDAC (any codes under C25 for ICD-10 and 157 for ICD-9) in their EHR and with clinical encounters in oncology departments at UCSF Medical Center at Mission Bay or at Mount Zion. We excluded patients with documented clinical codes or notes indicating pancreatic neuroendocrine tumors. Additionally, we generated a control cohort comprising patients without PDAC, propensity matched by age, gender, and race-ethnicity.

Our objective in this cohort was to characterize early-stage symptomatology to train predictive models of PDAC detection. Thus, we defined the index date as the earliest recorded diagnosis of PDAC in the EHR and extracted clinical notes documented between six months and five years prior to this index date. This earlier time gap is intended to facilitate identification of symptoms that significantly predate the PDAC diagnosis (Placido et al., 2023; Appelbaum et al., 2021; Malhotra et al., 2021).

Cohort 2: On-treatment patients and treatment-related symptoms

To expose models to a group of patients at a different stage of the cancer trajectory, we selected patients diagnosed with PDAC who were undergoing chemotherapy with FOLFIRINOX at UCSF Medical Center. Treatment status was determined based on drug administration data and by identifying the most recent medical oncologist's note within two weeks of a FOLFIRINOX infusion.

3.1. Data preparation

3.1.1. Note Sectionizing for Language model ingestion

We extracted relevant sections from the clinical notes using keyword-based segmentation to identify text corresponding to Subjective, Oncology History, and Assessment and Plan. We selected these three sections because: (1) the Subjective section captures the clinician's account of the patient's most prominent or concerning symptoms; (2) the Oncology History enriches the oncological context of the encounter; and (3) the Assessment and Plan often highlights the treatment plan or follow-up related to symptoms present during the encounter. The abridged notes were used for annotation, model training, and inference. Both datasets were partitioned into training, validation, and test sets at the note level.

3.1.2. Annotation of clinical notes

To establish ground truth for symptom documentation, we recruited seven physician annotators: two fellows, one attending gastrointestinal (GI) oncologist, and four residents. The annotation process began with a calibration phase involving two of these annotators, the GI oncologist and a fellow, who refined the guidelines to ensure consistency across annotators. Each annotator then independently annotated a random subset of clinical notes.

We annotated a total of 355 clinical notes from 127 patients in the pre-diagnosis cohort, including 312 notes from 113 patients with PDAC and 43 notes from 14 non-PDAC patients. For the on-treatment cohort, we annotated 142 notes from 94 patients receiving chemotherapy. The detailed of the symptom annotations are provided in Appendix A.

To further identify human discrepancies in annotations, we conducted a multi-fold LLM adjudication on the on-treatment cohort. We left out the original test data for the final hold-out validation. The original training and validation sets were used to create five folds of cross-validation. In each fold, there were training, validation, and test sets. The data points in the original train and validation sets were jointly covered by the fold-specific test sets, such that each training data point could potentially be adjudicated. After fine-tuning the model for each fold, the model was tested on the test set dedicated to that fold. We identified disagreements by comparing human annotations with zero-shot GPT-4o and with

the fine-tuned models under both prompts. In cases of discrepancies between human and machine annotations, a senior clinician specializing in pancreatic cancer care adjudicated all instances of human-machine disagreement with access to the note, original label and the label agreed upon by LLMs.

3.2. Training methods and procedures

3.2.1. fine-tuning Strategy

We fine-tuned the Llama-3.1-8B-Instruct model and Qwen3-8B on a HIPAA-compliant cluster using the framework proposed by (Liu et al., 2024). The framework utilizes quantized LoRA in the LLM fine-tuning automation pipeline (Dettmers et al., 2023). Hyperparameter settings are available in the codebase listed in the Data and Code availability. We evaluated performance in identifying the presence or absence of symptoms at the time of note writing using micro-averaged F1 score, precision, and recall.

3.3. Zero shot inference and Prompting Strategy

We investigated the zero-shot inference ability of GPT-4o, Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Gemma-9B-Instruct, Qwen3-8B, and Mixtral-8x7B-Instruct for extracting symptomatology from progress notes in our cohorts. Experiments were performed on a HIPAA-compliant computing cluster on Microsoft Azure with the Hugging Face API for open-source models and the UCSF Versa API for GPT-4o. We selected GPT-4o as our comparison point for the open-source models since it was the latest general instruction-following GPT available in the UCSF Versa API during the time of experimentation.

Prompt formatting and instruction framing have been shown to affect LLM performance in structured extraction and question-answering tasks (Sclar et al., 2023). Prior studies (Zhao et al., 2021; Reynolds and McDonell, 2021) suggest that even subtle prompt changes can lead to significant differences in behavior, motivating this comparative analysis. In clinical symptom extraction, where precision and structure are critical, we compared two prompting strategies with distinct representational goals. The first, **Mapped Symptom Extraction**, asks the model to extract only symptoms that are present and map them to predefined clinical codes. The second, **Binary Symptom Encoding**, prompts the model to exhaustively evaluate each symptom in a predefined list as either present (1) or absent (0) (see Appendix B). This comparison allows us to assess how different prompt framings impact sensitivity to presence/absence distinctions and alignment with clinical ontologies.

3.3.1. Disease-level adaptation vs. Task-specific adaptation

In practice, clinicians often copy and paste templates or prior notes into current documentation, inadvertently propagating outdated or irrelevant information and introducing inaccuracies or conflicting information. Understanding such notes often involves implicit reasoning about clinical context, temporality, and patient trajectory, in addition to EHR-specific nomenclature. We therefore evaluated whether fine-tuning LLMs on a broader set of clinical tasks, i.e., disease-level adaptation, enhances their ability to reason about these characteristics compared with task-specific adaptation on a narrower dataset.

We constructed the disease-level adaptation by combining data from two distinct clinical cohorts: one consisting of patients before treatment and another consisting of patients undergoing treatment. We fine-tuned Llama-3.1-8B-Instruct and Qwen3-8B on the combined dataset, using the same train, validation, and test splits defined for the task-specific experiments. The combined training data were randomly shuffled to ensure the model can converge on both tasks. In addition to the cohorts described above, we implemented and compared these two kinds of adaptations on an additional set of radiology reports detailed in Appendix D.

3.3.2. Training data augmentation

Given the limited availability of expert-annotated clinical data, we explored machine-generated annotations as a scalable alternative to manual labeling for fine-tuning. This approach follows knowledge distillation: a high-capacity teacher model produces training labels that supervise a smaller or more efficient student model. In our context, the teacher served as an automated annotator to alleviate annotation bottlenecks common in clinical NLP.

For experiments using machine-generated labels, we machine-annotated 1,285 previously unlabeled chemotherapy progress notes. We prompted GPT-4o with the Binary Symptom Encoding prompt to mark the presence/absence of each target symptom

based solely on note content. Specifically, a symptom was marked present if GPT-4o assigned a probability ≥ 0.8 to the token "1" for that symptom. To evaluate different machine-to-human label ratios, we incrementally added machine labels sampled without replacement from the machine-annotated pool to a fixed human-annotated training set at $0.4\times, 0.6\times, 0.8\times, 1\times, 2\times$, and $3\times$ the number of human labels. We did not perform human adjudication or apply any post-processing to the teacher outputs.

We trained separate Llama-3.1-8B-Instruct and Qwen3-8B models for each ratio using the augmented training sets and evaluated them on the human-annotated test set. To quantify variability, we ran the machine-label sampling and model training five times for each ratio. We report the mean and standard deviation of the micro-F1 across runs.

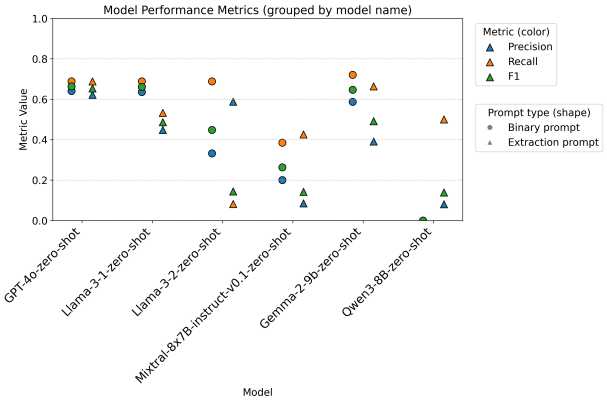
3.3.3. Statistical analysis

We report all the metrics including F1 score, precision and recall using sckit-learn (1.4.2) and SciPy (1.14.1). We obtain the 95% confidence intervals (CIs) using boostrap method with 5,000 samplings with replacement. We use Wilcoxon signed-rank tests for pairwise comparisons and correct for multiple comparisons with Holm–Bonferroni method.

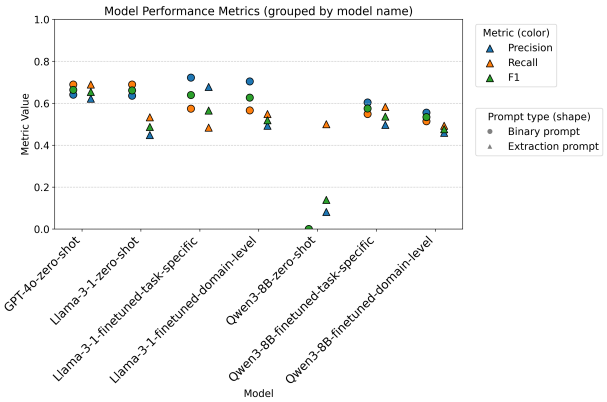
4. Results

We collected 497 notes, 355 from 127 patients prior to their diagnosis, and 142 from 94 patients on active systemic therapy. This provided a total of 10,721 data points for symptom presence or absence (633 presence, 10,088 absence) for 44 named symptoms (see Appendix A).

Using this dataset, we first assessed the zero-shot inference capabilities of GPT-4o, Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Gemma-9B-Instruct, Qwen3-8B, and Mixtral-8x7B-Instruct on symptom extraction from clinical progress notes (see Section 3). The performance metrics (micro-averaged F1 score, precision, recall) are detailed in Figures 3 and 2. For the pre-diagnosis cohort, we see that GPT-4o, Llama-3.1-8B-Instruct and Gemma-2-9B have similar performance when using the Binary Symptom Encoding prompt on zero-shot inference (overall F1 score: GPT-4o: 0.664, Llama-3.1-Instruct: 0.661, Gamma-2-9B: 0.647) (see Figure 2(a)). For the on-treatment cohort, zero-shot GPT-4o reaches an F1 score of 0.896, a precision of 0.904, and a recall of 0.888 when using the Mapped Symptom Extraction prompt (see Figure 3(a)).



(a) The pre-diagnosis cohort using pretrained models.

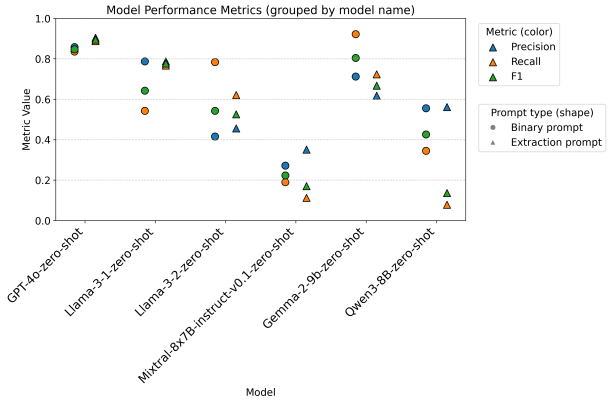


(b) The pre-diagnosis cohort using fine-tuned models with the baseline models on the left hand side (GPT-4o zero-shot, Llama-3.1-8B-instruct zero-shot and Qwen3-8B zero-shot).

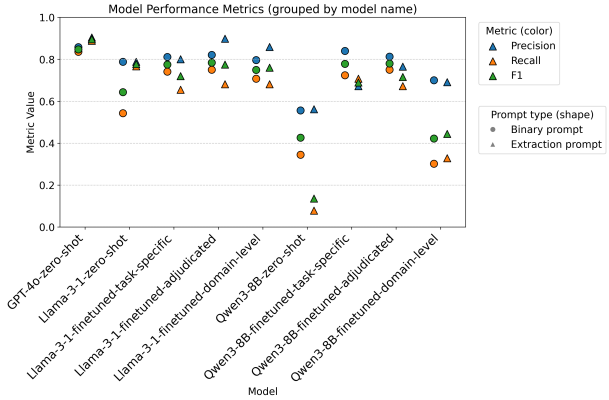
Figure 2: Performance scatter plots for the pre-diagnosis cohort with two different prompt formats.

4.1. Baseline Performance and Prompting Strategies

Next, we evaluated the hypothesis that a fine-tuned open-source LLM, Llama-3.1-8B-Instruct, can match or exceed proprietary models in symptom extraction performance and test whether specific prompting strategies showed benefit. We compared two prompting strategies, **Mapped Symptom Extraction** and **Binary Symptom Encoding**, evaluating their effect on fine-tuning LLMs on the symptom extrac-



(*a*) The on-treatment cohort using pretrained models.



(*b*) The on-treatment cohort using fine-tuned models with the baseline models on the left hand side (GPT-4o zero-shot, Llama-3.1-8B-instruct zero-shot and Qwen3-8B zero-shot).

Figure 3: Performance scatter plots for the on-treatment cohort with two different prompt formats.

tion task. For the on-treatment cohort, the task-specific fine-tuned Llama-3.1-8B-instruct using the Binary Symptom Encoding prompt performs better than the one using the Mapped Symptom Extraction prompt, where the F1 score for the former is 0.775, whereas it is 0.729 for the latter (see Figure 3(*b*)). The same also goes for the pre-diagnosis cohort (see Figure 2(*b*)). The task-specific fine-tuned Llama-3.1-8B-instruct using the Binary Symptom Encoding prompt outperforms the one using the other prompt, where the former achieved an F1 score of 0.639, and the latter 0.565. The performance of each LLM is detailed in Appendix C.

4.2. Identifying human error through multi-fold LLM adjudication

Human annotation, while considered the ground truth in many fields, does have a non-zero error rate. This is especially true for annotation tasks that require long periods of attention for nuanced language tasks, such as extracting information from lengthy clinical notes (Fairchild et al., 2020). Often, these errors can be mitigated by having multiple humans perform each tasks to ensure annotation quality and validity. However, in practice, human resources are scarce, and as the complexity of the task increases, it becomes even harder to collect annotations at scale in a timely fashion. To reduce the burden of second-pass annotation, we propose to direct human annotation efforts towards specific examples most likely to require multiple experts. We conducted this by fine-tuning 5 LLMs on the data and identifying annotations where all LLMs disagreed with the human annotations on the held out set (see Section 3), with the hypothesis that these would be enriched for human errors.

We found that in cases where all four models disagreed with the human annotator, a second human identified these as human error 48% of the time (37 out of 76). After the adjudication of the human annotator, we fine-tuned the models again on the adjudicated data with two different prompts, and compare the performance using the same test set. The Llama-3.1-8B-Instruct using the adjudicated data yields better performance in both prompt setups (see Figure 4). Specifically, for the Mapped Symptom Extraction prompt, the model fine-tuned using adjudication improves 12% of precision from 0.800 to 0.898, 4% of recall from 0.655 to 0.681, and 8% of F1 score from 0.720 to 0.775 suggesting these human errors were limiting accuracy of the fine-tuned models. Results were similar for Qwen3-8B (see Figure 4 and Appendix Table 4)

4.3. Disease-level vs. Task-specific Adaptation

We tested whether disease-level fine-tuning across multiple clinical cohorts improves symptom extraction accuracy compared to task-specific fine-tuning. For the on-treatment cohort, the disease-level model performs better than the task-specific model when trained using the Mapped Symptom Extraction prompt, with the former achieving an F1 score of 0.760 and 0.720 for the latter (see Figure 3(*b*)),

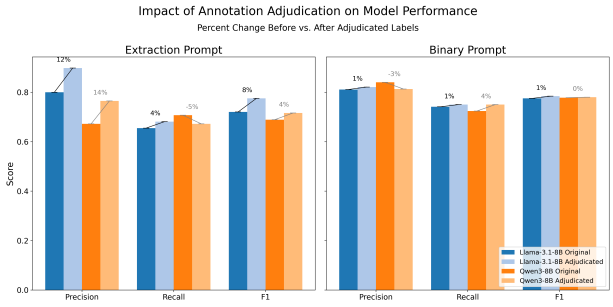


Figure 4: The percentage of change in the fine-tuned Llama-3.1-8B-Instruct and the fine-tuned Qwen3-8B before and after adjudicating the annotations using cross-validation with machine assistance on on-treatment cohort.

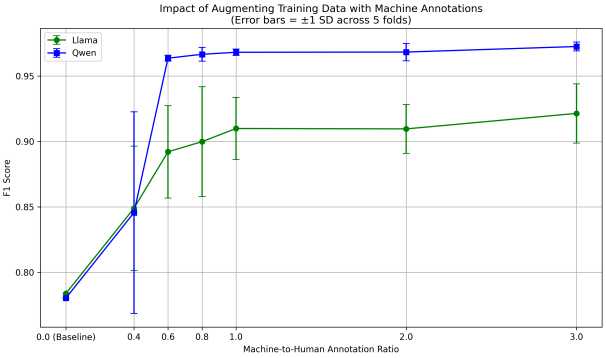


Figure 5: Augmenting human-labeled training data with machine-annotated notes improves F1. For each machine:human ratio p, we sampled machine labels without replacement from a pool disjoint from the human train/test split, fine-tuned a separate model using human and machine annotations, and evaluated on a human-labeled test set. Performance rises with p and plateaus around $p \approx 1$–2, consistently above the baseline ($p = 0$). The baseline F1 score are 0.784 and 0.780 from Llama-3.1-8B-instruct and Qwen3-8B fine-tuned on adjudicated human annotations using Binary Encoding prompt, respectively.

when using Llama-3.1-8B-Instruct. However, the disease-level fine-tuning did not improve the performance for the pre-diagnosis cohort (see Figure 2(b)). For Llama-3.1-8B-Instruct, the disease-level fine-tuned model using the Mapped Symptom Extraction prompt achieved an F1 score of 0.519, while the task-specific model using the same prompt achieved an F1 score of 0.565.

4.4. Impact of Training Data Augmentation

We systematically evaluated the impact of varying ratios of machine-generated annotations relative to human annotations in our training set across two open LLMs, Llama-3.1-8B-Instruct and Qwen3-8B. We considered machine-to-human annotation ratios of $0.4\times, 0.6\times, 0.8\times, 1\times, 2\times$, and $3\times$, selecting corresponding numbers of notes from the machine-annotated dataset to augment the training data. Additionally, to ensure that machine-generated labels have sufficient confidence, we impose a confidence threshold based on token-level probabilities (see Section 3). Notably, both models show a similar performance trend with respect to data augmentation. As the ratio of machine annotations increases, the model progressively improves its performance and eventually outperforms the baseline model fine-tuned only on human annotations. As the ratio of machine annotations becomes dominant, i.e., the machine annotations are at least two times more than the human annotations, the fine-tuned model reaches at least an F1 score over 0.9 but appears to plateau beyond this (see Figure 5 and Appendix Table 5).

5. Discussion

This work presents a practical playbook for adapting open-source LLMs to a specialized clinical symptom-extraction task, yielding insights for future clinical NLP deployments. First, our experiments showed that fine-tuned open-source models can approach the performance of a proprietary model in extracting complex symptom information from clinical notes. With appropriate prompt design and domain-specific training, the open-source model achieved micro-averaged F1 scores comparable to GPT-4o's zero-shot performance for certain cohorts. This suggests that health systems can leverage privacy-preserving, cost-effective open-source models with minimal loss in accuracy. We also found that prompt formatting meaningfully affects extraction performance: a binary symptom encoding strategy, which exhaustively evaluates each candidate symptom as present or absent, outperformed a mapped symptom extraction strategy that outputs only detected symptoms with coded labels. Explicitly forcing the model to consider every possible symptom yielded more complete and precise outputs, an insight that can guide prompt design for similar structured clinical tasks.

Second, a key methodological contribution is an LLM-based adjudication mechanism to improve annotation quality. We leveraged an ensemble of LLMs in a cross-validation scheme to flag instances where model predictions consistently disagreed with initial human labels. Flagged cases were enriched for true annotation errors; after correcting these labels and retraining, we observed a clear boost in the model's precision and overall F1 score. This finding has important implications for clinical NLP: even expert-generated data can contain non-trivial errors, especially for complex, time-consuming tasks. Using LLM consensus to identify likely mistakes offers a cost-effective way to focus expert review on the most problematic labels, thereby improving dataset quality with limited additional cost. Future applications could extend this technique to other domains, pairing human expertise with model feedback to iteratively refine labels in a human-in-the-loop process.

Third, our results shed light on the benefits and limits of using machine-generated annotations to augment training data. Augmenting the human-labeled dataset with additional notes annotated by GPT-4o markedly improved the fine-tuned model's performance up to a point, confirming the value of knowledge distillation from a stronger model. Notably, however, we observed a plateau in performance gains once synthetic data greatly exceeded human labels. Beyond roughly a two-to-one or three-to-one ratio of synthetic to human examples, the model's micro F1 score leveled off. This suggests diminishing returns from additional synthetic data, which may introduce repetitive patterns or noise that the student model cannot effectively leverage. For practitioners, this result indicates that while model-generated labels are an effective way to augment training data on limited annotations, one should not expect unlimited improvements by simply adding more synthetic examples. Instead, there may be an optimal mix of expert and synthetic annotations, enough to expand coverage beyond the expert set without overwhelming the distinctive signal in human labels.

Beyond the core findings, our evaluations offer practical guidance for deployment. Choosing between task-specific fine-tuning and multi-cohort (disease-level) fine-tuning depends on data similarity and noise. In our clinical-note experiments, combining related cohorts sometimes improved robustness on complex notes, for example, on-treatment chemotherapy notes with templated text and evolving patient status, whereas narrowly focused training performed best for certain subsets. By contrast, in the radiology experiments (see Appendix D), a model fine-tuned jointly across multiple radiology tasks outperformed a single-task model on every individual task. A likely reason is that radiology reports are well-structured and information-dense, enabling high-quality human annotations and more reliable pattern learning. On the other hand, progress notes tied to a specific context (e.g., chemotherapy) can poorly capture the complex temporality of a patient's trajectory, which is constrained by the clinician's capacities and the extensive care team involved. Overall, the benefit of multi-task or multi-cohort training appears to hinge on the similarity of tasks and the noise characteristics of the combined data.

Additionally, our work reinforces that relatively small open-source LLMs can often be further enhanced with modest, institution-specific adaptation to deliver strong performance, enabling data governance and cost control compared with third-party APIs.

Limitations and future directions This study was conducted at a single academic health system and centered on pancreatic cancer care; generalizability to other specialties, institutions, and languages remains to be tested. While we expect the framework to transfer, optimal prompts, model choices, and the effectiveness and limits of machine-annotated data will likely vary by domain. Future work could explore more automated or systematic methods to not only detect but also correct such discrepancies. For example, by having LLMs provide explanations or rationales for their predictions to help human annotators understand and resolve disagreements.

6. Conclusion

In sum, we offer actionable guidance for adapting LLMs to specialized healthcare extraction tasks. By combining prompt design, targeted error correction, and balanced data augmentation, an open-source model can achieve high symptom-extraction accuracy under privacy and cost constraints. Extracting nuanced symptom trajectories from unstructured text has immediate utility for earlier detection and monitoring. More broadly, our findings support a shift from generic pretrained LLMs to privacy-preserving, locally adapted clinical systems that continually learn under governance and have the potential to improve patient outcomes.

Acknowledgments

We appreciate reviewers for their valuable feedback. We gratefully acknowledge Gavin Hui and Zahra Heidary for their effort in annotating the dataset. We thank UCSF Information Commons for supporting the research with computing resources and a de-identified database. In particular, we thank the UCSF AI Tiger Team, Academic Research Services, Research Information Technology, the Chancellor's Task Force for their data released asset and technical support on computing environments and Versa API gateway (UCSF secure API gateway to large language models and generative AI).

This study was supported in part by Weill Cancer Hub West and the National Cancer Institute of the National Institutes of Health (R01CA277782), which had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Limor Appelbaum, José P Cambronero, Jennifer P Stevens, Steven Horng, Karla Pollick, George Silva, Sebastien Haneuse, Gail Piatkowski, Nordine Benhaga, Stacey Duey, Mary A Stevenson, Harvey Mamon, Irving D Kaplan, and Martin C Rinard. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. *Eur. J. Cancer*, 143:19–30, January 2021.

Ryzen Benson, Marianna Elia, Benjamin Hyams, Ji Hyun Chang, and Julian C Hong. A narrative review on the application of large language models to support cancer care and research. *Yearbook of Medical Informatics*, 33(01):090–098, 2024.

Ryzen Benson, Clodagh Kenny, Amir Ashraf Ganjouei, Michelle Zhao, Rami Darawsheh, Alexander Qian, and Julian C Hong. Large language models in population oncology: A contemporary review on the use of large language models to support data collection, aggregation, and analysis in cancer care and research. *JCO Clinical Cancer Informatics*, 9: e2500112, 2025.

Rajesh Bhayana, Bipin Nanda, Taher Dehkharghanian, Yangqing Deng, Nishaant Bhambra, Gavin Elias, Daksh Datta, Avinash Kambadakone, Chaya G Shwaartz, Carol-Anne Moulton, et al. Large language models for automated synoptic reports and resectability categorization in pancreatic cancer. *Radiology*, 311(3):e233117, 2024.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, 74(3):229–263, May 2024.

Li-Ching Chen, Travis Zack, Arda Demirci, Madhumita Sushil, Brenda Miao, Corynn Kasap, Atul Butte, Eric A Collisson, and Julian C Hong. Assessing large language models for oncology data inference from radiology reports. *JCO Clinical Cancer Informatics*, 8:e2400126, 2024.

Qinyu Chen, Daniel R Cherry, Vinit Nalawade, Edmund M Qiao, Abhishek Kumar, Andrew M Lowy, Daniel R Simpson, and James D Murphy. Clinical data prediction model to identify patients with early-stage pancreatic cancer. *JCO Clin. Cancer Inform.*, 5(5):279–287, March 2021.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

Dongkyu Cho, Miao Zhang, and Rumi Chunara. Expert-guided clinical text augmentation via query-based model collaboration. *arXiv preprint arXiv:2509.21530*, 2025.

Derek Chong, Jenny Hong, and Christopher D Manning. Detecting label errors by using pre-trained language models. *arXiv preprint arXiv:2205.12702*, 2022.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *arXiv [cs.LG]*, May 2023.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.97. URL https://aclanthology.org/2024.findings-acl.97/.

Andrew T Fairchild, Jarred P Tanksley, Jessica D Tenenbaum, Manisha Palta, and Julian C Hong. Interrater reliability in toxicity identification: Limitations of current standards. *Int. J. Radiat. Oncol. Biol. Phys.*, 107(5):996–1000, August 2020.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. LLMs accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR, December 2023.

Stefan Hegselmann, Shannon Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. A data-centric approach to generate faithful and high quality patient summaries with large language models. *arXiv preprint arXiv:2402.15422*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hiroshi Hirakawa, Koichiro Yasaka, Takuto Nomura, Rin Tsujimoto, Yuki Sonoda, Shigeru Kiryu, and Osamu Abe. Fine-tuned large language model for extracting pretreatment pancreatic cancer according to computed tomography radiology reports. *Journal of Imaging Informatics in Medicine*, pages 1–8, 2025.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv [cs.CL]*, June 2021.

Barbara Kenner, Suresh T Chari, David Kelsen, David S Klimstra, Stephen J Pandol, Michael Rosenthal, Anil K Rustgi, James A Taylor, Adam Yala, Noura Abul-Husn, Dana K Andersen, David Bernstein, Søren Brunak, Marcia Irene Canto, Yonina C Eldar, Elliot K Fishman, Julie Fleshman, Vay Liang W Go, Jane M Holt, Bruce Field, Ann Goldberg, William Hoos, Christine Iacobuzio-Donahue, Debiao Li, Graham Lidgard, Anirban Maitra, Lynn M Matrisian, Sung Poblete, Laura Rothschild, Chris Sander, Lawrence H Schwartz, Uri Shalit, Sudhir Srivastava, and Brian Wolpin. Artificial intelligence and early detection of pancreatic cancer: 2020 summative review. *Pancreas*, 50(3):251–279, March 2021.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Rev. Biomed. Eng.*, PP(99):1–20, May 2025.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, et al. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine*, 8(1):240, 2025.

Robert Korom, Sarah Kiptinness, Najib Adan, Kassim Said, Catherine Ithuli, Oliver Rotich, Boniface Kimani, Irene King'ori, Stellah Kamau, Elizabeth Atemba, Muna Aden, Preston Bowman, Michael Sharman, Rebecca Soskin Hicks, Rebecca Distler, Johannes Heidecke, Rahul K Arora, and Karan Singhal. AI-based clinical decision support for primary care: A real-world study. *arXiv [cs.CL]*, July 2025.

Longchao Liu, Long Lian, Yiyan Hao, Aidan Pace, Elaine Kim, Nour Homsi, Yash Pershad, Liheng

Lai, Thomas Gracie, Ashwin Kishtagari, Peter R Carroll, Alexander G Bick, Anobel Y Odisho, Maggie Chung, and Adam Yala. Human-level information extraction from clinical reports with fine-tuned language models. *medRxiv*, page 2024.11.18.24317466, November 2024.

Ananya Malhotra, Bernard Rachet, Audrey Bonaventure, Stephen P Pereira, and Laura M Woods. Can we screen for pancreatic cancer? identifying a subpopulation of patients at high risk of subsequent diagnosis using machine learning techniques applied to primary care data. *PLoS One*, 16(6):e0251876, June 2021.

Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J Butte, and Ahmed Alaa. Evaluating large language models as agents in the clinic. *NPJ Digit Med*, 7(1):84, April 2024.

Timo Minssen, Effy Vayena, and I Glenn Cohen. The challenges for regulating medical use of chatgpt and other large language models. *Jama*, 330(4):315–316, 2023.

Davide Placido, Bo Yuan, Jessica X Hjaltelin, Chunlei Zheng, Amalie D Haue, Piotr J Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, Alexander Chowdhury, Alexandra Franz, Lauren Brais, Elizabeth Andrews, Debora S Marks, Aviv Regev, Siamack Ayandeh, Mary T Brophy, Nhan V Do, Peter Kraft, Brian M Wolpin, Michael H Rosenthal, Nathanael R Fillmore, Søren Brunak, and Chris Sander. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat. Med.*, 29(5):1113–1122, May 2023.

Sivaraman Rajaganapathy, Shaika Chowdhury, Xiaodi Li, Vincent Buchner, Zhe He, Rui Zhang, Xiaoqian Jiang, Ping Yang, James R Cerhan, and Nansu Zong. Synoptic reporting by summarizing cancer pathology reports using large language models. *npj Health Systems*, 2(1):11, 2025.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv [cs.CL]*, February 2021.

Timo Schick and Hinrich Schütze. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*, 2021.

Jeppe Bennekou Schroll, Emma Maund, and Peter C Gøtzsche. Challenges in coding adverse events in clinical trials: a systematic review. *PLoS One*, 7(7):e41174, July 2012.

Jérôme Schwingel, Miriam Decker, Lisa Schneider, Carsten Johannes Stürmer, and Manfred P Lutz. Early detection of pancreatic cancer. *Oncol. Res. Treat.*, 46(6):259–267, April 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *arXiv [cs.CL]*, October 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Elena M Stoffel, Randall E Brand, and Michael Goggins. Pancreatic cancer: Changing epidemiology and new approaches to risk assessment, early detection, and prevention. *Gastroenterology*, 164(5):752–765, April 2023.

Aneeta Sylolypavan, Derek Sleeman, Honghan Wu, and Malcolm Sim. The impact of inconsistent human annotations on AI driven clinical decision making. *NPJ Digit. Med.*, 6(1):26, February 2023.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832, 2024.

University of California, San Francisco, Academic Research Systems. UCSF DeID CDW-OMOP, June 2022. URL https://data.ucsf.edu/research/deid-data. Available through UCSF Data Portal.

University of California, San Francisco, Academic Research Systems. UCSF DeID CDW (2024-December) [Dataset], 2024. URL https://data.ucsf.edu/research/deid-data.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020. URL https://arxiv.org/abs/1905.00537.

Andy Wang, Cong Liu, Jingye Yang, and Chunhua Weng. Fine-tuning large language models for rare disease concept normalization. *Journal of the American Medical Informatics Association*, 31(9): 2076–2083, 2024.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Isabella C Wiest, Marie-Elisabeth Leßmann, Fabian Wolf, Dyke Ferber, Marko Van Treeck, Jiefu Zhu, Matthias P Ebert, Christoph Benedikt Westphalen, Martin Wermke, and Jakob Nikolas Kather. Deidentifying medical documents with local, privacy-preserving large language models: the llm-anonymizer. *NEJM AI*, 2(4):AIdbp2400537, 2025.

Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K Meyer, Radhika Juglan, Zunamys I Carrero, Daniel Paech, Jens Kleesiek, Matthias P Ebert, et al. Privacy-preserving large language models for structured medical information retrieval. *NPJ Digital Medicine*, 7(1):257, 2024.

Xinyi Xu, Zhaoxuan Wu, Rui Qiao, Arun Verma, Yao Shu, Jingtan Wang, Xinyuan Niu, Zhenfeng He, Jiangwei Chen, Zijian Zhou, Gregory Kang Ruey Lau, Hieu Dao, Lucas Agussurja, Rachael Hwee Ling Sim, Xiaoqiang Lin, Wenyang Hu, Zhongxiang Dai, Pang Wei Koh, and Bryan Kian Hsiang Low. Position paper: Data-centric AI in the age of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11895–11913, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.695. URL https://aclanthology.org/2024.findings-emnlp.695/.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5HCnKDeTws.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *arXiv [cs.CL]*, February 2021.

Menglei Zhu, Hui Lin, Jue Jiang, Abbas J Jinia, Justin Jee, Karl Pichotta, Michele Waters, Doori Rose, Nikolaus Schultz, Sulov Chalise, et al. Large language model trained on clinical oncology data predicts cancer progression. *npj Digital Medicine*, 8(1):397, 2025.

Appendix A. Data Description

We constructed two cohorts who were in different phase of pancreatic cancer trajectory. The first cohort, called pre-diagnosis cohort, were a set of patients, who were not yet diagnosed with pancreatic cancer at the time of note writing. To validate their cancer status, we used the number of ICD codes they received for pancreatic cancer and their encounter history from the EHR as mention in Section 3. We collected 113 patients with PDAC diagnosis and annotated 312 notes written before their first cancer diagnosis. We also extracted a control group comprising 14 patients and annotated 43 notes for the pre-diagnosis cohort. The positive counts of each symptom of the pre-diagnosis cohort is listed in Table 1.

The second cohort, called on-treatment cohort, consisted of patients diagnosed with PDAC and were undergoing chemotherapy with FOLFIRINOX at UCSF Medical Center. We annotated 142 notes from 94 patients for this cohort. The positive counts of each symptom of the on-treatment cohort is listed in Table 2.

Table 1: Positive counts of each symptom and prevalence over a total 355 notes in the pre-diagnosis cohort.

Symptom	**Positive Count**	**Prevalence**
Lower Back Pain	72	0.205
Weight Loss	11	0.031
Appetite Loss	12	0.034
Jaundice	1	0.003
Pruritus	1	0.003
Indigestion	7	0.020
Steatorrhea	0	0.000
Urine Color Change	5	0.014
Constipation	9	0.026
Nausea	29	0.083
Vomiting	13	0.037
Diarrhea	21	0.060
Gas or Bloating	9	0.026
Fatigue/Malaise/Lethargy	29	0.083
Early Satiety	4	0.011
Blood Glucose	4	0.011
GI Bleed	1	0.003
Melena	6	0.017
Bright Red Blood per Rectum	5	0.014
Abdominal Pain	50	0.142
Upper Mid Back Pain	7	0.020

Appendix B. Clinically-Curated Prompts for Symptom Extraction

We created two types of prompt for two specific tasks. The first type of prompt called **Mapped Symptom Extraction** inquires LLMs to extract the symptoms that are presenting in the patient given the clinical text and maps the symptoms to the given list of symptom of interest. The second type of prompt called **Binary Symptom Encoding** asks the the model to explicitly list out the binary state of presence for every symptom of interest. The symptoms that we are interested in for the two particular tasks are listed below.

Table 2: Positive counts of each symptom and prevalence over a total 142 notes in the on-treatment cohort.

Symptom	**Positive Count**	**Prevalence**
Diarrhea	36	0.254
Constipation	23	0.162
Nausea	47	0.331
Vomiting	8	0.056
Abdominal Pain	36	0.254
Abdominal Distension	3	0.021
Fatigue	53	0.373
Allergic reaction	1	0.007
Weight loss	13	0.092
Erythema	0	0.000
Hair loss	5	0.035
Neutropenia	2	0.014
Anemia	13	0.092
Abnormal liver function	5	0.035
Dyspnea	2	0.014
Appetite Loss	21	0.148
Fever	3	0.021
Chills	0	0.000
Jaundice	5	0.035
Thrombocytopenia	9	0.063
Sensory Neuropathy	34	0.239
Motor Neuropathy	1	0.007
Cold-induced Neuropathy	17	0.120

Pre-diagnosis prompt

Mapped Symptom Extraction

```
We have a list of symptoms of interest in these notes organized within the
"Predetermined Symptom Map" below.
Please identify all symptoms that the patient is experiencing from this list during this visit.
If the patient is not experiencing the symptom currently,
even if it is mentioned or the patient has experienced it in the past, do not include it.
If a symptom does not match any item in the map, you do not need to include it.
    Predetermined Symptom Map:
    (A): EyeRedness
    (B): LowerBackPain
    (C): WeightLoss
    (D): AppetiteLoss
    (E): Jaundice
    (F): Pruritus
    (G): Indigestion
    (H): Steatorrhea
    (I): Urine Color Change
    (J): Constipation
    (K): Nausea
    (L): Vomiting
    (M): Diarrhea
    (N): GasorBloating
    (O): FatigueMalaiseLethargy
    (P): EarlySatiety
    (Q): BloodGlucose
    (R): GI_Bleed
    (S): Melena
    (T): BRBPR
    (U): AbdominalPain
    (V): UpperMidBackPain
Your answer should be in json format.
For example: {\n"Eye Redness": {\n"Symptom Map": "(A): EyeRedness"\n},\n"Nausea":
{\n"Symptom Map": "(K): Nausea"\n}\n}
If no symptoms are identified, return an empty response: {\n"N/A": {\n"Symptom Map":
"(Z): None"\n}\n}
Provide your answer in the given json format.
```

Binary Symptom Encoding

```
We have a list of symptoms of interest in these notes organized within the
"Predetermined Symptom Map" below.
Please identify all symptoms that the patient is experiencing from this list during this visit.
If the patient is not experiencing the symptom currently,
even if it is mentioned or the patient has experienced it in the past, do not include it.
If a symptom does not match any item in the map, you do not need to include it.
    Predetermined Symptom Map:
    (A): EyeRedness
    (B): LowerBackPain
    (C): WeightLoss
    (D): AppetiteLoss
    (E): Jaundice
    (F): Pruritus
```

```
    (G): Indigestion
    (H): Steatorrhea
    (I): Urine Color Change
    (J): Constipation
    (K): Nausea
    (L): Vomiting
    (M): Diarrhea
    (N): GasorBloating
    (O): FatigueMalaiseLethargy
    (P): EarlySatiety
    (Q): BloodGlucose
    (R): GI_Bleed
    (S): Melena
    (T): BRBPR
    (U): AbdominalPain
    (V): UpperMidBackPain
Your answer should be in json format.
For example: {"(A) EyeRedness": 0, "(B) LowerBackPain": 1, "(C)WeightLoss": 0,
"(D) AppetiteLoss": 0, "(E) Jaundice": 0, "(F) Pruritus": 0,"(G) Indigestion": 0,
"(H) Steatorrhea": 0, "(I) UrineColorChange": 0, "(J)Constipation": 0, "(K) Nausea": 0,
"(L) Vomiting": 0, "(M) Diarrhea": 0,"(N) GasorBloating": 0, "(O) FatigueMalaiseLethargy": 0,
"(P) EarlySatiety":0, "(Q) BloodGlucose": 0, "(R) GI_Bleed": 0, "(S) Melena": 0, "(T) BRBPR":0,
"(U) AbdominalPain": 0, "(V) UpperMidBackPain": 0}
Provide your answer in the given json format.
```

On-treatment prompt

Mapped Symptom Extraction

```
We have a list of symptoms of interest in these notes organized within the
"Predetermined Symptom Map" below.
Please identify all symptoms that the patient is experiencing from this list during this visit.
If the patient is not experiencing the symptom currently,
even if it's mentioned or the patient has experienced it in the past, do not include it.
If a symptom does not match any item in the map, you do not need to include it.
Predetermined Symptom Map:
      (A): Diarrhea
      (B): Constipation
      (C): Nausea
      (D): Vomiting
      (E): Abdominal Pain
      (F): Abdominal Distension
      (G): Fatigue
      (H): Allergic reaction
      (I): Weight loss
      (J): Erythema
      (K): Hair loss
      (L): Neutropenia
      (M): Anemia
      (N): Abnormal liver function
      (O): Dyspnea
      (P): Appetite Loss
      (Q): Fever
      (R): Chills
```

```
      (S): Jaundice
      (T): Thrombocytopenia
      (U): Sensory Neuropathy
      (V): Motor Neuropathy
      (W): Cold-induced Neuropathy
      (Z): Other (Specify the symptom name)
 Your answer should be in json format.
 For example: {\Abdominal Distension": {\Symptom Map":\(F)"},
    \Nausea":{\Symptom Map":"(C)"}}
 If no symptoms are identified: {"N/A": {"Symptom Map": "(Z)"}}
 Provide your answer in the given json format.
```

Binary Symptom Encoding

```
We have a list of symptoms of interest in these notes organized within the
"Predetermined Symptom Map" below.
Please identify all symptoms that the patient is experiencing from this list during this visit.
If the patient is not experiencing the symptom currently,
even if it's mentioned or the patient has experienced it in the past, do not include it.
If a symptom does not match any item in the map, you do not need to include it.
 Predetermined Symptom Map:
      (A): Diarrhea
      (B): Constipation
      (C): Nausea
      (D): Vomiting
      (E): Abdominal Pain
      (F): Abdominal Distension
      (G): Fatigue
      (H): Allergic reaction
      (I): Weight loss
      (J): Erythema
      (K): Hair loss
      (L): Neutropenia
      (M): Anemia
      (N): Abnormal liver function
      (O): Dyspnea
      (P): Appetite Loss
      (Q): Fever
      (R): Chills
      (S): Jaundice
      (T): Thrombocytopenia
      (U): Sensory Neuropathy
      (V): Motor Neuropathy
      (W): Cold-induced Neuropathy
Your answer should be in json format that contains all the symptoms with binary label.
For example: {"(A): Diarrhea": 1, "(B): Constipation": 0, "(C): Nausea": 1,
"(D): Vomiting": 0, "(E): Abdominal Pain": 1, "(F): Abdominal Distension": 0,
"(G): Fatigue": 1, "(H): Allergic reaction": 0, "(I): Weight loss": 1,
"(J): Erythema": 0, "(K): Hair loss": 1, "(L): Neutropenia": 0, "(M): Anemia": 1,
"(N): Abnormal liver function": 0, "(O): Dyspnea": 1, "(P): Appetite Loss": 0,
"(Q): Fever": 1, "(R): Chills": 0, "(S): Jaundice": 1, "(T): Thrombocytopenia": 0,
"(U): Sensory Neuropathy": 1, "(V): Motor Neuropathy": 0, "(W): Cold-induced Neuropathy": 1}
Provide your answer in the given json format.
```

Appendix C. LLMs Performance

We include the detailed results of each LLM in this section. We report the F1 score, precision and recall based on the overall performance across all symptoms. We compute the 95% CIs across 5,000 bootstrap samples.

Table 3: LLMs performance with zero-shot learning and fine-tuning with two different prompts on the pre-diagnosis cohort.

Model	F1 (95% CI)	Precision (95% CI)	Recall (95% CI)
Mapped Symptom Extraction			
GPT-4o zero-shot	0.654 (0.578–0.720)	0.622 (0.537–0.702)	0.689 (0.597–0.765)
Llama-3.1-8B zero-shot	0.487 (0.410–0.558)	0.448 (0.368–0.529)	0.533 (0.444–0.618)
Llama-3.2-3B zero-shot	0.144 (0.078–0.241)	0.588 (0.333–0.810)	0.082 (0.041–0.143)
Mixtral-8x7B-instruct-v0.1 zero-shot	0.142 (0.111–0.179)	0.085 (0.065–0.109)	0.426 (0.339–0.516)
Gemma-2-9b zero-shot	0.492 (0.425–0.559)	0.391 (0.327–0.460)	0.664 (0.574–0.746)
Qwen3-8B zero-shot	0.139 (0.110–0.172)	0.081 (0.063–0.102)	0.500 (0.41–0.587)
Llama-3.1-8B fine-tuned task-specific	0.565 (0.483–0.646)	0.678 (0.565–0.769)	0.484 (0.395–0.570)
Llama-3.1-8B fine-tuned domain-level	0.519 (0.441–0.592)	0.493 (0.407–0.578)	0.549 (0.461–0.636)
Qwen3-8B fine-tuned task-specific	0.536 (0.464–0.606)	0.497 (0.418–0.579)	0.582 (0.495–0.669)
Qwen3-8B fine-tuned domain-level	0.476 (0.404–0.545)	0.459 (0.379–0.540)	0.493 (0.411–0.575)
Binary Symptom Encoding			
GPT-4o zero-shot	0.664 (0.593–0.730)	0.641 (0.554–0.720)	0.689 (0.601–0.764)
Llama-3.1-8B zero-shot	0.661 (0.593–0.723)	0.636 (0.551–0.715)	0.689 (0.602–0.765)
Llama-3.2-3B zero-shot	0.448 (0.385–0.513)	0.332 (0.275–0.391)	0.689 (0.600–0.765)
Mixtral-8x7B-instruct-v0.1 zero-shot	0.263 (0.206–0.327)	0.200 (0.154–0.257)	0.385 (0.303–0.475)
Gemma-2-9b zero-shot	0.647 (0.574–0.711)	0.587 (0.506–0.665)	0.721 (0.636–0.797)
Qwen3-8B zero-shot	0.000 (N/A)	0.000 (N/A)	0.000 (N/A)
Llama-3.1-8B fine-tuned task-specific	0.639 (0.560–0.712)	0.722 (0.628–0.806)	0.574 (0.482–0.658)
Llama-3.1-8B fine-tuned domain-level	0.627 (0.545–0.698)	0.704 (0.608–0.787)	0.566 (0.473–0.653)
Qwen3-8B fine-tuned task-specific	0.575 (0.495–0.654)	0.604 (0.510–0.691)	0.549 (0.456–0.635)
Qwen3-8B fine-tuned domain-level	0.534 (0.456–0.607)	0.555 (0.468–0.639)	0.514 (0.431–0.602)

Appendix D. Radiology Report Experiments

In this section, we detail our implementation of the framework on another dataset. We selected 164 patients having the presence of Malignant Neoplasm of the Pancreas, excluding Malignant Neoplasm of Endocrine Pancreas, and treated for this disease at UCSF Medical Center, with 203 radiology reports. Reports were annotated by a medical oncologist specialized in treating pancreatic tumors for three tasks: (1) determining the disease status across 7 nuanced clinical categories, (2) locating the cancer, and (3) locating indeterminate findings potentially concerning cancer within 15 organs.

We elicit responses from LLMs for these three tasks with the prompts listed in Appendix D. We queried the GPT-4 model in March 2024 using the OpenAI API. We ran the Llama-3-1-8B-instruct model with 4-bit quantization. We fine-tuned the model with the learning weight set to 0.001 and training duration 3 epochs using the low-rank representation technique (Liu et al., 2024). We split the data into train, validation, and test sets in the proportions of 0.3, 0.4, and 0.3, respectively. We fine-tuned a model for each task individually (task-specific) and compared this to a model fine-tuned for all tasks available for the dataset (document-level).

Table 4: LLMs performance with zero-shot learning and fine-tuning with two different prompts on the on-treatment cohort.

Model	F1 (95% CI)	Precision (95% CI)	Recall (95% CI)
Mapped Symptom Extraction			
GPT-4o zero-shot	0.896 (0.844–0.932)	0.904 (0.835–0.949)	0.888 (0.817–0.935)
Llama-3.1-8B zero-shot	0.777 (0.712–0.833)	0.788 (0.707–0.857)	0.767 (0.682–0.838)
Llama-3.2-3B zero-shot	0.526 (0.452–0.593)	0.456 (0.377–0.535)	0.621 (0.528–0.707)
Mixtral-8x7B-instruct-v0.1 zero-shot	0.170 (0.100–0.262)	0.351 (0.206–0.515)	0.112 (0.064–0.181)
Gemma-2-9b zero-shot	0.667 (0.595–0.729)	0.618 (0.533–0.697)	0.724 (0.637–0.802)
Qwen3-8B zero-shot	0.136 (0.068–0.235)	0.562 (0.300–0.812)	0.078 (0.036–0.138)
Llama-3.1-8B fine-tuned task-specific	0.720 (0.644–0.782)	0.800 (0.708–0.874)	0.655 (0.565–0.739)
Llama-3.1-8B fine-tuned domain-level	0.760 (0.686–0.820)	0.859 (0.772–0.918)	0.681 (0.591–0.762)
Llama-3.1-8B fine-tuned adjudicated	0.775 (0.705–0.831)	0.898 (0.816–0.949)	0.681 (0.590–0.762)
Qwen3-8B fine-tuned task-specific	0.689 (0.619–0.754)	0.672 (0.586–0.754)	0.707 (0.616–0.784)
Qwen3-8B fine-tuned domain-level	0.444 (0.351–0.538)	0.691 (0.556–0.804)	0.328 (0.245–0.420)
Qwen3-8B fine-tuned adjudicated	0.716 (0.642–0.779)	0.765 (0.680–0.840)	0.672 (0.579–0.754)
Binary Symptom Encoding			
GPT-4o zero-shot	0.847 (0.791–0.891)	0.858 (0.782–0.913)	0.836 (0.760–0.896)
Llama-3.1-8B zero-shot	0.643 (0.560–0.717)	0.787 (0.688–0.866)	0.543 (0.451–0.631)
Llama-3.2-3B zero-shot	0.543 (0.476–0.605)	0.416 (0.352–0.482)	0.784 (0.704–0.850)
Mixtral-8x7B-instruct-v0.1 zero-shot	0.223 (0.154–0.305)	0.272 (0.184–0.380)	0.190 (0.128–0.271)
Gemma-2-9b zero-shot	0.805 (0.747–0.853)	0.713 (0.639–0.785)	0.922 (0.862–0.962)
Qwen3-8B zero-shot	0.426 (0.337–0.515)	0.556 (0.437–0.667)	0.345 (0.263–0.437)
Llama-3.1-8B fine-tuned task-specific	0.775 (0.713–0.832)	0.811 (0.727–0.881)	0.741 (0.651–0.815)
Llama-3.1-8B fine-tuned domain-level	0.749 (0.676–0.806)	0.796 (0.706–0.868)	0.707 (0.615–0.784)
Llama-3.1-8B fine-tuned adjudicated	0.784 (0.718–0.837)	0.821 (0.739–0.885)	0.750 (0.664–0.825)
Qwen3-8B fine-tuned task-specific	0.778 (0.713–0.833)	0.840 (0.758–0.905)	0.724 (0.636–0.798)
Qwen3-8B fine-tuned domain-level	0.422 (0.327–0.517)	0.700 (0.558–0.817)	0.302 (0.224–0.392)
Qwen3-8B fine-tuned adjudicated	0.780 (0.715–0.835)	0.813 (0.729–0.880)	0.750 (0.667–0.826)

Table 5: LLMs performance with zero-shot learning and fine-tuning with Binary symptom encoding prompt on the on-treatment cohort. F1 shows the mean with mean ± standard deviation (SD) in parentheses.

Model	Machine:Human Ratio	F1	Precision	Recall
Llama-3.1-8B	0.4	0.849 (0.048)	0.887 (0.008)	0.818 (0.093)
Llama-3.1-8B	0.6	0.892 (0.035)	0.882 (0.007)	0.904 (0.070)
Llama-3.1-8B	0.8	0.900 (0.042)	0.892 (0.014)	0.913 (0.092)
Llama-3.1-8B	1.0	0.910 (0.024)	0.882 (0.010)	0.941 (0.056)
Llama-3.1-8B	2.0	0.910 (0.019)	0.879 (0.003)	0.943 (0.040)
Llama-3.1-8B	3.0	0.921 (0.023)	0.894 (0.026)	0.952 (0.046)
Qwen3-8B	0.4	0.846 (0.077)	0.915 (0.031)	0.791 (0.117)
Qwen3-8B	0.6	0.964 (0.002)	0.964 (0.009)	0.963 (0.006)
Qwen3-8B	0.8	0.967 (0.005)	0.963 (0.004)	0.970 (0.010)
Qwen3-8B	1.0	0.968 (0.002)	0.964 (0.006)	0.972 (0.005)
Qwen3-8B	2.0	0.968 (0.007)	0.970 (0.005)	0.967 (0.010)
Qwen3-8B	3.0	0.972 (0.003)	0.970 (0.005)	0.975 (0.004)

Table 6: Wilcoxon signed-rank pairwise comparisons with Holm–Bonferroni correction paired per-note F1 scores for pre-diagnosis cohort. Δ is the median paired difference ($F1_A - F1_B$). We use L for short of Llama-3.1-8B-instruct and Q for short of Qwen3-8B. Families define the set of comparisons corrected together (e.g., all prompt comparisons within a cohort).

Family	**A vs. B**	Δ	W	$p_{\mathbf{raw}}$	$p_{\mathbf{adj}}$
Prompt	GPT-4o-zero-shot (Binary) vs (Mapped)	0.005	157.0	0.879	1.000
Prompt	L-zero-shot (Binary) vs (Mapped)	0.143	560.5	0.001	0.140
Prompt	L-task-specific (Binary) vs (Mapped)	0.022	589.0	0.812	1.000
Prompt	L-domain-level (Binary) vs (Mapped)	0.042	216.0	0.103	1.000
Prompt	Q-zero-shot (Binary) vs (Mapped)	0.075	711.0	0.058	1.000
Prompt	Q-task-specific (Binary) vs (Mapped)	-0.011	491.5	0.766	1.000
Prompt	Q-domain-level (Binary) vs (Mapped)	0.058	288.0	0.038	1.000
Model (Binary prompt)	GPT-4o-zero-shot vs L-task-specific	0.023	443.5	0.541	1.000
Model (Mapping prompt)	GPT-4o-zero-shot vs L-task-specific	0.041	779.5	0.548	1.000
Model (Binary prompt)	GPT-4o-zero-shot vs Q-task-specific	0.057	277.0	0.071	1.000
Model (Mapping prompt)	GPT-4o-zero-shot vs Q-task-specific	0.233	900.0	0.001	0.169
Adaptation	L (Disease-level) vs (Task-specific) (Binary)	0.002	149.0	0.977	1.000
Adaptation	Q (Disease-level) vs (Task-specific) (Binary)	0.015	0.0	1.000	1.000

Table 7: Wilcoxon signed-rank pairwise comparisons with Holm–Bonferroni correction paired per-note F1 scores for on-treatment cohort. Δ is the median paired difference ($F1_A - F1_B$). We use L for short of Llama-3.1-8B-instruct and Q for short of Qwen3-8B. Families define the set of comparisons corrected together (e.g., all prompt comparisons within a cohort).

Family	**A vs. B**	Δ	W	$p_{\mathbf{raw}}$	$p_{\mathbf{adj}}$
Prompt	GPT-4o-zero-shot (Binary) vs (Mapped)	-0.049	89	0.791	1.000
Prompt	L-zero-shot (Binary) vs (Mapped)	-0.326	47	< 0.05	0.042
Prompt	L-task-specific (Binary) vs (Mapped)	0.054	134	0.071	1.000
Prompt	L-domain-level (Binary) vs (Mapped)	0.016	205.5	0.795	1.000
Prompt	Q-zero-shot (Binary) vs (Mapped)	0.221	47	0.030	1.000
Prompt	Q-task-specific (Binary) vs (Mapped)	0.021	168	0.425	1.000
Prompt	Q-domain-level (Binary) vs (Mapped)	-0.033	26.5	0.562	1.000
Model (Binary prompt)	GPT-4o-zero-shot vs L-task-specific	0.073	113	0.068	1.000
Model (Mapping prompt)	GPT-4o-zero-shot vs L-task-specific	0.176	54.5	< 0.05	0.122
Model (Binary prompt)	GPT-4o-zero-shot vs Q-task-specific	0.073	113	0.068	1.000
Model (Mapping prompt)	GPT-4o-zero-shot vs Q-task-specific	0.170	108	0.003	0.508
Adaptation	L (Disease-level) vs (Task-specific) (Binary)	-0.054	44	0.040	1.000
Adaptation	Q (Disease-level) vs (Task-specific) (Binary)	-0.344	5.5	< 0.05	0.017

Table 8: Model performance on determining the disease status.

Model	F1 (95% CI)	Precision (95% CI)	Recall (95% CI)
Zero-shot GPT-4	0.662 (0.538–0.769)	0.662 (0.538–0.769)	0.662 (0.538–0.769)
Zero-shot Llama-3-1	0.381 (0.270–0.508)	0.381 (0.254–0.508)	0.381 (0.270–0.508)
Task-specific Llama-3-1	0.683 (0.556–0.794)	0.683 (0.571–0.794)	0.683 (0.571–0.794)
Document-level Llama-3-1	0.810 (0.698–0.889)	0.810 (0.698–0.889)	0.810 (0.698–0.889)

Table 9: Model performance across two tasks, identifying the disease location and the indeterminate nodules from the note.

		Task	
Metric	Model	Indeterminate nodules	Disease location
F1			
	Zero-shot GPT-4	0.468	0.609
	Zero-shot Llama-3-1	0.231	–
	Task-specific Llama-3-1	0.560	0.811
	Document-level Llama-3-1	0.632	0.791
Precision			
	Zero-shot GPT-4	0.688	0.977
	Zero-shot Llama-3-1	0.147	–
	Task-specific Llama-3-1	0.737	0.899
	Document-level Llama-3-1	0.692	0.895
Recall			
	Zero-shot GPT-4	0.355	0.442
	Zero-shot Llama-3-1	0.548	0.000
	Task-specific Llama-3-1	0.452	0.740
	Document-level Llama-3-1	0.581	0.708

We compared the micro F1 score of the inference from four LLMs: GPT-4o zero-shot, Llama-3-1-8B-instruct zero-shot, Llama-3-1-8B-instruct fine-tuned on the radiology dataset for three tasks (document-level), and Llama-3-1-8B-instruct fine-tuned on the radiology dataset for the specific task (task-specific). For determining the disease status, the Llama-3-1-8B-instruct fine-tuned document-level model achieves a micro F1 score of 0.801, outperforming GPT-4 zero-shot with a micro F1 score of 0.662 and the task-specific model with a micro F1 score of 0.683, as shown in Table 8. In both tasks, identifying the indeterminate findings and the disease location, the Llama-3-1-8B-instruct fine-tuned dataset-specific surpasses GPT-4 zero-shot with a micro F1 score of 0.632 and 0.791 for each task, respectively, while GPT-4 zero-shot with a micro F1 score of 0.468 and 0.609, as shown in Table 9.

Radiology report prompt for information retrieval

Disease Status

```
- If there is no evidence of cancer in the scan, please answer '{"Result":
"No evidence of disease"}'
- If cancer was present, but there are no previous scans to compare it to,
please answer '{"Result": "Disease present"}'
- If the cancer has gotten worse or tumors have grown in size compared to prior scans,
```

```
please answer '{"Result": "Progression"}'
- If the cancer has shrunk or there is evidence of response to treatment,
please answer '{"Result": "Treatment response"}'
- If there is cancer on the scan, but it has remained mostly unchanged,
please answer '{"Result": "Disease Stability"}'
- If some parts of the cancer have grown and others have shrunk,
please answer '{"Result": "Mixed"}'
- If cancer is present, but you cannot determine whether it has grown or shrunk,
please answer '{"Result": "Disease present"}'
- If there is concern for cancer, but it is unclear whether cancer is present or not,
please answer '{"Result": "Unclear"}'
- Finally, if the cancer is present, but you believe this is the first time it was seen on
imaging, please answer '{"Result": "Disease present"}'
```

Indeterminate Nodule

```
Indeterminate findings are those that do not clearly indicate malignancy but require
further assessment.

Exclude findings that are obviously cancerous or related
to known cancerous changes in organs.
Also exclude findings that are very likely benign.
Return the organs/compartments containing indeterminate
findings from the following list in the specified JSON format:
    {
      "Indeterminate_Findings": [
        {"Organ_ID": 1, "Organ_Name": "Adrenals"},
        {"Organ_ID": 2, "Organ_Name": "Pancreas"},
        {"Organ_ID": 3, "Organ_Name": "Upper GI lumen"},
        {"Organ_ID": 4, "Organ_Name": "Lower GI lumen"},
        {"Organ_ID": 5, "Organ_Name": "Spleen"},
        {"Organ_ID": 6, "Organ_Name": "Lungs"},
        {"Organ_ID": 7, "Organ_Name": "Extraperitoneal soft tissues"},
        {"Organ_ID": 8, "Organ_Name": "Gall Bladder"},
        {"Organ_ID": 9, "Organ_Name": "Liver"},
        {"Organ_ID": 10, "Organ_Name": "Mesentery"},
        {"Organ_ID": 11, "Organ_Name": "Kidney"},
        {"Organ_ID": 12, "Organ_Name": "Peritoneum"},
        {"Organ_ID": 13, "Organ_Name": "Lymph nodes"},
        {"Organ_ID": 14, "Organ_Name": "Bones"},
        {"Organ_ID": 15, "Organ_Name": "Pelvis"}
      ]
    }

Note: Each of the above represents a SINGLE organ system.
Return only the organs with indeterminate nodules.
Do not report benign findings or other abnormalities.
Return the organs in a pipe-delimited list.

Examples:
If there are indeterminate nodules in the pancreas and lungs, return: {"Indeterminate_Findings":
[{"Organ_ID": 2, "Organ_Name": "Pancreas"},
{"Organ_ID": 6, "Organ_Name": "Lungs"}]}
```

```
If there are no indeterminate nodules, return: {"Indeterminate_Findings": []}
```

Disease Location

```
Cancer presence within an organ can be inferred in cases where the radiologist documents
that findings consistent with or suggestive of malignancy.
Cancer presence should also be documented when a radiology notes a known presence of cancer
in that organ that is confirmed on imaging.
Exclude other abnormal findings, including findings that are indeterminate and could be cancer,
but could also be other benign findings according to the reading radiologist.
Return the organs/compartments where cancer is present from the following
list in the specified JSON format:
    {
      "Cancer_presence": [
        {"Organ_ID": 1, "Organ_Name": "Adrenals"},
        {"Organ_ID": 2, "Organ_Name": "Pancreas"},
        {"Organ_ID": 3, "Organ_Name": "Upper GI tract"},
        {"Organ_ID": 4, "Organ_Name": "Lower GI tract"},
        {"Organ_ID": 5, "Organ_Name": "Spleen"},
        {"Organ_ID": 6, "Organ_Name": "Lungs"},
        {"Organ_ID": 7, "Organ_Name": "Extraperitoneal soft tissues"},
        {"Organ_ID": 8, "Organ_Name": "Gall Bladder"},
        {"Organ_ID": 9, "Organ_Name": "Liver"},
        {"Organ_ID": 10, "Organ_Name": "Mesentery"},
        {"Organ_ID": 11, "Organ_Name": "Kidney"},
        {"Organ_ID": 12, "Organ_Name": "Peritoneum"},
        {"Organ_ID": 13, "Organ_Name": "Lymph nodes"},
        {"Organ_ID": 14, "Organ_Name": "Bones"},
        {"Organ_ID": 15, "Organ_Name": "Pelvis"}
      ]
    }

Note: Each of the above represents a SINGLE organ system. Return only the organs with cancer.
Do not report benign findings or other abnormalities. Return the organs in a pipe-delimited list.

Examples:
If there is cancer in the pancreas and lungs, return:
{"Cancer_presence": [{"Organ_ID": 2, "Organ_Name": "Pancreas"},
{"Organ_ID": 6, "Organ_Name": "Lungs"}]}
If there is no cancer, return: {"Cancer_presence": []}
```