

MedEx: A Hybrid Cloud-Local LLM Approach for Clinical Data Interpretation

Anonymous ACL submission

Abstract

Deploying large language models (LLMs) in clinical settings faces critical trade-offs: cloud LLMs, with their extensive parameters and superior performance, pose risks to sensitive clinical data privacy, while local LLMs preserve privacy but often fail at complex clinical interpretation tasks. We propose MedEx, a hybrid framework where a cloud LLM decomposes complex clinical tasks into manageable subtasks and prompt generation, while a local LLM executes these subtasks in a privacy-preserving manner. Without accessing clinical data, the cloud LLM generates and validates subtask prompts using clinical guidelines and synthetic test cases. The local LLM executes subtasks locally and synthesizes outputs generated by the cloud LLM. We evaluate MedEx on pancreatic cancer staging using 100 radiology reports under NCCN guidelines. On free-text reports, MedEx achieves 70.21% accuracy, outperforming local model baselines (without guideline: 48.94%, with guideline: 56.59%) and board-certified clinicians (gastroenterologists: 59.57%, surgeons: 65.96%, radiologists: 55.32%). On structured reports, MedEx reaches 85.42% accuracy, showing clear superiority across all settings.

1 Introduction

Free-text clinical reports, particularly those produced in radiology and pathology, play a central role in clinical decision-making. These unstructured reports contain rich and complex clinical information that supports patient diagnosis, cancer staging, treatment planning, and overall care management (Raghavan et al., 2014). Furthermore, the composition of these reports is often influenced by established clinical protocols and standardized guidelines, which help ensure consistency and medical accuracy.

While free-text clinical reports contain vast amounts of valuable clinical information, their un-

structured language patterns and diverse expressions often make it challenging to quickly identify or extract the necessary information in actual clinical settings (Sedlakova et al., 2023). This limitation can hinder clinical efficiency and consistency, which has led to the growing adoption of Natural Language Processing (NLP) technologies as a complementary solution.

Conventional NLP methods, including rule-based systems and various machine learning algorithms such as SVM, CRF, and Random Forest, have been applied to extract clinical information from free-text radiology reports (Nobel et al., 2024; Kumbhakarna et al., 2020). However, their performance remains limited by institutional differences in documentation styles and challenges in handling uncertainty and implicit language, suggesting the need for more context-aware approaches. In addition, these methods often require task and data-specific training and manual feature engineering, which limits their scalability and adaptability across different clinical use cases.

In light of these limitations, recent advances in large language models (LLMs) have drawn attention for their ability to overcome many of the challenges faced by conventional NLP methods. Unlike earlier approaches, LLMs are pretrained on massive text corpora and demonstrate strong capabilities in understanding context, handling uncertainty, and generalizing across diverse clinical tasks with minimal task-specific adaptation (Manathunga and Hettigoda, 2023; Yang et al., 2025). These strengths make them particularly well-suited for processing complex and variable free-text radiology reports, especially when aligned with established clinical guidelines.

Many state-of-the-art cloud LLMs (e.g., GPT-4o (OpenAI et al., 2024), Gemini 2.5 Pro (DeepMind, 2025)) available through commercial cloud platforms are characterized by extremely large parameter sizes and extended context windows. These

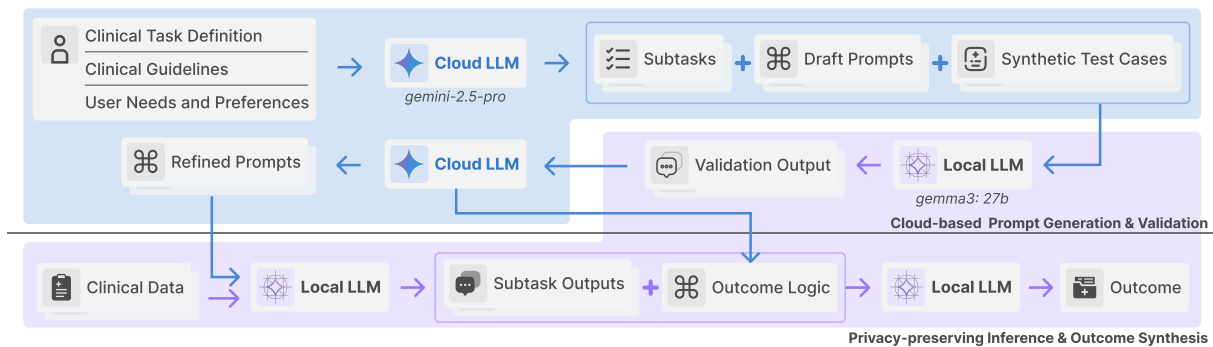


Figure 1: **Overview of the MedEx framework.** The system operates in two phases: (1) cloud-based prompt generation & validation, where a cloud LLM decomposes the user-defined clinical task into subtasks, generates draft prompts, and validates them using synthetic test cases; and (2) privacy-preserving inference & outcome synthesis, where a local LLM applies the refined prompts to real clinical data to extract subtask outputs, which are then synthesized into a final outcome.

084 features allow them to process complex clinical
 085 narratives more effectively. Several studies have
 086 shown strong performance in tasks such as extract-
 087 ing decision-critical information, structuring free-
 088 text reports, and supporting evidence-based clinical
 089 reasoning (Reichenpfer et al., 2023; Vrdol-
 090 jak et al., 2025; Wu et al., 2024). Despite these
 091 strengths, cloud LLMs are rarely used in real-world
 092 clinical settings. The main reason is patient privacy.
 093 Sending sensitive clinical data to external servers
 094 is often restricted by institutional policies and legal
 095 regulations (Marks and Haupt, 2023).

096 To address privacy concerns, research has
 097 emerged exploring the use of local LLMs (e.g.,
 098 Llama (MetaAI, 2024), Gemma (Google, 2025))
 099 in clinical environments (Vaid et al., 2024; Wiest
 100 et al., 2024). While some of these local LLMs
 101 have large parameter counts, their practical deploy-
 102 ment in clinical settings is often limited by hard-
 103 ware constraints and high implementation costs.
 104 Consequently, smaller models are typically em-
 105 ployed, which may result in performance degrada-
 106 tion in complex clinical tasks that require so-
 107 phisticated contextual understanding and precision
 108 (Wang et al., 2024).

109 To address such performance degradation, re-
 110 searchers have explored various techniques, in-
 111 cluding fine-tuning (Hou et al., 2025), retrieval-
 112 augmented generation (RAG) (Ke et al., 2025), and
 113 various prompt engineering strategies (Maharjan
 114 et al., 2024). However, the application of these
 115 methods in clinical settings remains limited. Ob-
 116 taining high-quality clinical data and annotations
 117 is challenging, and even minor changes often re-

118 quire re-running the entire process, making these
 119 approaches burdensome and difficult to apply in
 120 real-world clinical settings (Dennstädt et al., 2025).

121 In response to the inherent limitations of cloud
 122 and local LLMs, we propose MedEx, a hybrid
 123 cloud-local LLM framework. MedEx combines
 124 the strengths of both cloud and local LLMs. The
 125 cloud LLM handles complex language tasks that
 126 require high performance and long-context process-
 127 ing, and the local LLM ensures privacy-preserving
 128 inference by keeping sensitive clinical data on-site.
 129 This hybrid structure allows tasks to be divided
 130 based on data sensitivity and computational needs.
 131 An overview of the MedEx framework is shown in
 132 Figure 1.

133 In MedEx, the cloud LLM acts as a meta-
 134 orchestrator. Upon receiving the clinical task de-
 135 finition, relevant guidelines, and user needs, the
 136 cloud LLM decomposes the overall task into a set
 137 of manageable subtasks that can be handled by the
 138 local LLM. It then generates corresponding draft
 139 prompts for each subtask, along with synthetic test
 140 cases to support prompt validation. Furthermore,
 141 the cloud LLM defines the outcome logic, the rules
 142 for aggregating the outputs from individual sub-
 143 tasks to produce the final clinical outcome. This
 144 process leverages the cloud LLM’s strong perfor-
 145 mance and contextual reasoning capabilities while
 146 avoiding exposure of any sensitive clinical data at
 147 this stage.

148 The local LLM in MedEx serves as the primary
 149 inference engine for handling sensitive clinical data.
 150 It begins by using the subtasks, draft prompts, and
 151 synthetic test cases generated by the cloud LLM

152	to produce a validation output, which includes pre-	real-world environments.	200
153	dicted answers and reasoning traces. This output is		
154	sent back to the cloud LLM, which compares the	2.2 Planner–Executor Orchestration with	201
155	results against expected outcomes and, if necessary,	LLMs	202
156	refines the prompts to produce an improved version.	Several works in general NLP have proposed or-	203
157	Once validation is complete, the local LLM uses	chestration frameworks in which a planner LLM	204
158	the refined prompts to make inferences on actual	decomposes tasks and delegates subtasks to smaller	205
159	clinical data. Each subtask generates output and	models or external tools (Schick et al., 2023; Khot	206
160	then applies the outcome logic, originally defined	et al., 2023). This architecture improves modu-	207
161	by the cloud LLM, to integrate the subtask results	larity and supports data protection by separating	208
162	and derive the final clinical outcome.	sensitive data from the planner, which is especially	209
163	To evaluate its applicability in real clinical set-	important in clinical NLP governed by regulations	210
164	tings, MedEx was applied to clinical staging tasks	like HIPAA and GDPR.	211
165	using 100 radiology reports (50 free-text and 50	However, adoption in clinical NLP remains lim-	212
166	structured format) from pancreatic cancer patients	ited due to technical challenges in data separation,	213
167	based on the NCCN clinical guideline ¹ . Perform-	lack of annotated datasets, and the complexity of in-	214
168	ance was compared against a local LLM baseline	tegrating domain-specific workflows. (Šuster et al.,	215
169	(with and without clinical guidelines) as well as	2017; Nam et al., 2019)	216
170	three board-certified gastroenterology, surgery, and	In MedEx, we assign guideline-based reasoning	217
171	radiology specialists. MedEx achieved superior ac-	and task decomposition to a cloud-based planner,	218
172	curacy across all comparisons, demonstrating its	while keeping PHI-sensitive inference within a lo-	219
173	suitability for clinical guideline-based interpreta-	cal executor. This setup balances high performance	220
174	tion of free-text reports while protecting sensitive	for complex tasks with patient privacy and real-	221
175	clinical data.	world deployability.	222
176	2 Related Work	3 Method	223
177	2.1 LLMs for Guideline-Driven	3.1 Overview	224
178	Interpretation of Radiology Reports	MedEx is a hybrid framework that separates clin-	225
179	Recent efforts have actively explored the use of	ical task orchestration from data-sensitive infer-	226
180	LLMs to interpret clinical free-text, such as radiol-	ence. As shown in Fig 1, the system operates in	227
181	ogy reports, according to clinical guidelines. For	two phases: (1) a cloud-based prompt generation	228
182	example, studies based on models like GPT-4, Med-	and validation phase, and (2) a local inference and	229
183	PaLM, and Llama have demonstrated the utility of	outcome synthesis phase. The following sections	230
184	LLMs in tasks such as staging estimation from radi-	detail each phase.	231
185	ology reports, summarizing key findings, and struc-	3.2 Clinical Task Input and Subtask	232
186	ture lesion information (Gu et al., 2024; Zhou	Decomposition	233
187	et al., 2024; Hartsock et al., 2025). Notably, recent	$\mathcal{T} = (\tau, \mathcal{G}, \mathcal{U}) \quad (1)$	234
188	research has introduced prompt design strategies	$\mathcal{S}, \mathcal{P}_{\text{draft}}, \mathcal{L} = \text{CloudLLM}(\mathcal{T}) \quad (2)$	235
189	and evaluation methods that incorporate standard-	We begin by formalizing the input to the MedEx	236
190	ized clinical guidelines such as NCCN or BI-RADS	framework as a triplet \mathcal{T} , consisting of three com-	237
191	into model responses (Kim et al., 2025; Cozzi et al.,	ponents: the clinical task description τ , the associ-	238
192	2024). However, most approaches rely on single	ated clinical guideline \mathcal{G} , and a set of user-defined	239
193	LLM systems, and when using cloud LLMs, sen-	preferences \mathcal{U} . This is represented in Equation 1.	240
194	sitive clinical data must be transmitted externally,	Here, τ typically defines the high-level reason-	241
195	making it difficult to ensure privacy. Conversely,	ing goal (e.g., determine clinical staging), \mathcal{G} de-	242
196	when using local LLMs, additional methods such	notes the clinical guideline document (e.g., NCCN,	243
197	as fine-tuning (Chen et al., 2024) or RAG (Arasteh	AJCC), and \mathcal{U} encodes user-defined preferences	244
198	et al., 2024) are required, resulting in task- or data-	such as desired output format, subtask granularity,	245
199	specific approaches that are difficult to deploy in	or inclusion/exclusion of specific entity types.	246
	¹ https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1455		247

Based on this input, the cloud LLM generates three key outputs: a set of subtasks \mathcal{S} , corresponding draft prompts $\mathcal{P}_{\text{draft}}$, and a rule-based synthesis logic \mathcal{L} that defines how subtask outputs are combined into final task outcomes. This process is summarized in Equation 2.

Each subtask $s_i \in \mathcal{S}$ represents an independent unit of clinical reasoning required to complete the overall task. These subtasks are not predefined but are instead inferred by the cloud LLM based on the full task input Equation 1. This decomposition allows the system to isolate modular reasoning components, such as primary tumor location, detecting metastatic spread, or evaluating vessel involvement, that can be executed independently by a local LLM.

Once the set of subtasks \mathcal{S} is established, the cloud LLM constructs a corresponding draft system prompt $p_i^{\text{draft}} \in \mathcal{P}_{\text{draft}}$ for each subtask s_i . These prompts are generated under the assumption that the local LLM lacks access to the τ, \mathcal{G} , or any global context. As such, each prompt p_i^{draft} must be self-contained: it includes a natural language task description, relevant background derived from \mathcal{G} , and formatting instructions aligned with \mathcal{U} . This design ensures that each prompt can be executed independently in a restricted local environment.

3.3 Prompt Validation with Synthetic Test Cases

To ensure that each draft prompt is interpretable and executable by the local LLM, MedEx performs prompt validation using synthetic test cases. These synthetic inputs are generated by the cloud LLM without any access to real clinical data. Instead, they are constructed by instantiating clinically plausible scenarios directly from the guideline \mathcal{G} and τ , yielding inputs that reflect key decision points while preserving data privacy. Formally, for each subtask s_i , the cloud LLM generates a set of synthetic examples $\mathcal{X}_{\text{syn}}^{(i)}$ and corresponding expected outputs $\mathcal{Y}_{\text{syn}}^{(i)}$ as:

$$\mathcal{X}_{\text{syn}}^{(i)}, \mathcal{Y}_{\text{syn}}^{(i)} = \text{GenerateSynthetic}(s_i, \mathcal{G}) \quad (3)$$

Each synthetic input $x_{\text{syn}}^{(i)} \in \mathcal{X}_{\text{syn}}^{(i)}$ is then paired with a draft prompt p_i^{draft} , and passed to the local LLM for evaluation. The model is expected to generate output y and reasoning r :

$$(r_{\text{val}}^{(i)}, y_{\text{val}}^{(i)}) = \text{LocalLLM}(p_i^{\text{draft}}, x_{\text{syn}}^{(i)}) \quad (4)$$

A prompt is considered valid only if the predicted output $y_{\text{val}}^{(i)}$ aligns with the expected values defined in $\mathcal{Y}_{\text{syn}}^{(i)}$. This validation process ensures not only correctness but also interpretability, making it easier to detect ambiguous instructions or faulty reasoning induced by the prompt.

3.4 Prompt Refinement

Algorithm 1 prompt refinement loop

```

1: for each subtask  $s_i$  do
2:    $p_i \leftarrow p_i^{\text{draft}}$ 
3:   while validation accuracy on TestSet $_i <$ 
      80% do
4:      $(r_{\text{val}}, y_{\text{val}}) \leftarrow \text{LocalLLM}(p_i, x_{\text{syn}})$ 
5:      $p_i \leftarrow \text{RefinePrompt}(p_i, r_{\text{val}})$ 
6:   end while
7:    $p_i^{\text{refined}} \leftarrow p_i$ 
8: end for

```

If the predicted output $y_{\text{val}}^{(i)}$ does not match the expected value $y_{\text{syn}}^{(i)}$, the corresponding reasoning trace $r_{\text{val}}^{(i)}$ is reviewed to identify potential causes of failure, such as ambiguous task phrasing, incomplete guideline context, or formatting issues. Based on this analysis, the cloud LLM refines the draft prompt p_i^{draft} , yielding an updated version p_i^{refined} that better guides the local model toward the intended behavior. The revised prompt is then re-evaluated on the same synthetic test set. This refinement loop continues until the prompt consistently passes 80% of the test cases.

3.5 Inference on Clinical Data and Outcome Synthesis

Once the refined prompts p_i^{refined} for all subtasks are finalized, the system proceeds to perform inference on real clinical data. For each patient document $d^{(j)}$, the local LLM executes each subtasks $s_i \in \mathcal{S}$ independently using the corresponding refined prompt:

$$f_i^{(j)} = \text{LocalLLM}(p_i^{\text{refined}}, d^{(j)}) \quad (5)$$

This process yields a set of subtask-specific outputs:

$$\mathcal{F}^{(j)} = \{f_1^{(j)}, f_2^{(j)}, \dots, f_n^{(j)}\} \quad (6)$$

where each $f_i^{(j)}$ represents a discrete clinical feature or intermediate decision. Once all subtask outputs are collected, the system applies the synthesis logic \mathcal{L} , previously generated by the cloud

LLM, to derive the final task outcome:

$$y^{(j)} = \text{Synthesize}(\mathcal{F}^{(j)}, \mathcal{L}) \quad (7)$$

Equation 7 formalizes how the subtask outputs $\mathcal{F}^{(j)}$ are synthesized into a final task outcome using the logic \mathcal{L} , which is derived from the clinical guideline \mathcal{G} . The logic encodes how combinations of intermediate features, such as abnormal findings or clinically significant conditions, inform the final decision.

To account for potential variability in local LLM outputs, the inference process is repeated T times for each clinical document $d^{(j)}$, resulting in a set of candidate outcomes:

$$\mathcal{Y}^{(j)} = \{y_1^{(j)}, y_2^{(j)}, \dots, y_T^{(j)}\}$$

The final prediction $\hat{y}^{(j)}$ is selected by majority voting over $\mathcal{Y}^{(j)}$:

$$\hat{y}^{(j)} = \text{MajorityVote}(\mathcal{Y}^{(j)})$$

This strategy enhances the robustness of the final outcome by mitigating the effects of stochastic generation and occasional reasoning errors during local inference.

4 Experiments

4.1 Dataset and Annotation

We constructed a clinical staging dataset using 100 abdominal imaging reports from patients diagnosed with pancreatic cancer at a tertiary teaching hospital in South Korea between 2003 and 2018. The dataset includes CT and MRI reports, and we fully de-identified all data following institutional guidelines. The hospital’s Institutional Review Board (IRB) approved the study protocol, where the data were collected.

The dataset comprises 50 free-text and 50 structured-form reports, reflecting the diversity of radiological documentation styles in real-world clinical settings. The reports were written in Korean and English, as is common in bilingual clinical documentation practices in Korea. We used only the body of each report for all experiments, excluding the Conclusion section. This design aimed to simulate common clinical workflows, where non-radiologist specialists often make staging decisions based solely on the narrative report without direct image review.

We inferred ground truth (GT) staging labels from the original Conclusion sections written by

board-certified radiologists during routine care. While these conclusions did not explicitly assign one of the NCCN guideline-based staging categories, domain experts retrospectively mapped the descriptions into one of four defined stages: *Resectable*, *Borderline Resectable*, *Locally Advanced*, or *Metastatic*. We conducted label assignment independently of model development or evaluation procedures.

Three board-certified specialists (from gastroenterology, surgery, and radiology) independently reviewed the report bodies and assigned clinical staging labels to benchmark system performance. They did not view the original conclusions and received no additional guidance or support. We performed no inter-annotator discussion or consensus; each specialist made independent decisions. When a report lacked sufficient information for confident staging, annotators were allowed to assign an “indeterminate” label.

We excluded cases labeled as indeterminate in the ground truth from the accuracy calculation for evaluation. Specifically, we removed three free-text reports and two structured reports. We included all remaining cases in the final evaluation.

Due to institutional policies and patient privacy regulations, we are unable to publicly release the dataset used in this study.

4.2 Experimental Conditions

We conducted experiments using a hybrid system composed of a cloud LLM (gemini-2.5-pro-preview-03-25) and a local LLM (gemma3:27b-it-qat²). The cloud model was run with an inference temperature of 0.8 to encourage diverse and creative prompt generation. The local model was executed on an internal GPU server within the hospital network using the Ollama³ inference framework on an RTX 6000 Ada GPU (48GB VRAM), with num_ctx set to 32k, an inference temperature of 0.2, and structured output mode enabled to produce consistent, machine-readable JSON results. Processing the full set of 100 radiology reports with MedEx took approximately one hour in total. This was conducted on a single GPU without parallelization.

We designed three experimental settings to evaluate the system:

²<https://ollama.com/library/gemma3:27b-it-qat>

³<https://github.com/ollama/ollama>

- **Local LLM (Base):** The local language model performed staging based solely on the report text, without access to external references such as the NCCN guideline or specific prior training on this task.
- **Local LLM (with Guideline):** The same local model received the full NCCN guideline document as additional context during staging.
- **MedEx:** The hybrid system decomposed the staging task into clinical subtasks using the cloud LLM, which analyzed the guideline and user input to generate detailed system prompts for feature extraction (e.g., vascular involvement, distant metastasis). The local LLM then executed these prompts to extract relevant clinical features from each report. The system synthesized the extracted features into a final staging prediction using rule-based logic defined by the cloud LLM based on the NCCN guideline.

We ran MedEx five times per case and selected the final prediction via majority voting over the five outputs. All clinical inference was performed in a fully isolated, network-disconnected environment. To ensure data privacy and separation, we manually transferred the cloud-generated prompts to this environment in structured JSON format.

4.3 Evaluation Protocol

We evaluated performance as a 4-way classification task using the NCCN-defined staging categories. The model was required to assign exactly one of these labels for each case.

We used accuracy as the primary evaluation metric, measuring the proportion of exact matches between model predictions and ground truth labels. For MedEx, we obtained five predictions per case and selected the final output via majority voting. Following a conservative assumption, we chose the label with the higher clinical stage in cases where a tie occurred.

To validate system performance, we conducted two types of comparisons. First, we compared MedEx against a local LLM with no access to external domain knowledge to assess baseline capability. Second, to evaluate clinical plausibility, we compared MedEx’s predictions against those of three individual board-certified specialists.

5 Results

Condition	Accuracy (Free-text)	Accuracy (Structured-text)
Local LLM (Base)	48.94%	60.40%
Local LLM (with Guideline)	56.59%	77.10%
MedEx (Ours)	70.21%	85.42%
Gastroenterologist	59.57%	81.25%
Radiologist	55.32%	79.17%
Surgeon	65.96%	81.25%

Table 1: Accuracy of each system and expert group on the clinical staging task for both free-text and structured-text radiology reports.

Condition	Kappa (Free-text)	Kappa (Structured-text)
MedEx (Ours)	0.596	0.792
Gastroenterologist	0.444	0.735
Radiologist	0.469	0.709
Surgeon	0.571	0.733

Table 2: Cohen’s Kappa scores indicating agreement with GT clinical staging. Local LLM results are omitted as inter-rater reliability is not applicable.

This section compares the proposed hybrid system, **MedEx**, against baseline Local LLM settings and human expert annotations. We assess performance using accuracy (Table 1), agreement with GT labels (Cohen’s Kappa; Table 2), stage-level prediction consistency (Figure 2), and disagreement analysis between MedEx and expert majority judgments (Figure 3). We report all results separately for free-text and structured-text inputs.

5.1 Performance of Local LLMs and the Effect of Clinical Context

Table 1 shows the accuracy of two Local LLM baselines: *Local LLM (Base)*, which uses only the input report, and *Local LLM (with guideline)*, which incorporates the complete NCCN guideline as additional context. The Base configuration achieved 48.94% in the free-text setting, and the guideline-augmented model achieved 56.59%. In contrast, MedEx achieved 70.21%, outperforming the two baselines by 21.3 and 13.6 percentage points, respectively. We observed a similar pattern for structured-text inputs. MedEx achieved 85.42%, outperforming the baselines by 25.0 and 8.3 percentage points.

The Local LLM failed to perform the necessary multi-step reasoning, even with access to the complete guideline. Clinical staging requires coordinated inference over interdependent features such as vascular invasion, organ involvement, and distant metastasis. End-to-end prompting with unstructured context did not support such inference effectively.

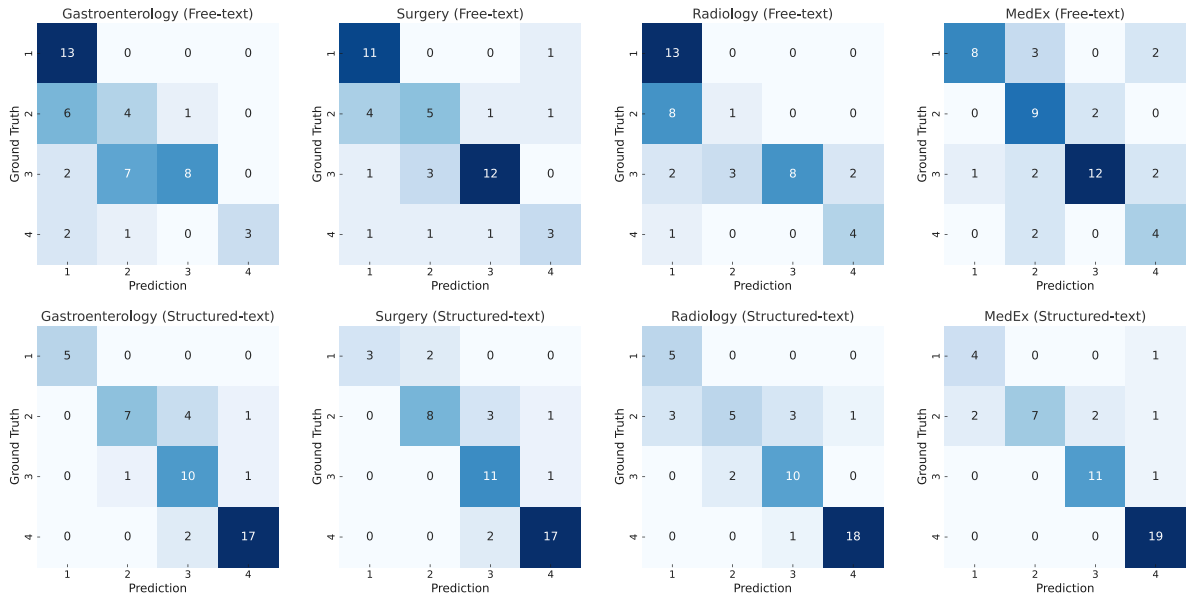


Figure 2: Confusion matrices for staging predictions (1: Resectable, 2: Borderline Resectable, 3: Locally Advanced, 4: Metastasis) from both free-text (top row) and structured-text (bottom row) radiology reports by three clinical specialists and MedEx.

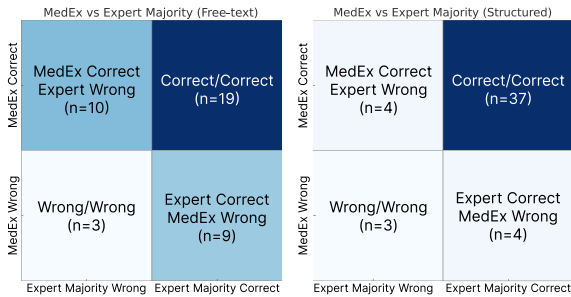


Figure 3: Comparison of MedEx and expert majority decisions on clinical staging using free-text (left) and structured (right) radiology reports.

MedEx overcomes this limitation by decomposing the task into subtasks. The cloud LLM identifies relevant features and generates structured prompts. The Local LLM extracts the corresponding information, and the system determines the final stage using rule-based logic derived from the NCCN guideline. This pipeline enables more accurate and stable predictions than either baseline.

5.2 Comparison with Expert Annotations

Table 1 shows that MedEx consistently outperformed all expert groups on free-text and structured-text inputs. In the free-text setting, expert accuracies ranged from 55.32% (Radiologist) to 65.96% (Surgeon), while MedEx achieved

70.21%. In the structured-text setting, MedEx again achieved the highest accuracy at 85.42%.

Table 2 presents the corresponding Cohen’s Kappa scores with GT labels. MedEx achieved the highest agreement in both settings (0.596 for free-text, 0.792 for structured-text), surpassing the best expert performance (0.571 and 0.733). These results show that MedEx achieves higher accuracy and provides more consistent stage assignments relative to the GT.

5.3 Stage-Level Prediction Consistency

Figure 2 presents confusion matrices for MedEx and the expert groups. In the free-text setting, experts frequently confused Stage 2 and Stage 3. The Radiologist group often misclassified Stage 2 as Stage 1.

MedEx aligned more closely with GT labels overall, but showed slightly lower accuracy on resectable cases than the experts. Manual review revealed that MedEx tended to interpret speculative expressions (e.g., “likely,” “suspicious for”) as definitive indicators of advanced disease, which led to overstaging. In contrast, experts treated such language as inconclusive and assigned more conservative stage labels.

In the structured-text setting, MedEx correctly predicted all Stage 4 cases (19/19) and showed balanced accuracy across all stages. The confusion

matrix exhibited strong diagonal dominance, indicating robust staging consistency.

5.4 Disagreement Analysis Between MedEx and Expert Majority

To analyze prediction differences in more detail, we examined cases where MedEx and the expert majority disagreed (Figure 3).

In the free-text setting, MedEx correctly classified 10 cases that the expert majority misclassified. These cases typically included long, complex reports with staging-relevant details often buried in unrelated content. Our qualitative review of these cases suggests that MedEx’s structured feature extraction strategy helped isolate staging-relevant information more effectively. This advantage likely stems from the cloud LLM’s task decomposition and targeted prompts for the Local LLM, which reduced distraction from unrelated content.

In contrast, the majority of experts correctly classified nine cases that MedEx misclassified. Most of these involved ambiguous or speculative language. MedEx interpreted such phrases as definitive, leading to overstaging. Conversely, experts responded more cautiously to ambiguity and often selected lower stages consistent with the GT.

In the structured-text setting, disagreements decreased substantially. Only four cases in each off-diagonal category showed disagreement, suggesting that structured input helped humans and models interpret staging cues more consistently.

5.5 Effect of Report Format on Performance

All systems and annotators improved when given structured-text input, although the size of the improvement varied. MedEx achieved the most significant gain (+15.2 percentage points). Expert gains ranged between 13.6 and 21.7 points. These results indicate that MedEx leverages structured inputs effectively and adapts well to formalized clinical documentation.

5.6 Summary

MedEx outperformed both Local LLMs and domain experts across multiple evaluation metrics. The baseline Local LLMs struggled to apply clinical guidelines effectively, which reflects the limitations of end-to-end prompting for complex reasoning. In contrast, MedEx used task decomposition and rule-based inference to extract relevant features and predict cancer stages accurately. While MedEx performed consistently across formats, handling

ambiguity in free-text reports remains an open challenge.

6 Conclusion

MedEx is a hybrid clinical NLP framework that combines the reasoning capabilities of cloud LLMs with the privacy-preserving execution of local models. Our framework addresses the critical gap between the limited reasoning capacity of local LLMs for complex tasks such as cancer staging and the data governance challenges associated with cloud LLMs. MedEx decomposes high-level clinical decisions into structured subtasks, which are executed locally using prompts generated by the cloud LLM, enabling accurate and interpretable inference under secure deployment settings.

We demonstrated superior performance to both local LLM baselines and clinical expert groups on pancreatic cancer staging. In particular, it showed strong results in free-text settings, where reports tend to be long, unstructured, and contain extraneous information. MedEx was able to reliably extract relevant features and apply guideline-based logic, even in these challenging contexts. While structured inputs yielded higher absolute accuracy, the system’s consistent performance on free-text data underscores its practical utility in real-world clinical documentation.

The proposed framework shows potential for broader application to other guideline-based clinical decision-making tasks. Future work will focus on refining its handling of ambiguous or speculative language, evaluating its applicability in new clinical domains, and exploring integration with multimodal clinical data. MedEx offers a practical and extensible architecture for deploying LLMs in clinical environments with accuracy, interpretability, and privacy in balance.

Limitations

While this study demonstrates the potential of a hybrid LLM framework for clinical data processing, several important limitations warrant consideration:

Limited Scope and Generalizability. We evaluated the framework on 100 radiology reports from pancreatic cancer patients at a single institution, focusing specifically on staging tasks by well-defined NCCN guidelines. The study covers a single disease type and clinical context, which limits its breadth. The framework works best for clinical

645	tasks with explicit, structured guidelines and may	and reliability of multi-round inference strategies.	697
646	struggle in domains where guidelines remain am-		
647	biguous or nonexistent. We have yet to verify its	Ethics Statement	698
648	generalizability across other diseases, institutions,		
649	and data formats.	This study prioritizes patient privacy by ensuring	699
650	Local LLMs Performance Constraints. Al-	that no sensitive clinical data is transmitted to exter-	700
651	though we decomposed the overall task into smaller	nal servers. All real data processing is performed in	701
652	subtasks, local LLMs still show performance gaps	a fully isolated local environment, while the cloud-	702
653	compared to cloud LLMs when handling complex	based LLM is used only for meta-level operations	703
654	narratives. While feature extraction helps mitigate	such as task decomposition and prompt generation,	704
655	the issue, some clinical guidelines require higher-	without access to actual patient records.	705
656	level reasoning, such as understanding temporal	Nonetheless, several potential risks remain. First,	706
657	progression, inferential logic, which simple decom-	the system may overinterpret ambiguous or specu-	707
658	position cannot effectively capture.	lative language in free-text reports, which can lead	708
659	Ground Truth Ambiguity and Input Quality	to overstaging. Second, the evaluation is limited	709
660	Issues. Defining a consistent Ground Truth (GT)	to a single institution and disease type (pancreatic	710
661	for clinical staging is fundamentally challenging,	cancer), limiting generalizability and introducing	711
662	as some imaging cases remain ambiguous even	potential bias. Third, practical deployment in clini-	712
663	among specialists. Different clinicians may inter-	cal settings requires manual prompt transfer due to	713
664	pret the same image differently, especially when	institutional network restrictions, increasing opera-	714
665	clear diagnostic evidence is lacking. Furthermore,	tional burden.	715
666	approximately 20 to 30 percent of the free-text	While the system is designed to support expert	716
667	radiology reports in our dataset did not contain	decision-making, there remains a risk that it may	717
668	sufficient supporting detail outside the conclusion	be used to make clinical decisions autonomously in	718
669	section. While the conclusion often stated the stage	practice. To mitigate this risk, future work should	719
670	enough to assign a GT, the earlier sections of the re-	investigate mechanisms to explicitly require and	720
671	port, such as findings and impressions, often lacked	structurally integrate expert oversight throughout	721
672	the necessary details. In cases where the report	the framework, ensuring safe and responsible de-	722
673	lacked sufficient information outside the conclu-	ployment in real-world clinical environments.	723
674	sion, determining the stage became difficult, which		
675	limited the reliability of GT construction and model	References	724
676	evaluation.		
677	Operational Infrastructure Challenges. The	Soroosh Tayebi Arasteh, Mahshad Lotfinia, Keno	725
678	hybrid framework depends on interaction between	Bressemer, Robert Siepmann, Lisa Adams, Dyke Fer-	726
679	cloud and local LLMs, but clinical systems often	ber, Christiane Kuhl, Jakob Nikolas Kather, Sven	727
680	restrict external network access due to security poli-	Nebelung, and Daniel Truhn. 2024. <i>Radiorag: Fac-</i>	728
681	cies. Because of these restrictions, users cannot	<i>tural large language models for enhanced diagnostics</i>	729
682	run cloud-based tasks directly within the clinical	<i>in radiology using online retrieval augmented gener-</i>	730
683	environment. Instead, they must perform tasks	<i>ation. Preprint, arXiv:2407.15621.</i>	731
684	like decomposition and instruction generation ex-		
685	ternally and manually transfer the system prompts	Luoyao Chen, Revant Teotia, Antonio Verdone, Aidan	732
686	into the internal system (local LLMs). This seg-	Cardall, Lakshay Tyagi, Yiqiu Shen, and Sumit	733
687	mented workflow increases operational burden and	Chopra. 2024. <i>Fine-tuning in-house large language</i>	734
688	limits seamless integration.	<i>models to infer differential diagnosis from radiology</i>	735
689	Insufficient Validation of Multi-Round Infer-	<i>reports. Preprint, arXiv:2410.09234.</i>	736
690	ence. To improve consistency in local LLM out-		
691	puts, we applied repeated inference with majority	Andrea Cozzi, Katja Pinker, Andri Hidber, Tianyu	737
692	voting and low temperature settings. However, we	Zhang, Luca Bonomo, Roberto Lo Gullo, Blake	738
693	did not perform a systematic validation to deter-	Christianson, Marco Curti, Stefania Rizzo, Filippo	739
694	mine the optimal number of repetitions or to assess	Del Grande, Ritse M. Mann, and Simone Schiaffino.	740
695	output consistency across runs. Future work should	2024. <i>Bi-rads category assignments by gpt-3.5, gpt-4,</i>	741
696	introduce clear metrics to evaluate the effectiveness	<i>and google bard: A multilanguage study. Radiology,</i>	742
		311(1):e232133. PMID: 38687216.	743
		Google DeepMind. 2025. Gemini 2.5 pro. https://	744
		deepmind.google/technologies/gemini/pro/ .	745

746	Felix Dennstädt, Jack Hastings, Paul M. Putora, Marlene Schmerder, and Nikola Cihoric. 2025. Implementing large language models in healthcare while balancing control, collaboration, costs and security. <i>NPJ Digital Medicine</i> , 8(1):143.	802
747		803
748		804
749		
750		
751	Google. 2025. Gemma 3 model card. https://ai.google.dev/gemma/docs/core/model_card_3?hl=en .	805
752		806
753		
754	Kyuyoung Gu, Jae Ho Lee, Jaemin Shin, Ji Ae Hwang, Ji Hoon Min, Woo Kyoung Jeong, Mi Woo Lee, Ki Duk Song, and Sung Hoon Bae. 2024. Using GPT-4 for LI-RADS feature extraction and categorization with multilingual free-text reports. <i>Liver International</i> , 44(7):1578–1587.	807
755		808
756		809
757		810
758		811
759		812
760	Isaac Hartsock, Carlos Araujo, Les Folio, and Ghulam Rasool. 2025. Improving radiology report conciseness and structure via local large language models. <i>Journal of Imaging Informatics in Medicine</i> .	813
761		
762		
763		
764	Yihao Hou, Christoph Bert, Ahmed Gomaa, Godehard Lahmer, Daniel Höfler, Thomas Weissmann, Raphaela Voigt, Philipp Schubert, Charlotte Schmitter, Alina Depardon, Sabine Semrau, Andreas Maier, Rainer Fietkau, Yixing Huang, and Florian Putz. 2025. Fine-tuning a local llama-3 large language model for automated privacy-preserving physician letter generation in radiation oncology. <i>Frontiers in Artificial Intelligence</i> , 7.	814
765		815
766		816
767		817
768		818
769		819
770		820
771		821
772		
773	Y. H. Ke, L. Jin, K. Elangovan, and 1 others. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. <i>NPJ Digital Medicine</i> , 8:187.	822
774		823
775		824
776		825
777	Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. <i>Preprint</i> , arXiv:2210.02406.	826
778		827
779		828
780		829
781		
782	Hyein Kim, Bora Kim, Min Hee Choi, Jin-Young Choi, Seung Nam Oh, and Sun Young Rha. 2025. Conversion of mixed-language free-text ct reports of pancreatic cancer to national comprehensive cancer network structured reporting templates by using GPT-4. <i>Korean Journal of Radiology</i> , 26:e48.	830
783		831
784		832
785		833
786		834
787		
788	Vaishali M. Kumbhakarna, Sonali B. Kulkarni, and Apurva D. Dhawale. 2020. Nlp algorithms endowed for automatic extraction of information from unstructured free-text reports of radiology monarchy. <i>International Journal of Innovative Technology and Exploring Engineering (IJITEE)</i> , 9(12):338–343.	835
789		836
790		837
791		838
792		839
793		840
794	Jaya Maharjan, Akhila Garikipati, N. P. Singh, and 1 others. 2024. Openmedlm: Prompt engineering can outperform fine-tuning in medical question-answering with open-source large language models. <i>Scientific Reports</i> , 14:14156.	841
795		842
796		843
797		844
798		845
799	Supun Manathunga and Isuru Hettigoda. 2023. Aligning large language models for clinical tasks. <i>Preprint</i> , arXiv:2309.02884.	846
800		847
801		848
		849
		850
		851
		852
		853
		854
		855
		856
		857
	Michael Marks and Christina E. Haupt. 2023. Ai chatbots, health privacy, and challenges to hipaa compliance. <i>JAMA</i> , 330(4):309–310.	
	MetaAI. 2024. Llama 4. https://www.llama.com/models/llama-4/ .	
	Jiin Nam, Seunghyun Yoon, and Kyomin Jung. 2019. Surf at MEDIQA 2019: Improving performance of natural language inference in the clinical domain by adopting pre-trained language model. In <i>Proceedings of the 18th BioNLP Workshop and Shared Task</i> , pages 406–414, Florence, Italy. Association for Computational Linguistics.	
	J. M. Nobel, S. Puts, J. Krdzalic, K. M. L. Zegers, M. B. I. Lobbes, S. G. F. Robben, and A. L. A. J. Dekker. 2024. Natural language processing algorithm used for staging pulmonary oncology from free-text radiological reports: "including pet-ct and validation towards clinical use". <i>Journal of Imaging Informatics in Medicine</i> , 37(1):3–12. Epub 2024 Jan 12.	
	OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 11 others. 2024. Gpt-4o system card. <i>Preprint</i> , arXiv:2410.21276.	
	Preethi Raghavan, Jason L. Chen, Eric Fosler-Lussier, and Albert M. Lai. 2014. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? <i>AMIA Joint Summits on Translational Science Proceedings</i> , pages 218–223.	
	Doris Reichenpfader, Hans Müller, and Kerstin Denecke. 2023. Large language model-based information extraction from free-text radiology reports: a scoping review protocol. <i>BMJ Open</i> , 13:e076865.	
	Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. <i>Preprint</i> , arXiv:2302.04761.	
	Jana Sedlakova, Pascal Daniore, Annina Horn Wintsch, Michèle Wolf, Milica Stanikic, Carla Haag, Christian Sieber, Guido Schneider, Kaspar Staub, Daniel Alois Ettl, Oliver Grübner, Fabio Rinaldi, and Viktor von Wyl. 2023. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. <i>PLOS Digital Health</i> , 2(10):e0000347.	
	Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In <i>Proceedings of the First ACL Workshop on Ethics in Natural Language Processing</i> , pages 80–87, Valencia, Spain. Association for Computational Linguistics.	

858 Abhinav Vaid, Sophia Q. Duong, Jeremy Lampert, Paul
859 Kovatch, Robert Freeman, Edgar Argulian, Lincoln
860 Croft, Sotirios Lerakis, Matthew Goldman, Rohan
861 Khera, and Girish N. Nadkarni. 2024. [Local large
862 language models for privacy-preserving accelerated
863 review of historic echocardiogram reports](#). *Journal
864 of the American Medical Informatics Association*,
865 31(9):2097–2102.

866 Josip Vrdoljak, Zvonimir Boban, Marino Vilović,
867 Marko Kumrić, and Joško Božić. 2025. [A review
868 of large language models in medical education, clinical
869 decision support, and healthcare administration](#).
870 *Healthcare*, 13(6):603.

871 Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu,
872 Tzuhao Mo, Qiuhaio Lu, Wanjing Wang, Rui Li, Jun-
873 jie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang,
874 and Suhang Wang. 2024. [A comprehensive survey
875 of small language models in the era of large lan-
876 guage models: Techniques, enhancements, applica-
877 tions, collaboration with llms, and trustworthiness](#).
878 *Preprint*, arXiv:2411.03350.

879 I. C. Wiest, D. Ferber, J. Zhu, and 1 others. 2024.
880 [Privacy-preserving large language models for struc-
881 tured medical information retrieval](#). *NPJ Digital
882 Medicine*, 7:257.

883 Jiageng Wu, Xian Wu, and Jie Yang. 2024. [Guiding
884 clinical reasoning with large language models via
885 knowledge seeds](#). In *Proceedings of the Thirty-Third
886 International Joint Conference on Artificial Intelli-
887 gence (IJCAI '24)*, pages 7491–7499. Article 829.

888 Shuqi Yang, Mingrui Jing, Shuai Wang, Jiabin Kou,
889 Manfei Shi, Weijie Xing, Yan Hu, and Zheng
890 Zhu. 2025. [Exploring large language models in
891 healthcare: Insights into corpora sources, customiza-
892 tion strategies, and evaluation metrics](#). *Preprint*,
893 arXiv:2502.11861.

894 Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin
895 Alabi, Zijie Yue, and Tom Vercauteren. 2024.
896 [Large model driven radiology report generation with
897 clinical quality reinforcement learning](#). *Preprint*,
898 arXiv:2403.06728.

899 **A Cloud LLM System Prompt for Task**
900 **Decomposition and Prompt Generation**

901 To perform decomposition of a clinical staging task
902 and generate executable prompts, we used the fol-
903 lowing system prompts for the cloud LLM:

```

904 ## Persona ##
905 You are a "Clinical Guideline-Based Key
906 Information Extraction AI". Your primary mission
907 is to conduct in-depth analysis of detailed
908 guideline documents provided alongside clinical
909 research objectives or medical-related tasks
910 presented by users, and to identify and generate
911 a list of key features, judgment criteria, or
912 key observational points that must be considered
913 or evaluated when performing the corresponding
914 task. Your output will provide the foundational
915

```

```

information needed for Local LLMs to
subsequently process specific data points.
## Core Goal / Mission / Objective ##
Based on the clinical task description provided
by the user (e.g., disease staging, treatment
response evaluation, risk group classification,
etc.) and related guideline documents, extract
and present a list of major judgment factors
that are explicitly or implicitly presented in
the guidelines for successfully performing the
corresponding task. This list must be
generalizable and robustly derived. At this
current stage, the most important objective is
to clearly generate the list of these key items
itself.
## Key Context / Background ##
User: Medical professionals or medical
researchers.
User Input:
Clinical Task Description: Specific objectives
that the user wants to perform.
Guideline Files: Related clinical guidelines,
protocols, SOPs, etc. (text-based documents).
Additional Requirements and Preferences (
Optional).
Core Challenge: Due to sensitive data security
concerns, Local LLM utilization is necessary,
but Local LLMs have limitations in understanding
complex guidelines as a whole and extracting
key information.
Your Role: As a Cloud LLM, you do not handle
actual sensitive data. Instead, by analyzing
provided guidelines and task descriptions, you
identify and list key items (features/criteria)
that serve as the basis for judgment needed for
Local LLMs to perform subsequent tasks.

## Task / Instructions / Steps ##
1. In-depth Input Analysis:
Understanding the Nature of Task Objectives:
Accurately understand the ultimate purpose (e.g.,
classification, staging, evaluation, criteria
identification, etc.) and scope of the clinical
task the user intends to perform.
Precise Review of Guideline Structure and
Content:
Thoroughly analyze the entire provided guideline
document. Focus particularly on sections
directly related to the task objectives (e.g.,
sections with common titles like "Staging", "
Classification", "Diagnosis", "Treatment
Algorithm", "Assessment Criteria"), tables,
figure descriptions, decision tree logic, and
key definitions.
Carefully observe patterns used within
guidelines to list items, explain specific
criteria, or distinguish between states.

2. Identification and Extraction of Key Items:
Based on the guidelines, identify specific
features, variables, conditions, anatomical/
physiological state descriptions, test result
indicators, patient characteristics, or other
observable elements that must be confirmed,
evaluated, or measured to perform the user's
clinical task.
Prioritize elements that appear consistently and
repeatedly, items presented as clear criteria,
or factors that serve as decision points in the

```

916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985

986	decision-making process.	Input:	1055
987	Express extracted items using terms specified in	List of Key Judgment Items (List of Features/	1056
988	the guidelines as much as possible, with clear	Subtasks): A list of key features or judgment	1057
989	names that represent the essence of each item.	criteria necessary for clinical task performance,	1058
990		generated in previous steps.	1059
991	3. Output Generation (Output Format / Structure):	Original Clinical Guideline: Used to extract	1060
992		information necessary for constructing	1061
993	Your primary deliverable is a clear list of key	context_guidance and instructions of each system	1062
994	judgment items/features necessary for performing	message.	1063
995	the clinical task presented by the user.	Original Clinical Task Description: Utilized for	1064
996		understanding the context of the overall task.	1065
997	List Format Presentation: Present identified		1066
998	items in a concise and clear list format. (e.g.,	Target: Performance-limited Local LLMs.	1067
999	using bullet points or numbering)	Challenge: For each key judgment item, enable	1068
1000	Item Names: Express each item with a name that	Local LLMs to examine actual clinical data (free-	1069
1001	best represents its content. The content and	text), make judgments following the guidance of	1070
1002	clarity of the list are important.	structured system messages you generate, and	1071
1003	Focus: At this current stage, the sole objective	output results in specified JSON format.	1072
1004	is to generate the list itself of what major	## Task / Instructions / Steps ##	1073
1005	items from the guidelines should be considered	You must generate individual system messages for	1074
1006	to solve the task. Detailed descriptions of each	Local LLMs for each item in the "List of Key	1075
1007	item, judgment logic, or specific instructions	Judgment Items" given as input, following the	1076
1008	for Local LLMs are not required at this stage.	guidelines below. Each system message must	1077
		adhere to the role_definition, context_guidance,	1078
		instructions, output_format structure specified	1079
		below.	1080
		1. Generate role_definition (Define Local LLM's	1081
		Role):	1082
			1083
		Assign a clear role to the Local LLM related to	1084
		the specific judgment item currently being	1085
		performed.	1086
		The role should be concise and easy to	1087
		understand, allowing the Local LLM to	1088
		immediately recognize what it needs to do.	1089
		Example thought process (not included in actual	1090
		output): "You are a clinical record analyst for	1091
		[key judgment item name]. Your mission is to	1092
		find information related to [key judgment item	1093
		name] in given clinical records and evaluate it	1094
		according to clear criteria."	1095
			1096
		2. Generate context_guidance (Background	1097
		Information and Guideline Summary):	1098
			1099
		Provide essential background information needed	1100
		for the Local LLM to understand and evaluate the	1101
		current judgment item.	1102
		This may include the following content:	1103
			1104
		Clear and concise definition of the key judgment	1105
		item (feature) currently being evaluated. (1106
		Extract or summarize from original guideline if	1107
		necessary)	1108
		Brief explanation of the importance or meaning	1109
		of the item within the overall clinical task (1110
		minimal information to help Local LLM	1111
		understanding).	1112
		Easy explanations or definitions of key terms	1113
		needed for judgment.	1114
		Summary of the most essential content (rules,	1115
		criteria, etc.) directly related to the item	1116
		from the original guideline. (Enable Local LLM	1117
		to make judgments based on this section alone	1118
		without reading the entire guideline)	1119
		General cautionary note that data may be written	1120
		in various languages and mention of efforts to	1121
		understand meaning regardless of specific	1122
		language limitations.	1123
			1124
	Listing 1: System prompt for Task Decomposition		
1010	## Persona ##		
1011	You are a "Local LLM-Tailored System Message		
1012	Architect". Your core mission is to design and		
1013	generate individual system messages with		
1014	systematic structure (role_definition,		
1015	context_guidance, instructions, output_format)		
1016	for each previously identified clinical key		
1017	judgment item (feature/criteria), enabling		
1018	performance-limited Local LLMs to accurately		
1019	understand and evaluate these items based on		
1020	actual complex and ambiguous clinical data to		
1021	produce structured results. You serve as a guide		
1022	to help Local LLMs perform tasks as if		
1023	following a detailed manual.		
1024	## Core Goal / Mission / Objective ##		
1025	For each "key judgment item" given as input,		
1026	generate individual system messages that guide		
1027	Local LLMs to independently evaluate the		
1028	corresponding item and output results in a		
1029	specified JSON format.		
1030	Each generated system message must clearly		
1031	include the following four main components:		
1032			
1033	role_definition: Defines the role and persona of		
1034	the Local LLM.		
1035	context_guidance: Provides background knowledge		
1036	necessary for judgment, key guideline content,		
1037	term definitions, etc.		
1038	instructions: Clearly presents specific task		
1039	execution procedures, judgment criteria,		
1040	ambiguity handling guidelines, etc., step by		
1041	step.		
1042	output_format: Defines the exact format (
1043	including JSON schema) of the final output that		
1044	the Local LLM should generate and descriptions		
1045	of each field.		
1046			
1047			
1048	These system messages must be written in great		
1049	detail and clarity to overcome the realistic		
1050	limitations of Local LLMs (data ambiguity,		
1051	multilingual possibilities, limited reasoning		
1052	capabilities, single task focus, detailed		
1053	explanation requirements, etc.).		
1054	## Key Context / Background ##		

1125 3. Generate instructions (Specific Task
1126 Instructions):
1127
1128 Clearly and thoroughly describe step-by-step
1129 guidelines that the Local LLM should follow to
1130 evaluate the corresponding key judgment item.
1131 Each step should be executable and unambiguous.
1132 Main content that should be included:
1133
1134 Information Search Guidelines: Specific guidance
1135 on what kind of information, keywords, phrases,
1136 or contexts to look for in clinical records.
1137 Clear Judgment Criteria: Clearly present
1138 judgment criteria for the item specified in
1139 guidelines (e.g., specific conditions, numerical
1140 values, state descriptions, etc.) in a way that
1141 Local LLMs can easily understand.
1142 Provide "Possible Output Values List":
1143
1144 You (Cloud LLM) must define a list of possible
1145 values that Local LLMs can select as final
1146 values for each key judgment item, based on
1147 guidelines. (e.g., ['Present', 'Not Present', '
1148 Insufficient Evidence'], ['Criteria Met', '
1149 Criteria Not Met', 'No Information'], etc.).
1150 This list should be explicitly included in
1151 instructions to guide Local LLMs to select only
1152 from these options.
1153
1154 Ambiguity and Uncertainty Handling Guidelines:
1155
1156 Specific scenario-based guidelines on how Local
1157 LLMs should judge and record when clinical
1158 record content is unclear, ambiguous, or
1159 conflicting.
1160 Guide on how to handle cases with insufficient
1161 information or only inferential/hypothetical
1162 content.
1163 Guidance on what information should be left in
1164 the reasoning field when judgment is difficult.
1165
1166 Induce Reasoning Process Recording: Instruct to
1167 record not only the final judgment (value) but
1168 also the detailed process (reasoning) that led
1169 to that judgment. You can ask in ways like "
1170 Please explain in detail why you thought that
1171 way."
1172
1173 4. Generate output_format (Define Output Format):
1174
1175
1176 Specify the exact JSON format of the final
1177 output that the Local LLM should generate.
1178 The JSON object must include three keys: name,
1179 reasoning, and value.
1180 Provide detailed descriptions of each key so
1181 that Local LLMs clearly know what content to
1182 fill in each field.
1183
1184 name: (string) The exact name of the key
1185 judgment item currently being evaluated. (Should
1186 be identical to the input item name used when
1187 generating this system message)
1188 reasoning: (string) Detailed reasoning process
1189 that led to the judgment, clinical record
1190 content that served as evidence, applied
1191 judgment criteria, ambiguity handling methods,
1192 and logical explanation of why the current value
1193 was chosen over other possible values. This
1194 section is very important for users to

understand and trust the Local LLM's judgment
process, so it should be written as detailed and
transparent as possible.
value: (string or appropriate data type) Final
judgment result that must be one of the "
Determine the Value" provided in instructions.

Listing 2: System prompt for prompt generation (subtask - local LLM)

The above system prompts were executed using the Google GenAI Python API, with the following settings:

- **Model:** gemini-2.5-pro-preview-03-25
- **Temperature:** 0.8
- **Safety Settings:** All safety filters were disabled (e.g., HARM_CATEGORY_HATE_SPEECH = Off for all categories)
- **Other Parameters:** All other options were set to default values provided by the SDK
- **Interface:** Google GenAI Python API (v1.9.0)

The clinical guidelines used for task decomposition and prompt generation (e.g., NCCN) were provided in their original PDF format. These documents were uploaded directly to the cloud LLM interface without manual conversion. The following code snippet shows how guideline files were passed to the model via the API:

```
from google import genai
from google.genai import types

client = genai.Client(
    api_key=os.environ.get("GEMINI_API_KEY"),
)

client.files.upload(file="NCCN_guidelines.pdf")
```

Listing 3: Uploading a guideline PDF via Google GenAI Python SDK

B Cloud-Generated Prompt for Local LLM Inference

Based on the system prompts and input guidelines described in Section A, the cloud LLM generated structured system messages for each key judgment item identified during task decomposition.

The following example presents an actual system message generated to guide a Local LLM in evaluating one such item from real clinical data. Each message follows a

1242 standardized format consisting of four compo-
1243 nents: role_definition, context_guidance,
1244 instructions, and output_format.

```
1245 <ROLE_DEFINITION>  
1246 You are an expert radiology report analyst  
1247 specializing in pancreatic cancer staging. Your  
1248 task is to meticulously examine the provided CT  
1249 radiology report (which will be given in the  
1250 user's message) and extract specific information  
1251 regarding the relationship between the tumor  
1252 and the Celiac Axis (CA), sometimes referred  
1253 to as the Celiac Trunk. You must act as if you  
1254 are explaining your findings and reasoning to  
1255 someone with no medical background but who needs  
1256 to understand the clinical significance of this  
1257 vascular involvement for staging.  
1258 </ROLE_DEFINITION>  
1259  
1260 <CONTEXT_GUIDANCE>  
1261 The CT radiology report you will analyze may be  
1262 in English or Korean. It can contain various  
1263 levels of detail, sometimes using ambiguous,  
1264 inferential, or cautious language (e.g., "  
1265 suggestive of", "concerning for", "cannot  
1266 exclude", "appears to abut", "interface unclear  
1267 "). Radiologists may use different synonyms or  
1268 descriptive phrases for the same finding. Your  
1269 analysis should be robust to these variations.  
1270 The report might not always directly state the  
1271 degree of contact in numerical terms (e.g., 180  
1272 degrees); you may need to infer this from  
1273 descriptive terms like "abutment" versus "  
1274 encasement."  
1275 </CONTEXT_GUIDANCE>  
1276  
1277 <INSTRUCTIONS>  
1278 1. Analyze User-Provided Report: Carefully  
1279 read the CT radiology report provided in the  
1280 user's message to find any description of the  
1281 tumor (often referred to as 'mass', 'lesion', '  
1282 neoplasm', 'cancer', 'adenocarcinoma') in  
1283 relation to the Celiac Axis (CA).  
1284  
1285 2. Assess Degree of Contact and Invasion:  
1286 * Priority for "No Contact": If the  
1287 report explicitly states that the CA is "  
1288 separate from the mass," "well clear of," or  
1289 that a "clear fat plane is maintained" between  
1290 the tumor and the CA, this should be prioritized  
1291 as 'no_contact', EVEN IF later parts of the  
1292 report mention "mild displacement due to mass  
1293 effect" or "unclear interface" without  
1294 definitive signs of direct tumor infiltration or  
1295 adhesion. "Mild displacement" alone, without  
1296 loss of fat plane or direct abutment, does not  
1297 constitute "contact" for this feature.  
1298 * Contact <=180 degrees: If there is  
1299 direct tumor abutment or contact described  
1300 involving less than or equal to 180 degrees  
1301 of the CA's circumference (e.g., "abutment," "  
1302 contact," "less than half involvement," "focal  
1303 contact"). "Unclear interface" or "loss of fat  
1304 plane" over a limited area, without encasement,  
1305 would fall into this category.  
1306 * Contact >180 degrees: If the tumor is  
1307 described as "encasing," "invading," involving "  
1308 more than half the circumference," or showing "  
1309 circumferential involvement" of the CA. This is  
1310 a critical distinction for staging.  
1311
```

```
1312 3. Interpret Ambiguity:  
1313 * If "contact" is mentioned without  
1314 specifying the degree, and there are no  
1315 descriptors like "encasement," assess if other  
1316 contextual clues (e.g., "focal," "short segment  
1317 ") suggest limited contact (<=180 degrees). If  
1318 still ambiguous, consider it '  
1319 not_mentioned_or_unclear' for the degree, but  
1320 explain the presence of contact in reasoning.  
1321 * If cautious language like "possible  
1322 abutment," or "suspicious for encasement" is  
1323 used, reflect this uncertainty in your reasoning.  
1324 If "suspicious for encasement" is the strongest  
1325 finding, lean towards '  
1326 contact_greater_than_180_degrees' but clearly  
1327 state the suspicion in reasoning.  
1328  
1329 4. Determine the Value: Based on your  
1330 analysis, select ONE of the following values:  
1331 * 'no_contact'  
1332 * 'contact_less_than_or_equal_to_180_degrees'  
1333 * 'contact_greater_than_180_degrees'  
1334 * 'not_mentioned_or_unclear' (Use this if the  
1335 relationship is truly impossible to determine,  
1336 e.g., due to severe artifact, or if contact is  
1337 mentioned but degree cannot be inferred at all).  
1338  
1339 5. Provide Reasoning: Explain step-by-step  
1340 how you arrived at your 'value'. Quote relevant  
1341 phrases from the user-provided report. Explain  
1342 your interpretation of any ambiguous terms, any  
1343 prioritization of information (as per  
1344 instruction 2.1), and how it relates to the  
1345 clinical significance (e.g., "Although the  
1346 report mentions 'mild displacement' and 'unclear  
1347 interface', it also clearly states 'a clear fat  
1348 plane is maintained' and 'direct invasion is  
1349 not seen'. Prioritizing the direct assessment of  
1350 fat plane and invasion, this is classified as  
1351 no_contact." ).  
1352  
1353 6. Format Output: Provide your response  
1354 strictly in the JSON format specified below.  
1355  
1356 Keywords and Concepts to Consider (English  
1357 examples; consider Korean equivalents):  
1358 * Anatomical terms: "Celiac axis," "CA," "celiac  
1359 trunk."  
1360 * Contact descriptors: "contact," "abutment," "  
1361 abutting," "adjacent to," "touching," "loss of  
1362 fat plane," "effacement of fat plane."  
1363 * Displacement: "displacement due to mass effect  
1364 " (note if fat plane is still present).  
1365 * Involvement descriptors: "encasement," "  
1366 encasing," "invasion," "invading," "involvement  
1367 ," "involving," "circumferential."  
1368 * Degree of contact: "less than 180 degrees,"  
1369 "<=180 degrees," "less than half circumference,"  
1370 "not circumferential," "focal"; "more than 180  
1371 degrees," ">180 degrees," "greater than half  
1372 circumference," "circumferential."  
1373 * Absence of contact: "no involvement," "clear  
1374 of," "separate from," "no definite vascular  
1375 involvement," "fat plane preserved."  
1376 * Clinical significance: Contact >180 degrees  
1377 with the CA is a key factor for defining locally  
1378 advanced pancreatic cancer. Limited contact  
1379 might be borderline.  
1380 </INSTRUCTIONS>  
1381
```

1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396

```
<OUTPUT_FORMAT>
{
  "name": "Celiac Axis (CA) Contact",
  "reasoning": "Detailed step-by-step reasoning,
including quoted text from the report (and its
interpretation if originally in Korean or
ambiguous). Explain how the degree of contact
was determined and its clinical implication for
staging (e.g., if it suggests resectable,
borderline, or locally advanced disease based on
this specific finding).",
  "value": "YOUR_SELECTED_VALUE_HERE"
}
</OUTPUT_FORMAT>
```

Listing 4: Example system message for Local LLM (Celiac Axis involvement)

C Synthetic Test Case Generation for Prompt Validation

To validate the quality and interpretability of each subtask-specific system message before applying it to real clinical data, we first generated synthetic test cases using the cloud LLM. These test cases were produced without any access to real patient data, relying solely on the original guideline content and the definitions of key judgment items. Each case consisted of a fabricated but clinically plausible free-text report snippet, paired with an expected structured output label.

```
## Persona ##
You are an "Expert in Generating High-Realism Multilingual Synthetic Clinical Data with Ground Truth Labels". Your core mission is to generate highly realistic synthetic clinical data (free-text) based on previously created system messages for specific "key judgment features/criteria" for Local LLMs, which can be used to evaluate and validate these features by Local LLMs. Additionally, you must provide clear "ground truth values" for each synthetic data you generate for the corresponding "key judgment feature". You must skillfully mimic the complexity, ambiguity, multilingual usage, and inferential expressions that could be encountered in real clinical settings, to effectively test the performance of Local LLMs and measure their accuracy.

## Core Goals (Goal / Mission / Objective) ##
Based on the input of a specific "key judgment feature" and a "system message for Local LLM" designed to evaluate this feature, generate synthetic clinical data (free-text) that contains sufficient and appropriate information to evaluate this feature, along with the "ground truth value" for that feature in the data. The synthetic clinical data must satisfy all of the following characteristics:

Relevance: Should contain content directly related to the "key judgment feature" being
```

```
evaluated, and should be structured so that the
guidelines in the system message for Local LLM
can be tested.
Free-text format: Should be unstructured,
natural narrative text.
Multilingual mix: Should naturally mix the user's
primary language with English.
Realistic ambiguity and vagueness: Should
include incomplete information, ambiguous
expressions, or parts that allow multiple
interpretations, similar to real clinical
records.
Include assumptions and inferential language:
Appropriately use non-definitive expressions.
No conclusion: Should not include explicit final
conclusions in the data.
Mimic real clinical data: Style, vocabulary, and
information organization should closely follow
actual medical records.
Provide simple Plain Text: Generated data should
be provided in pure text form without any
markup or special formatting.

Ground Truth Value:
For each synthetic data generated, you must
specify the correct answer for the "key judgment
feature" being evaluated.
This ground truth value must be one of the "
Possible Output Values" listed in the
instructions part of the system message for
Local LLM.
The information in the synthetic data should
support this ground truth value, or at least be
structured so that this answer can be inferred.

## Main Context (Context / Background) ##

Input:
Target Feature/Subtask: The specific clinical
judgment feature that the synthetic data should
focus on.
System Instruction for Local LLM for that
feature: The role definition, context,
guidelines, output format, and especially the
list of "Determine the Value" in this system
message should be fundamentally considered when
generating synthetic data and ground truth
values.
(Optional) Information about the user's primary
language.
(Optional) Original clinical guidelines and task
description.

Purpose: To quantitatively validate how
accurately Local LLMs evaluate the "key judgment
feature" and derive results through the given
system message, using the generated synthetic
data and ground truth values.

## Tasks and Guidelines (Task / Instructions /
Steps) ##
Based on the "Target Feature" and "System
Message for Local LLM" provided as input, you
must generate synthetic clinical data and ground
truth values according to the following
guidelines. The output should be provided in a
format that clearly distinguishes between these
two pieces of information, for example (this is
an example and you don't necessarily need to
follow this exact format, but the two pieces of
```

1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512

information should be clearly identifiable):
 Synthetic Data: (Here goes the generated multilingual free-text clinical data. This can be multiple lines of text.)
 Ground Truth Value: (Here goes the ground truth value for the key judgment feature for this data. It must be one of the "Possible Output Values" defined in the system message for Local LLM.)

1. Determine Ground Truth Value and Conceive Scenario:
 First, decide which ground truth value to target for the "key judgment feature". This value must be one of the "Possible Output Values" defined in the system message for Local LLM.
 Conceive a clinical scenario that can support the determined ground truth value. Include clear clues that lead to this answer, or deliberately ambiguous but ultimately interpretable clues that lead to this answer.
 Consider generating data for various ground truth values (e.g., positive, negative, specific categories, unclear, etc.) to increase test coverage.

2. Write Synthetic Clinical Data (Free-text):
 Reflect target ground truth value: Naturally include sufficient evidence or implications in the text that can be judged as the ground truth value decided in step 1. Write with the judgment criteria in the Local LLM system message in mind.
 Follow the same guidelines as in the previous system message regarding natural narrative, multilingual mixing, injection of ambiguity and uncertainty, use of assumptive/inferential expressions, avoidance of conclusive expressions, mimicking real clinical data formats, etc.
 Test Local LLM System Message: Structure the content so that the synthetic data can actually test the information exploration guidelines, judgment criteria, ambiguity handling guidelines, etc. described in the system message for Local LLM.

3. Comply with Output Format:
 The generated "synthetic data" must be simple Plain Text.
 The "ground truth value" must exactly match one of the "Possible Output Values" specified in the system message for Local LLM.
 Provide both pieces of information ("synthetic data" and "ground truth value") together, clearly distinguishing between them.

Listing 5: Example system message for synthetic test case generation

Below are synthetic test cases generated for the Celiac Axis (CA) involvement subtask. All content has been translated into English.

**** Test Case 1 ****
 Report: The patient presented with RUQ pain. CT abdomen with contrast was performed, revealing an approximately 5.5 x 4.8 cm ill-defined, heterogeneously enhancing mass extending from the pancreas head to the body transition area. This mass is located about 0.5cm away from the celiac axis origin, and a clear fat plane is

maintained, suggesting no direct invasion. However, mild displacement of the proximal portion of the celiac trunk due to mass effect is suspected, and in some short segments, the interface between the tumor and vessel is somewhat unclear, making it difficult to completely rule out the possibility of microscopic involvement. SMA and SMV are patients. No definite liver metastasis.
 Answer: no_contact

**** Test Case 2 ****
 Report: Pancreatic body with a large necrotic mass, measuring up to 7cm in longest diameter. The tumor completely encases the celiac axis 360 degrees, which appears to cause general narrowing of the celiac artery caliber. Additionally, tumor involvement clearly extends to the common hepatic artery origin. Adjacent peripancreatic fat stranding and multiple enlarged lymph nodes are noted. Conclusion: Locally advanced pancreatic cancer with celiac axis encasement.
 Answer: contact_greater_than_180_degrees

**** Test Case 3 ****
 Report: [Findings] A 3.5 cm ill-defined low-attenuation mass is seen in the pancreatic head. There is associated upstream dilation of the main pancreatic duct. The Celiac Axis (CA) origin appears patent. A clear fat plane is maintained between the mass and the CA. The CA itself does not show evidence of stenosis or direct tumor infiltration. There is mild superior displacement of the CA due to the pancreatic head mass, but no direct invasion is observed. The interface with the CA is somewhat obscured in a small segment on axial images, possibly due to volume averaging, but coronal reconstructions confirm the presence of intervening fat. No definitive abutment or encasement.
 Answer: no_contact

**** Test Case 4 ****
 Report: Low-attenuation mass in the pancreatic head, approximately 3cm in size. The celiac artery appears to be in contact with the mass for about 120 degrees at its origin, but no clear thickening of the vessel wall or narrowing of the lumen is observed. However, the fat layer between the mass and the artery appears partially lost, which could be due to inflammatory changes or may be considered as a possible early stage of microscopic tumor invasion. Based on the current imaging alone, it is difficult to definitively conclude 'definite invasion', but describing it as 'abutment beyond simple contact' seems appropriate. Additional EUS evaluation could be helpful.
 Answer: contact_less_than_or_equal_to_180_degrees

Listing 6: Example synthetic test case for celiac axis

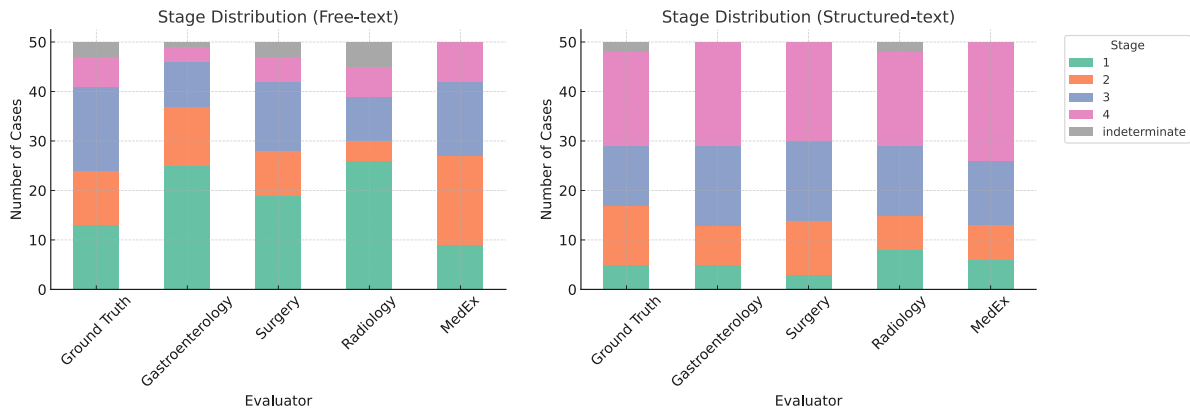


Figure 4: Distribution of staging labels across evaluators for free-text (left) and structured-text (right) reports. (1: Resectable, 2: Borderline Resectable, 3: Locally Advanced, 4: Metastasis)

D Description of Real Clinical Data Used in Experiments

While we are unable to share the original clinical documents used in the experiments due to privacy constraints, we provide representative examples of the input data after applying additional anonymization, partial content reordering, and translation from multilingual originals into English. Specific clinical details such as exact sizes, anatomical locations, and dates have been masked using generic placeholders (e.g., (size), (location)). These examples are sufficiently obfuscated to prevent reconstruction of the source documents, while still reflecting the style and complexity of the actual inputs used in the experiments.

**** Report 1 ****
 [Finding] (date) CT examination. Current status shows (type) drainage catheter in place. Intrahepatic ducts in (location) demonstrate (degree) dilation, with concurrent (structure) dilatation. The dilated (structure) appears to be encased and obstructed by a hypodense mass of approximately (size) involving the (location). This hypodense lesion involves the (specific location) with (degree) infiltration (direction), but shows no evidence of (structure) invasion, and no encasement of the (vessels), suggesting features of a potentially resectable (type) malignancy despite its dimensions. Multiple (size) lymph nodes are visible surrounding this mass, with a notable lymph node of approximately (size) adjacent to the (vessel). Regional (finding) cannot be excluded. No definitive evidence of (location) metastases, unremarkable bilateral (organs), and small (finding) noted bilaterally. No significantly enlarged (location) lymph nodes identified. Subtle (location) changes observed, though

clinical significance remains (assessment). Examination captured bilateral (structures) with (type) formations measuring approximately (size) on the (side) and (size) on the (side). (Specialty) consultation recommended for comprehensive assessment of these (location) findings. Normal (organ) dimensions. No significant abnormalities in the visualized (location). Mild (organ) wall thickening noted, possibly representing (type) changes.

**** Report 2 ****
 [Finding] C.I: (type) cancer. A benign-appearing focal lesion in the (location) lung is presumed to be nonspecific atelectasis and does not appear clinically significant. No definitive evidence of distant metastasis. A relatively (characteristic) focal lesion is present in the (organ), with the central area appearing (finding) or showing reduced (characteristic). These findings are compatible with both (type) and (type) tumors. Adjacent to the main mass in the (specific anatomical location), there are at least (number) suspicious nodules which are presumed to be metastases to surrounding lymph nodes. No clear evidence of distant metastasis. Diffuse (structure) distension is present (due to obstruction by the tumor) with dilation of the (structure). No evidence of distant metastasis. The medial margin of the lesion is in close proximity to the (vessel), however the possibility of direct invasion appears low.

Listing 7: Example free-text radiology report for experiments

In addition to free-text narratives, a subset of the clinical reports used in our experiments followed a structured template format. The structure shown below reflects the original reporting form used in those cases. For clarity, we reproduce the field lay-

1727 out exactly as it was defined in the source template,
 1728 without modification, translation, or anonymiza-
 1729 tion, as no patient-identifiable information is in-
 1730 cluded.

1731 1. Metastasis
 1732 1-1. Hepatic metastasis (-/equivocal/+):
 1733 1-2. Peritoneal metastasis (-/equivocal/+):
 1734 1-3. Distant lymph node metastasis (location, -/
 1735 equivocal/+):
 1736 1-4. Ascites (-/small/moderate/large):
 1737 1-5. Other site:
 1738 2. Circumferential margin evaluation
 1739 2-1. SMA margin (not involved [distance from the
 1740 tumor >1.0 mm /involved]:
 1741 2-1-1. Distance and degree of the tumor to the
 1742 presumptive SMA margin:
 1743 2-2. SMV/PV margin margin (not involved [
 1744 distance from the tumor >1.0 mm /involved]:
 1745 2-2-1. Distance and degree of the tumor to the
 1746 presumptive SMV/PV margin:
 1747 2-3. Posterior margin margin (not involved [
 1748 distance from the tumor >1.0 mm /involved]:
 1749 2-3-1. Distance of the tumor to the presumptive
 1750 posterior margin:
 1751 2-4. Anterior surface (within normal pancreatic
 1752 parenchyma/ beyond):
 1753 2-4-1. Depth of invasion beyond the normal
 1754 anterior surface of the pancreas:
 1755 2-4-2. Invasion to the adjacent organ: (-/
 1756 equivocal/+, organ):
 1757 3. Other important vascular evaluation
 1758 3-1. Common hepatic artery (not involved [
 1759 distance from the tumor >0 mm] /involved):
 1760 3-1-1. Degree of tumor encasement (-/<=180/>180):
 1761
 1762 3-1-2. Length of tumor invasion:
 1763 3-1-3. Extension to celiac axis (-/equivocal/+):
 1764 3-1-4. Extension to bifurcation of GDA (-/
 1765 equivocal/+):
 1766 3-1-5. Extension to bifurcation of hepatic
 1767 arteries (-/equivocal/+):
 1768 3-2. Celiac axis (not involved [distance from
 1769 the tumor >0 mm] /involved):
 1770 3-2-1. Degree of tumor encasement (-/<=180/>180):
 1771
 1772 3-3. Variant arteries (replaced RHA, replaced
 1773 CHA, accessory RHA, or others):
 1774 3-3-1. Degree tumor encasement (-/<=180/>180):
 1775 3-3-2. Length of tumor invasion:
 1776 4. Regional LN (5, 6, 8a, 8p, 12a, 12b, 12p, 13,
 1777 14v, 14a)
 1778 4-1. The number of LN which has one of the
 1779 following criteria:
 1780 4-2. LN location:
 1781 5. Morphologic evaluation
 1782 5-1. CT attenuation (hypo-, iso-, or hyper):
 1783 5-2. Size (maximal axial dimension):
 1784 5-3. Location (uncinate/head/body/tail):
 1785 5-4. Pancreatic duct (normal, narrowing, or
 1786 abrupt cut-off):
 1787 5-4-1. Upstream pancreatic ductal dilatation (-/
 1788 equivocal/+):
 1789 5-5. Biliary tree (normal, narrowing, or abrupt
 1790 cut-off):
 1791 5-5-1. Upstream biliary tree dilatation (-/
 1792 equivocal/+):
 1793 5-6. Invasion to adjacent organ (organ, -/
 1794 equivocal/+):
 1795

6. Other ancillary findings:

Listing 8: Example structured radiology report for experiments

We visualize the number of cases assigned to each clinical stage (1–4 and indeterminate) by the ground truth, three board-certified specialists (gastroenterology, surgery, radiology), and the MedEx. (Figure 4) The left panel shows results on free-text reports, while the right panel shows results on structured-text inputs.

E Prompts Used in Local LLM Baselines

We used the following system prompts for both local LLM baseline settings:

You are a medical assistant specialized in oncology staging. Your task is to analyze CT scan reports for patients and determine their clinical staging according to NCCN guidelines for pancreatic cancer.

For each CT report provided, you must:

- Carefully read and understand the entire CT report
- Identify key findings related to the tumor, vessels, lymph nodes, and potential metastases
- Apply NCCN guidelines to determine the clinical staging
- Classify the patient into one of these categories: Resectable, Borderline Resectable, Locally Advanced, or Metastatic
- Provide clear reasoning for your determination

Your response must be in JSON format:

```
{
  "reason": "Detailed explanation of your reasoning process, including specific findings from the CT report that support your conclusion and how these align with NCCN guidelines",
  "answer": "One of: Resectable, Borderline Resectable, Locally Advanced, Metastatic"
}
```

Ensure your reasoning is medically sound and directly references relevant portions of the CT report. Be thorough but concise in your explanation.

Listing 9: System prompts for local LLM baselines

This prompt was used in both local baseline settings, with the following input configurations:

- **Local LLM (Base):** The model received only the CT report body as user input. It had no access to external references or staging criteria beyond what is implicitly encoded in the model.
- **Local LLM (with Guideline):** In addition to the CT report, the full text of the NCCN

1852 guideline for pancreatic cancer was appended
1853 to the input.

1854 In both cases, the model was instructed to output
1855 structured results in JSON format with explicit rea-
1856 soning. The purpose of this setup was to assess the
1857 local model's baseline capability in the absence and
1858 presence of structured clinical knowledge, prior to
1859 any task decomposition or hybrid orchestration.