A Multi-Agent Open-Source LLM for Structured Cancer Registry Information Extraction from Pathology and Medical Reports

Riham Jeeballah[1], **Adhari Al Zaabi**[2], **Habiba El Keraby**[1], **and Abdulrahman AAl Abdulsalam**[1,*]

Corresponding author: a.aalabdulsalam@squ.edu.om

[1]Sultan Qaboos University, Department of Computer Science, Oman

[2]Sultan Qaboos University, Department of Clinical and Human Anatomy, Oman

Abstract

Extracting structured cancer registry information from pathology and medical reports is challenging due to heterogeneous reporting styles and implicit clinical reasoning. We propose a modular multi-agent framework that decomposes registry abstraction into semantic chunking, retrieval, field-specific extraction, validation, evaluation, and aggregation stages.

The dataset includes 818 annotated cancer cases from Sultan Qaboos University Hospital. Evaluation in this study focuses on breast (n=454) and colorectal (n=174) reports across grade, morphology, TNM staging, and laterality extraction tasks. The framework is compared against prompt-based LLaMA 3.3 baselines using accuracy and weighted/macro F1-score metrics.

The proposed framework improved performance in context-dependent tasks, particularly grade extraction, where weighted F1-score increased from 0.71 to 0.78 for breast cancer and from 0.56 to 0.67 for colorectal cancer. Improvements were also observed for colorectal laterality extraction. For other extraction tasks, particularly highly structured tasks such as TNM staging and morphology extraction, the multi-agent framework achieved performance comparable to direct prompting.

Although the baseline achieved slightly higher average weighted F1-scores overall, the proposed framework provides improved modularity, traceability, and pipeline-level interpretability through explicit intermediate reasoning stages, supporting error analysis and future clinician-guided refinement.

1 Introduction

Cancer registries are essential for epidemiological surveillance, clinical research, and healthcare planning. However, key registry variables such as grade, morphology, TNM staging, and laterality are often embedded in free-text pathology reports, making abstraction labor-intensive, difficult to scale, and prone to inter-annotator variability.

Extracting structured information from clinical narratives remains challenging due to heterogeneous documentation styles, implicit terminology, and contextual ambiguity. While transformer models and large language models (LLMs) improved clinical text understanding and structured output generation, single-pass prompting approaches remain vulnerable to hallucination, inconsistency, and weak grounding when extracting multiple interdependent variables.

Recent advances in agentic LLM systems offer an alternative by decomposing complex tasks into modular reasoning stages. The ReAct paradigm demonstrated that combining reasoning and action improves performance and interpretability (Yao et al., 2022). Building on this idea, biomedical agentic systems incorporated planning, orchestration, and tool use (Abbasian et al., 2024; Li et al., 2025), while surveys and benchmark studies characterized these paradigms in biomedicine (Lin et al., 2024; Xu and Sankar, 2025; Yang et al., 2025; Radi et al., 2025; Gorenshtein et al., 2025). In oncology, agentic frameworks have been explored for decision support, diagnostic reasoning, and treatment planning (Ferber et al., 2025; Chen et al., 2025; Çağatay Umut Öğdü et al., 2025; Kuerbanjiang et al., 2025; Inoue et al., 2025; Zhao et al., 2025a). In parallel, LLM-based approaches have also been investigated for structured extraction from pathology and oncology reports (Chow et al., 2025; Gupta et al., 2025; Hart and Bergamaschi, 2026).

Despite this progress, existing approaches often focus on isolated tasks, single cancer types, or direct prompting strategies with limited visibility into intermediate reasoning stages and failure sources.

In this work, we propose an open-source multi-agent framework for structured cancer registry extraction from pathology and medical reports,

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 531–551

focusing on breast and colorectal cancer. Implemented as a sequential LangGraph[1] workflow, the framework performs semantic chunking, field-conditioned retrieval, field-specific extraction, validation, evaluation, and aggregation. Unlike direct prompting, the system stores intermediate outputs across stages, enabling pipeline-level traceability and allowing failures to be linked to specific stages such as retrieval, extraction, normalization, or aggregation. This design supports structured auditing and future clinician-guided refinement for clinical information extraction.

2 Related Work

2.1 Clinical Information Extraction from Pathology Reports

Clinical information extraction evolved from rule-based systems to transformer-based and LLM-based approaches. Although transformer models improved generalization across tasks such as named entity recognition and relation extraction, pathology reports remain challenging due to dense terminology, implicit structure, and contextual ambiguity.

Recent studies demonstrated that LLMs can perform structured extraction from clinical and pathology documents without extensive task-specific training. Chow et al. (2025) demonstrated strong performance of open-weight LLMs for schema-aligned pathology abstraction. Hart and Bergamaschi (2026) proposed an agent-based framework for breast cancer pathology extraction, while Gupta et al. (2025) introduced a hierarchical agentic system for large-scale oncology data extraction. However, these approaches mainly focus on single-task extraction, institution-specific workflows, or direct prompting strategies with limited pipeline-level interpretability.

2.2 LLM Agents in Biomedicine and Oncology

Agentic AI expands the capabilities of LLMs by enabling iterative reasoning, external tool interaction, memory utilization, and coordinated multi-step problem solving (Yao et al., 2022; Xu and Sankar, 2025; Yang et al., 2025; Radi et al., 2025; Zhao et al., 2025b; Abdollahi et al., 2025). Foundational work such as ReAct demonstrated the advantages of integrating reasoning with action execution (Yao et al., 2022), while subsequent studies proposed biomedical agentic frameworks incorporating orchestration and tool-assisted workflows (Abbasian et al., 2024; Lin et al., 2024; Li et al., 2025; Inoue et al., 2025). Recent surveys further emphasized the growing applications of agentic AI in clinical decision-making, research automation, and biomedical knowledge synthesis (Xu and Sankar, 2025; Yang et al., 2025; Radi et al., 2025; Zhao et al., 2025b; Abdollahi et al., 2025).

In oncology, agentic frameworks have been explored for decision support, diagnostic reasoning, treatment planning (Ferber et al., 2025; Chen et al., 2025; Kuerbanjiang et al., 2025; Liu et al., 2026), and biomedical research applications such as drug discovery and biomarker reasoning (Inoue et al., 2025; Zuo et al., 2025; Jin et al., 2025; Huang et al., 2025; Bazgir et al., 2025). However, most work focuses on decision support or research workflows rather than structured cancer registry extraction from heterogeneous pathology documentation.

2.3 Research Gap

Despite progress in clinical information extraction and agentic AI, several gaps remain at their intersection. Existing approaches often rely on direct prompting strategies that process entire reports in a single inference step. While effective for explicit patterns such as TNM expressions, these approaches provide limited visibility into intermediate reasoning stages and extraction failures. In cancer registry abstraction, clinically relevant evidence may be distributed across multiple report sections, expressed implicitly, or embedded within noisy longitudinal documentation.

To address these limitations, we propose a modular multi-agent framework implemented as a sequential LangGraph workflow that performs semantic chunking, field-conditioned retrieval, field-specific extraction, validation, evaluation, and aggregation. The framework additionally stores intermediate outputs across stages, enabling pipeline-level traceability, structured auditing, and future clinician-guided refinement.

3 Methodology

3.1 Study Overview

We propose a modular multi-agent framework for structured cancer registry extraction from pathology and medical reports, targeting clinically relevant variables including grade, morphology, TNM staging, and laterality. Unlike single-pass LLM

[1] https://github.com/langchain-ai/langgraph

prompting, the framework decomposes extraction into sequential specialized stages implemented as a directed LangGraph workflow. Agents operate over a shared state and progressively refine intermediate outputs, enabling modular execution, structured auditing, and pipeline-level traceability.

3.2 Dataset

The dataset was collected from Sultan Qaboos University Hospital and includes 818 annotated cancer cases. Clinical reports were independently reviewed by trained pathologists, with disagreements adjudicated by a senior pathologist to establish the final gold standard annotations. Inter-annotator agreement analysis across registry variables and cancer types is provided in Appendix B.

Although the overall dataset contains multiple cancer types, the current study focuses on breast (n=454) and colorectal (n=174) reports for evaluation across four registry variables: grade, morphology, TNM staging, and laterality.

3.3 Multi-Agent Architecture

The proposed system consists of sequential specialized agents, as illustrated in Figure 1. Additional implementation details and source code availability are provided in Code Availabilty section.

- **Chunking Agent:** Segments reports into semantically coherent chunks and assigns field-relevance labels.
- **Retriever Agent:** Selects field-specific evidence relevant to grade, morphology, TNM staging, and laterality extraction. Retrieval is used to reduce irrelevant context and improve extraction focus for context-dependent variables.
- **Field-Specific Extraction Agents:** Dedicated agents independently extract grade, morphology, T, N, M, and laterality using task-specific prompts and structured output constraints.
- **Reviewer Agent:** Validates outputs, enforces schema consistency, and repairs malformed predictions.
- **Evaluator Agent:** Computes evaluation metrics when reference labels are available, or estimates confidence using evidence alignment and internal consistency.
- **Aggregation Agent:** Consolidates validated outputs into a unified structured record containing predictions, normalized labels, confidence scores, and supporting evidence.

All prompts are provided in Appendix D. Prompt design was iteratively refined through experimental evaluation and error inspection to improve schema consistency and extraction reliability across fields.

3.4 Workflow Orchestration

The pipeline is implemented using LangGraph as a directed execution graph operating over a shared state containing reports, retrieved evidence, intermediate outputs, predictions, and confidence scores. This design enables modular execution, transparent information flow, and stage-level error tracing.

3.5 Model Communication Protocol (MCP)

We introduce a lightweight Model Communication Protocol (MCP) that standardizes communication between agents through unified schemas for inputs, outputs, intermediate states, and metadata. MCP supports consistent state updates, logging, and reproducible execution across the pipeline.

3.6 Persistent Audit Database

A SQLite-based persistence module stores reports, intermediate outputs, predictions, and confidence values across runs, supporting auditing, traceability, and future refinement.

3.7 Implementation Details

The framework is implemented in Python using LangGraph. LLaMA 3.3 (70B) is deployed locally through Ollama (v0.4.7) to preserve data privacy. The system runs on Ubuntu 22.04 LTS with Python 3.10, Streamlit for visualization, and Pandas for data processing. Experiments were conducted using two NVIDIA RTX 6000 Ada GPUs (48 GB each) with CUDA 13.0.

3.8 Baseline Method

As a baseline, we implement direct prompt-based extraction using LLaMA 3.3 without task decomposition, retrieval, or intermediate reasoning stages. Multiple prompting strategies were evaluated, including simple/prefix prompting (direct extraction instructions), cloze prompting (fill-in-the-blank style extraction), anticipatory prompting (guiding the model with preparatory contextual instructions), chain-of-thought prompting (step-wise reasoning

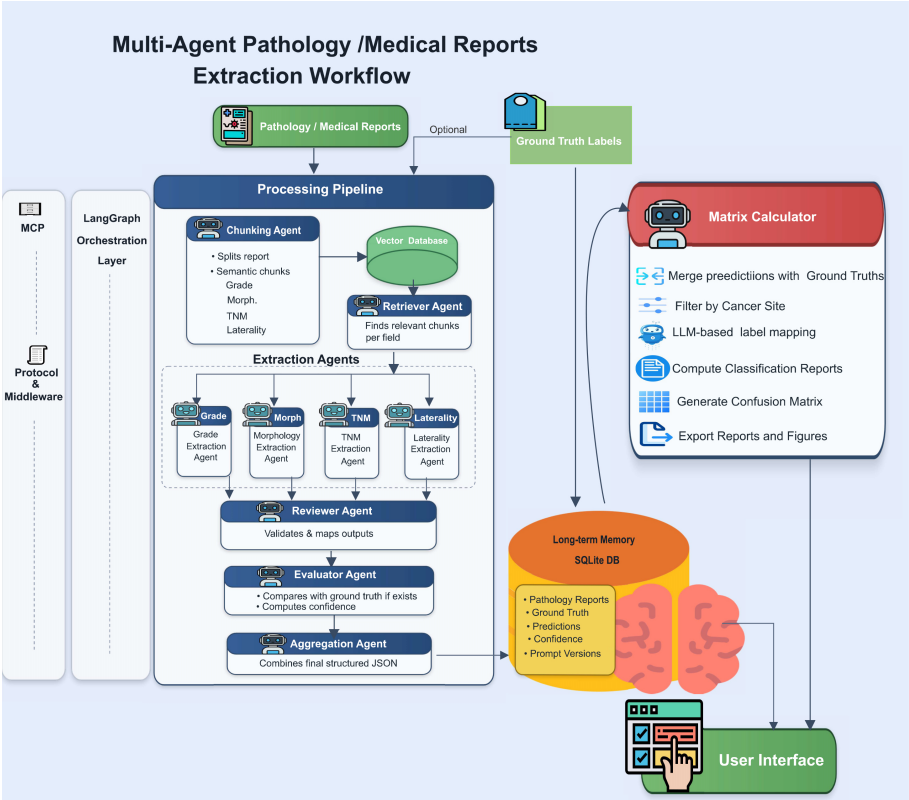


Figure 1: Overview of the proposed multi-agent framework. The pipeline consists of sequential agents for chunking, retrieval, field-specific extraction, validation, evaluation, and aggregation coordinated through a shared state.

before prediction), and heuristic prompting (rule-guided extraction instructions). For each extraction task, the best-performing baseline prompting strategy was selected for comparison against the proposed multi-agent framework. Detailed baseline prompts are provided in Appendix C.

Compared to direct prompting, the proposed multi-agent framework introduces additional computational overhead due to sequential agent execution, retrieval stages, and intermediate validation steps. However, the current study primarily focuses on extraction quality, interpretability, and traceability rather than runtime optimization.

3.9 Evaluation Metrics

Performance is evaluated using accuracy, macro-averaged, and weighted precision, recall, and F1-score. Macro metrics assess performance equally across classes, while weighted metrics account for class imbalance and reflect real-world clinical distributions.

Predictions are evaluated using exact-label matching against ground-truth registry annotations.

4 Results and Discussion

This section compares the proposed multi-agent framework with prompt-based LLaMA 3.3 baselines for breast and colorectal cancer registry extraction across grade, morphology, TNM staging, and laterality tasks. Because several fields are imbalanced, weighted F1-score is used as the primary metric, while macro metrics evaluate class-balanced performance.

Overall, results reveal task-dependent behavior. The multi-agent framework improves context-dependent tasks such as grade extraction and colorectal laterality, while achieving performance comparable to direct prompting for highly structured fields such as TNM staging and morphology.

4.1 Grade Extraction Performance

Table 1 summarizes grade extraction performance across breast and colorectal cancer sites. Figure 2 presents weighted metrics, while weighted F1-only and macro-level comparisons are provided in Appendix A (Figures 8 and 9).

For breast cancer, the multi-agent framework improved weighted F1-score from 0.71 to 0.78 and accuracy from 0.64 to 0.73. For colorectal cancer,

Table 1: Comparative performance for grade extraction in breast and colorectal cancer.

Site	Method	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$
Breast	Baseline-Anticipatory	0.64	0.83	0.64	0.71	0.55	0.61	0.51
Breast	Multi-Agent	0.73	0.83	0.73	0.78	0.58	0.54	0.56
Colorectal	Baseline-Anticipatory	0.44	0.85	0.44	0.56	0.40	0.34	0.34
Colorectal	Multi-Agent	0.59	0.80	0.59	0.67	0.34	0.30	0.31

Table 2: Comparative performance for morphology extraction in breast and colorectal cancer.

Site	Method	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$
Breast	Baseline-Anticipatory	0.75	0.91	0.75	0.82	0.34	0.40	0.33
Breast	Multi-Agent	0.72	0.92	0.72	0.80	0.35	0.54	0.34
Colorectal	Baseline-Simple	0.86	0.87	0.86	0.87	0.31	0.31	0.31
Colorectal	Multi-Agent	0.67	0.99	0.67	0.80	0.33	0.23	0.27

weighted F1-score increased from 0.56 to 0.67 and accuracy from 0.44 to 0.59. These improvements suggest that decomposition and field-specific evidence retrieval are beneficial when grade-related information is context-dependent or distributed across multiple report sections.

However, the slight decrease in colorectal macro F1-score indicates reduced sensitivity for less frequent classes. Overall, grade extraction represents the strongest advantage of the proposed multi-agent framework over direct prompting.

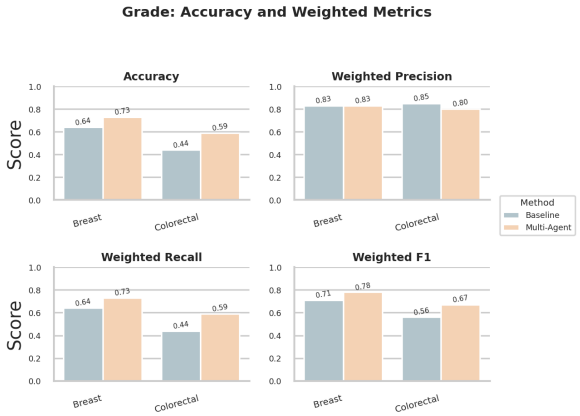


Figure 2: Accuracy, weighted precision, weighted recall, and weighted F1-score comparison between the baseline and the multi-agent framework for grade extraction across breast and colorectal cancer sites.

4.2 Morphology Extraction Performance

Table 2 and Figure 3 summarize morphology extraction performance. Additional weighted F1-only and macro-level comparisons are provided in Figures 10 and 11. For breast cancer, the multi-agent framework remained close to the baseline, with weighted F1-score decreasing slightly from 0.82 to 0.80 while macro F1-score improved marginally from 0.33 to 0.34. In colorectal cancer, the baseline achieved higher weighted F1-score, increasing from 0.80 to 0.87 (+0.07) compared to the multi-agent framework.

These findings suggest that morphology extraction often depends on explicit histologic terminology that can already be captured effectively through direct prompting. Additional retrieval and normalization stages may occasionally introduce information loss or normalization inconsistencies, particularly for less frequent morphology classes.

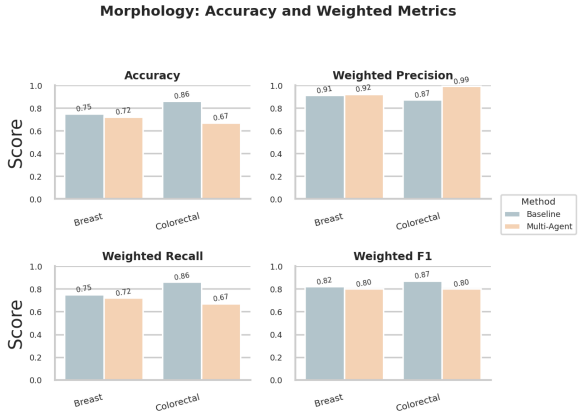


Figure 3: Accuracy, weighted precision, recall, and weighted F1-score comparison between the baseline and the multi-agent framework for morphology extraction across breast and colorectal cancer sites.

4.3 TNM Extraction Performance

Table 3 and Figure 4 summarize TNM extraction performance. Additional weighted F1-only and macro-level analyses are provided in Figures 12, 13, 14, 15, 16, and 17. Across TNM components, the baseline generally achieved stronger weighted performance, particularly for T and M categories. This is expected because TNM staging often appears in concise and standardized forms that are well suited to direct prompting. In contrast, the multi-agent framework introduces additional chunking, retrieval, and normalization stages that may occasionally reduce useful contextual information.

4.3.1 T Category

For T-stage extraction, the baseline outperformed the multi-agent framework in both cancer sites. In breast cancer, weighted F1-score decreased from 0.75 to 0.66, while in colorectal cancer it decreased slightly from 0.60 to 0.59. The T sub-figure in Figure 4 confirms that the baseline achieves stronger weighted metrics across both sites, particularly in

breast cancer. These findings suggest that T-stage extraction relies heavily on concise and standardized expressions that can already be captured effectively through direct prompting.

4.3.2 N Category

For N-stage extraction, the multi-agent framework achieved performance comparable to the baseline in breast cancer, with weighted F1-scores of 0.62 for both approaches while slightly improving macro F1-score. In colorectal cancer, the baseline achieved higher weighted F1-score, whereas the multi-agent framework showed slightly improved macro-level behavior. The N sub-figure in Figure 4 illustrates this site-dependent trend.

4.3.3 M Category

For M-stage extraction, the baseline consistently outperformed the multi-agent framework across both cancer sites. In breast cancer, weighted F1-score decreased from 0.60 to 0.48, while in colorectal cancer it decreased from 0.66 to 0.48. The M sub-figure in Figure 4 highlights a clear precision–recall trade-off: although the multi-agent framework achieved higher precision in some settings, it suffered from substantially lower recall, suggesting conservative predictions that missed metastasis-related cases.

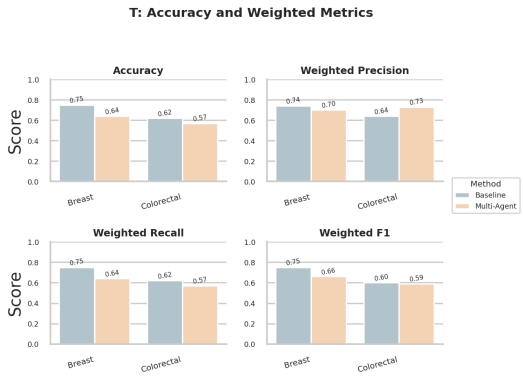
4.4 Laterality Extraction Performance

Table 4 and Figure 5 summarize laterality extraction performance. Weighted F1-only and macro-level comparisons are shown in Figures 18 and 19.

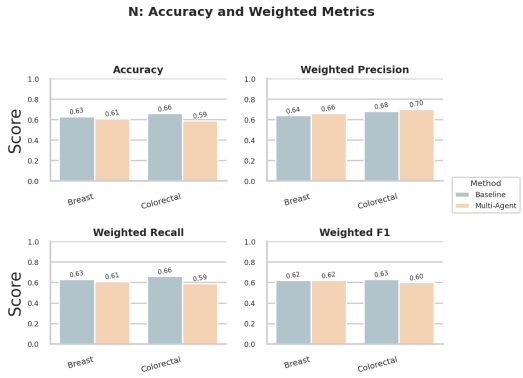
Laterality extraction showed site-dependent behavior. In breast cancer, both approaches achieved high performance, with weighted F1-score of 0.92 for the baseline and 0.91 for the multi-agent framework. Figure 5 confirms that differences across weighted metrics are minimal in this setting.

In colorectal cancer, the multi-agent framework improved weighted F1-score from 0.45 to 0.50 and macro F1-score from 0.22 to 0.27. The corresponding figure suggests that these gains are associated with improved recall and better handling of less explicit laterality expressions.

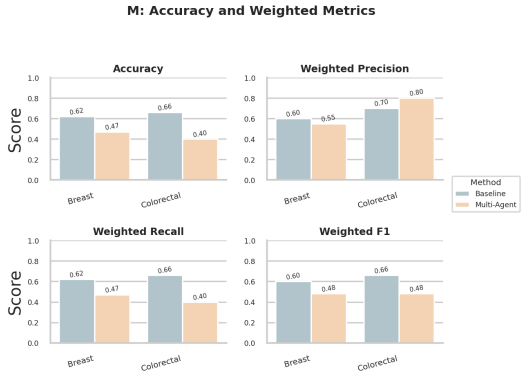
Overall, laterality emerged as one of the more stable extraction tasks, particularly in breast cancer where laterality is often explicitly stated. These findings further suggest that decomposition and field-specific retrieval are beneficial when laterality information is expressed inconsistently or implicitly across report sections.



T-stage



N-stage



M-stage

Figure 4: Accuracy, weighted precision, recall, and weighted F1-score comparison for TNM extraction across breast and colorectal cancer sites.

4.5 Overall Cross-Field Performance Analysis

To assess global system behavior, we analyze weighted F1-scores across all extraction tasks (grade, morphology, TNM, and laterality) using radar plots for each cancer site.

Figures 6 and 7 show that the multi-agent framework redistributes performance across tasks rather than uniformly improving all of them. In breast cancer, the largest gain is observed in grade extraction, while morphology and laterality remain competitive and TNM categories, particularly T

Table 3: Comparative performance for TNM extraction in breast and colorectal cancer.

Cat.	Site	Method	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$
T	Breast	Baseline-Cloze	0.75	0.74	0.75	0.75	0.65	0.60	0.62
T	Breast	Multi-Agent	0.64	0.70	0.64	0.66	0.51	0.47	0.47
T	Colorectal	Baseline-Cloze	0.62	0.64	0.62	0.60	0.61	0.48	0.51
T	Colorectal	Multi-Agent	0.57	0.73	0.57	0.59	0.55	0.46	0.47
N	Breast	Baseline-Cloze	0.63	0.64	0.63	0.62	0.46	0.44	0.43
N	Breast	Multi-Agent	0.61	0.66	0.61	0.62	0.47	0.46	0.45
N	Colorectal	Baseline-Anticipatory	0.66	0.68	0.66	0.63	0.49	0.53	0.47
N	Colorectal	Multi-Agent	0.59	0.70	0.59	0.60	0.58	0.49	0.50
M	Breast	Baseline-Cloze	0.62	0.60	0.62	0.60	0.56	0.50	0.51
M	Breast	Multi-Agent	0.47	0.55	0.47	0.48	0.44	0.47	0.40
M	Colorectal	Baseline-Anticipatory	0.66	0.70	0.66	0.66	0.60	0.58	0.57
M	Colorectal	Multi-Agent	0.40	0.80	0.40	0.48	0.64	0.47	0.40

Table 4: Comparative performance for laterality extraction in breast and colorectal cancer.

Site	Method	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$
Breast	Baseline-Heuristic	0.90	0.94	0.90	0.92	0.58	0.42	0.45
Breast	Multi-Agent	0.89	0.92	0.89	0.91	0.47	0.46	0.46
Colorectal	Baseline-Simple	0.36	0.72	0.36	0.45	0.27	0.48	0.22
Colorectal	Multi-Agent	0.40	0.73	0.40	0.50	0.35	0.47	0.27

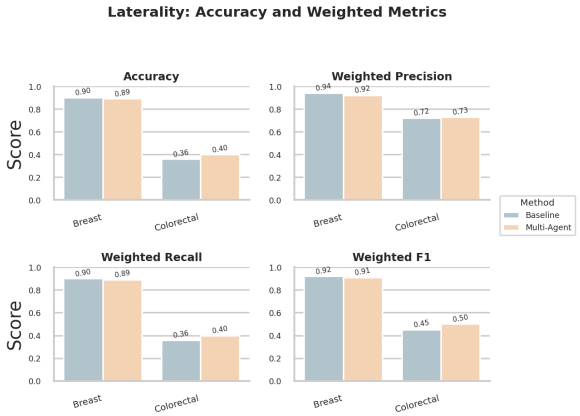


Figure 5: Accuracy, weighted precision, recall, and weighted F1-score comparison between the baseline and the multi-agent framework for laterality extraction across breast and colorectal cancer sites.

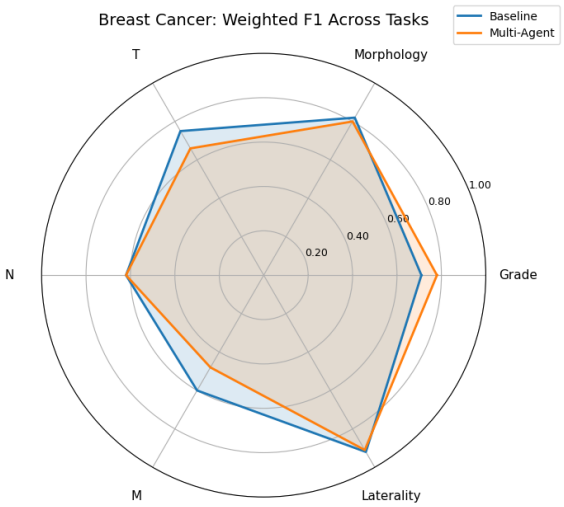


Figure 6: Weighted F1-score comparison across extraction tasks for breast cancer.

Table 5: Average weighted F1-score across all extraction tasks for each cancer site.

Site	Baseline	Multi-Agent
Breast	0.74	0.71
Colorectal	0.63	0.61

and M, favor the baseline. In colorectal cancer, the multi-agent framework improves grade and laterality but underperforms in morphology and M-stage extraction.

Table 5 reports the average weighted F1-score across tasks. The baseline achieves slightly higher overall averages in both breast (0.74 vs. 0.71) and colorectal (0.63 vs. 0.61). However, these aggregate results mask important task-dependent differences and varying levels of clinical complexity.

The baseline performs better in highly structured tasks such as TNM staging and morphology extraction, where clinically relevant information often appears in concise and standardized forms. In contrast, the multi-agent framework demonstrates clear advantages in more context-dependent tasks such as grade extraction and colorectal laterality, where evidence may be distributed across sections or expressed implicitly.

Pipeline-level inspection suggests that performance degradation primarily originates from retrieval and normalization stages rather than extraction itself. For structured fields, chunking and retrieval may occasionally remove useful context or introduce unnecessary normalization steps. In contrast, decomposition improves extraction when contextual reasoning and evidence localization are required.

A formal ablation study isolating the contribution of chunking, retrieval, and normalization stages was not conducted in the current study, as the primary objective was to evaluate the overall

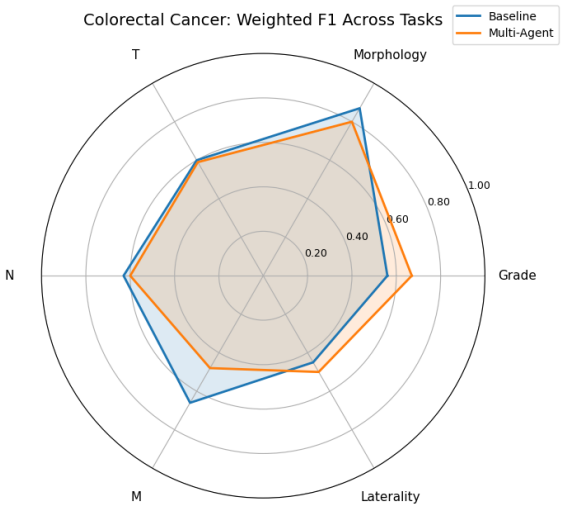


Figure 7: Weighted F1-score comparison across extraction tasks for colorectal cancer.

behavior and clinical applicability of the complete end-to-end framework.

Overall, the proposed framework should not be viewed solely as a score-maximizing alternative, but as a modular and auditable extraction pipeline that supports structured reasoning, traceability, and systematic error analysis.

5 Future Work

Future work will focus on improving retrieval-aware extraction, adaptive chunk selection, and rare-class calibration, particularly for conservative settings such as M-stage extraction where recall degradation was observed. Since performance varies across tasks, future optimization should target field-specific pipeline behavior rather than uniformly increasing model complexity.

Future work will also investigate clinician-guided validation and interactive correction workflows to support targeted refinement of ambiguous predictions. Such feedback may help improve schema normalization, reduce context-related extraction errors, and support iterative prompt and model refinement.

Future evaluation across institutions with different reporting conventions is necessary to better assess generalization and robustness.

Additional directions include evaluating domain-specific biomedical LLMs, comparing alternative prompting strategies more systematically, extending the framework to additional cancer registry fields and cancer sites, and evaluating generalization across institutions with different reporting conventions.

6 Conclusion

In this study, we proposed a modular multi-agent framework for extracting structured cancer registry information from pathology and medical reports. Implemented as a sequential LangGraph workflow, the system decomposes extraction into semantic chunking, retrieval, field-specific extraction, validation, evaluation, and aggregation stages, enabling structured intermediate outputs and pipeline-level traceability.

Experimental results demonstrate that performance is strongly task dependent. The multi-agent framework provides clear advantages in context-dependent tasks such as grade extraction and colorectal laterality, while direct prompting performs better in highly structured tasks such as TNM staging and morphology extraction. These findings highlight an important trade-off between reasoning-driven decomposition and direct pattern-based extraction.

Beyond aggregate performance, the proposed framework provides practical benefits through modular execution, structured auditing, and interpretable intermediate outputs that support error analysis and future refinement. Overall, this work positions multi-agent clinical information extraction as a promising and extensible direction for reliable cancer registry abstraction and future clinically guided NLP systems.

Acknowledgments

This research was funded by Sultan Qaboos University Research Fund under grant number SR-SCI-COMP-22-01.

Code Availability

The implementation of the proposed multi-agent framework, including the LangGraph workflow, extraction agents, evaluation pipeline, and Streamlit interfaces, is publicly available at:
https://github.com/RihamJeeballah/multi-agent-cancer-registry-extraction

References

Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, and Ramesh Jain. 2024. Conversational health agents: A personalized llm-powered agent framework.

Ali Abdollahi, Mohammad Amin Rezaei, Xi Wang, Sunan He, Seyed Moein Ayyoubzadeh, Yuxiang Nie, and Hao Chen. 2025. The next pradigm in medical ai: A survey of agentic ai in biomedicine.

Adib Bazgir, Amir Habibdoust Lafmajani, and Yuwen Zhang. 2025. Beyond correlation: Towards causal large language model agents in biomedicine.

Xi Chen, Huahui Yi, Mingke You, Wei Zhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8.

Nan-Haw Chow, Han Chang, Hung-Kai Chen, Chen-Yuan Lin, Ying-Lung Liu, Po-Yen Tseng, Li-Ju Shiu, Yen-Wei Chu, Pau-Choo Chung, and Kai-Po Chang. 2025. Comprehensive structured abstraction of pathology reports is now feasible using local large language models.

Dyke Ferber, Omar S.M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann, Marie Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn, and Jakob Nikolas Kather. 2025. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer*, 6:1337–1349.

Alon Gorenshtein, Mahmud Omar, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. 2025. Ai agents in clinical medicine: A systematic review.

Shashi Kant Gupta, Arijeet Pramanik, Jerrin John Thomas, Regina Schwind, Lauren Wiener, Avi Raju, Jeremy Kornbluth, Yanshan Wang, Zhaohui Su, and Hrituraj Singh. 2025. Harmon-e: Hierarchical agentic reasoning for multimodal oncology notes to extract structured data.

Steven N. Hart and Teya S. Bergamaschi. 2026. Agent-based large language model system for extracting structured data from breast cancer synoptic reports: a dual-validation study. *JAMIA Open*, 9.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N Carter, Xin Zhou, Matthew Wheeler, Jonathan A Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, and 4 others. 2025. Biomni: A general-purpose biomedical ai agent. *bioRxiv : the preprint server for biology*.

Yoshitaka Inoue, Tianci Song, Xinling Wang, Augustin Luna, and Tianfan Fu. 2025. Drugagent: Multi-agent large language model-based reasoning for drug-target interaction prediction.

Ruofan Jin, Zaixi Zhang, Mengdi Wang, and Le Cong. 2025. Stella: Self-evolving llm agent for biomedical research.

Warisijiang Kuerbanjiang, Xinyu Wang, Yiershatijiang Jiamaliding, Musitapa Maimaitiaili, and Yuexiong Yi. 2025. Multidisciplinary large language model agent teams for precision oncology enhance complex gynecologic oncology decision support.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2025. Agent hospital: A simulacrum of hospital with evolvable medical agents.

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z. Li, and Kaicheng Yu. 2024. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science.

Qicai Liu, Zhichao Hu, Tao Huang, Yupeng Niu, Xinche Zhang, Shanwu Ma, Chutong Lin, Goh Kim Huat, Hyeokkoo Eric Kwon, Feng Gao, Xianfu Sun, Zhitao Ying, and Guangliang Qiang. 2026. Evomdt: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *npj Digital Medicine*.

Abdul Mohaimen Al Radi, Xu Cao, Fanyang Yu, Yuyuan Liu, Fengbei Liu, Chong Wang, Yuanhong Chen, Jintai Chen, Hu Wang, Yanda Meng, Zhenyi Wang, Chen Chen, Mubarak Shah, Tianyu Han, Christos Davatzikos, MacLean P. Nasrallah, and Yu Tian. 2025. Agentic large-language-model systems in medicine: A systematic review and taxonomy.

Xiaoran Xu and Ravi Sankar. 2025. Large language model agents for biomedicine: A comprehensive review of methods, evaluations, challenges, and future directions.

Tiantian Yang, Yihang Xiao, Zhijie Bao, Jianye Hao, and Jiajie Peng. 2025. The rise and potential opportunities of large language model agents in bioinformatics and biomedicine.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models.

Lina Zhao, Jiaxing Bai, Zihao Bian, Qingyue Chen, Yafang Li, Guangbo Li, Min He, Huaiyuan Yao, and Zongjiu Zhang. 2025a. Autonomous multi-modal llm agents for treatment planning in focused ultrasound ablation surgery.

Wenqi Zhao, Shansong Wang, Mojtaba Safari, Mingzhe Hu, and Xiaofeng Yang. 2025b. Medical ai agents: A comprehensive survey of architectures, cognitive modules, and clinical workflows.

Kaiwen Zuo, Zixuan Zhong, Peizhou Huang, Shiyan Tang, Yuyan Chen, and Yirui Jiang. 2025. Healkggen: A hierarchical multi-agent llm framework with knowledge graph enhancement for genetic biomarker-based medical diagnosis.

Çağatay Umut Öğdü, Kübra Arslanoğlu, and Mehmet Karaköse. 2025. An adaptive multi-agent llm-based clinical decision support system integrating biomedical rag and web intelligence. *IEEE Access*, 13:167390–167404.

A Additional Results

This appendix provides supplementary visualizations for weighted F1-only comparisons and macro-level comparisons across extraction tasks.

A.1 Grade Extraction

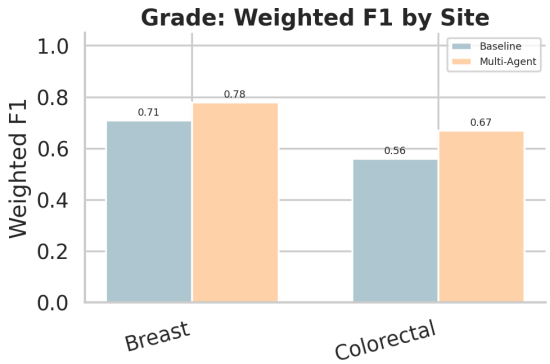


Figure 8: Weighted F1-score comparison between the prompt-based baseline and the multi-agent framework for grade extraction across breast and colorectal cancer sites.

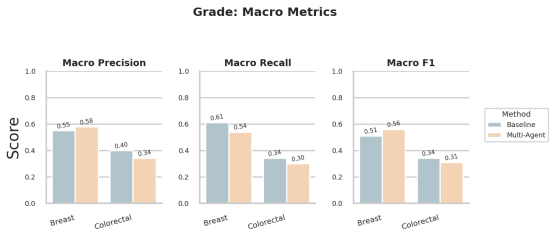


Figure 9: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for grade extraction across breast and colorectal cancer sites.

A.2 Morphology Extraction

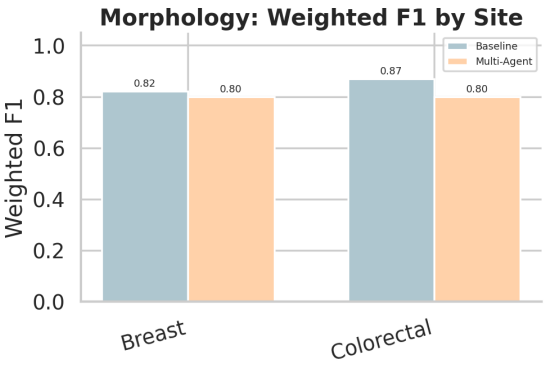


Figure 10: Weighted F1-score comparison between the baseline and the multi-agent framework for morphology extraction across breast and colorectal cancer sites.

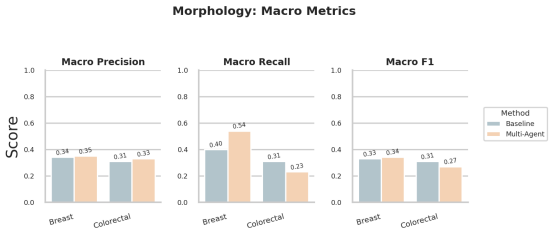


Figure 11: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for morphology extraction across breast and colorectal cancer sites.

A.3 T-Stage Extraction

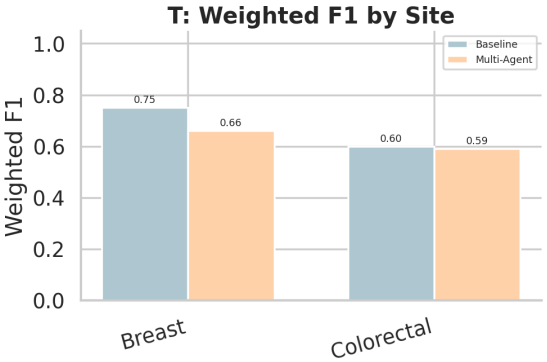


Figure 12: Weighted F1-score comparison between the baseline and the multi-agent framework for T-stage extraction across breast and colorectal cancer sites.

Figure 13: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for T-stage extraction across breast and colorectal cancer sites.

A.4 N-Stage Extraction

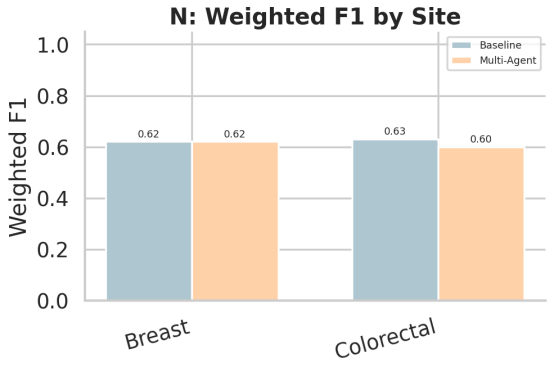


Figure 14: Weighted F1-score comparison between the baseline and the multi-agent framework for N-stage extraction across breast and colorectal cancer sites.

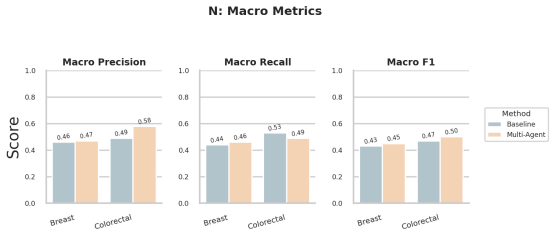


Figure 15: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for N-stage extraction across breast and colorectal cancer sites.

A.5 M-Stage Extraction

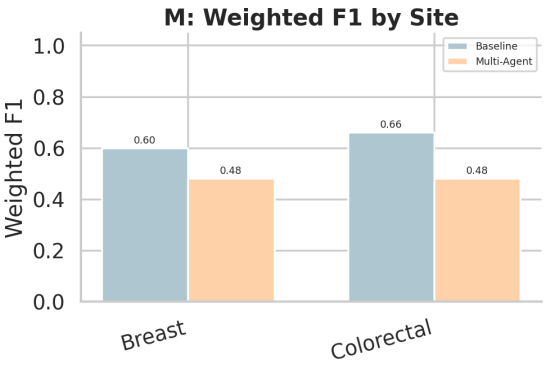


Figure 16: Weighted F1-score comparison between the baseline and the multi-agent framework for M-stage extraction across breast and colorectal cancer sites.

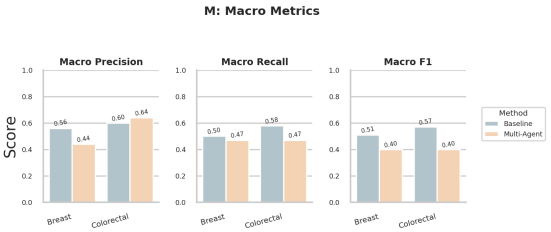


Figure 17: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for M-stage extraction across breast and colorectal cancer sites.

A.6 Laterality Extraction

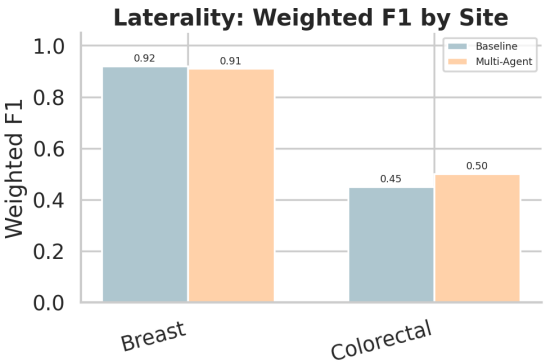


Figure 18: Weighted F1-score comparison between the baseline and the multi-agent framework for laterality extraction across breast and colorectal cancer sites.

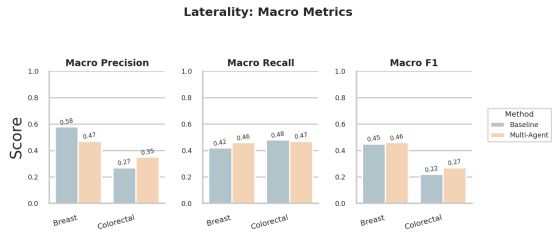


Figure 19: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for laterality extraction across breast and colorectal cancer sites.

B Inter-Annotator Agreement (IAA)

To assess the reliability of the gold-standard annotations, inter-annotator agreement (IAA) was evaluated across all registry variables using Cohen's κ coefficient and percentage agreement. The overall agreement across all fields was $\kappa = 0.917$, with a percentage agreement of 91.99%, indicating high consistency between annotators.

B.1 Overall Agreement Across Registry Fields

Table 6 presents the overall agreement for each registry variable. All fields demonstrate strong agreement, with most achieving "almost perfect" consistency.

Table 6: Overall inter-annotator agreement across registry variables.

Field	Cohen's κ	% Agreement	Interpretation
GRADE	0.968	97.83	Almost perfect
STAGE	0.909	92.05	Almost perfect
MORPHOLOGY	0.981	98.71	Almost perfect
T	0.820	86.80	Almost perfect
N	0.904	92.64	Almost perfect
M	0.812	88.48	Substantial
LATERALITY	0.869	90.33	Almost perfect

B.2 IAA by Cancer Type

Table 7 reports agreement stratified by cancer type. Agreement remains consistently high across most cancer sites, with slight variability observed in more complex fields such as T, N, and M staging, particularly for colorectal and thyroid cases.

C Baseline Prompt Designs

This appendix presents the prompt designs used for the baseline prompt-based LLM extraction. Multiple prompting strategies were explored, including prefix, cloze, anticipatory, chain-of-thought, and heuristic prompting.

Table 7: Inter-annotator agreement stratified by cancer type.

Cancer Type	Field	κ	%	Interpretation
Breast	GRADE	0.975	98.83	Almost perfect
	STAGE	0.944	95.48	Almost perfect
	MORPHOLOGY	0.914	96.47	Almost perfect
	T	0.916	94.96	Almost perfect
	N	0.933	95.88	Almost perfect
	M	0.930	95.88	Almost perfect
	LATERALITY	0.958	97.50	Almost perfect
Colorectal	GRADE	0.852	91.49	Almost perfect
	STAGE	0.889	91.49	Almost perfect
	MORPHOLOGY	0.964	97.87	Almost perfect
	T	0.576	73.49	Moderate
	N	0.918	92.42	Almost perfect
	M	0.784	85.38	Substantial
	LATERALITY	0.854	91.49	Almost perfect
Prostate	GRADE	0.992	99.03	Almost perfect
	STAGE	0.983	98.06	Almost perfect
	MORPHOLOGY	0.986	98.71	Almost perfect
	T	0.879	87.74	Almost perfect
	N	0.903	90.32	Almost perfect
	M	0.735	80.65	Substantial
	LATERALITY	0.873	90.32	Almost perfect
Thyroid	GRADE	0.934	97.80	Almost perfect
	STAGE	0.416	24.53	Moderate
	MORPHOLOGY	0.979	99.72	Almost perfect
	T	0.702	71.49	Substantial
	N	0.255	34.84	Fair
	M	0.512	70.57	Moderate
	LATERALITY	0.701	86.88	Substantial

C.1 Grade Extraction Prompts

Table 8: Prompt designs for grade extraction.

Prompt Type	Prompt Text
Prefix (p1)	Extract only the numerical tumor grade from the medical report and return it in the JSON template. Ignore additional details. If multiple grades are mentioned, select the final diagnostic grade from the *Impression* section when available. Use the highest numerical tumor grade from the most recent relevant pathology-based report (e.g., pathology or biopsy) and ignore non-pathology reports.
Cloze (p2)	The tumor grade in the text is ___, the grading system used (if available) is ___, and the individual score components (if available) are ___. From the following concatenated clinical notes, extract the highest tumor grade mentioned across all relevant reports. Ignore unrelated or irrelevant sections. The highest tumor grade in the text is ___, the grading system used (if available) is ___, and the individual score components (if available) are ___.
Anticipatory (p3)	[Input 1:] How can tumor grade information be identified in a medical report? [Input 2:] For each tumor mentioned in the given medical report, extract the reported grade. If multiple grades are mentioned, select the final diagnostic grade from the *Impression* section and return only the numerical grade. [Input 1:] How can tumor grade information be identified in a medical history? [Input 2:] For each tumor mentioned in the patient's medical history, extract the highest relevant tumor grade from the most recent pathology report. If multiple grades are mentioned in that report, select the final diagnostic grade from the *Impression* or equivalent summary section. Ignore irrelevant or outdated reports and return only the numerical grade.
Chain-of-Thought (p4)	**EXAMPLE:** Identify and extract the tumor grade from the medical report. If the grading system and its components are mentioned, include them. **TEXT:** "Microscopic examination reveals an invasive ductal carcinoma with moderate tubule formation (score = 2), marked pleomorphism (score = 3), and 5 mitoses per 10 high-power fields (score = 2). Based on these features, the tumor is classified as Nottingham Grade II." **QUESTION 1:** What is the tumor grade mentioned in the medical report? **ANSWER 1:** The tumor grade is II (explicitly stated as Nottingham Grade II). **QUESTION 2:** What grading system is used in the report, if available? **ANSWER 2:** The grading system used is Nottingham. **QUESTION 3:** Is there any additional information about the tumor that should be noted? **ANSWER 3:** No additional features such as necrosis were mentioned. **QUESTION:** Using the stored example, extract the tumor grade, grading system, and score components from the following medical report. **TEXT:** `<text>`
Heuristic (p5)	[Input 1:] First, store these rules in memory: If the tumor grade is explicitly mentioned with a number, use that as the grade. If the grading system is Nottingham, check for component scores such as tubule formation, pleomorphism, and mitotic count. If the grade is described as "Grade I," "Grade II," or "Grade III," classify it accordingly. If it is described as "low," "moderate," or "high," map these to Grade I, II, or III, respectively. If the grading system is not mentioned but descriptors such as "well differentiated," "moderately differentiated," or "poorly differentiated" are used, assign Grade I, II, or III based on those terms. If the grade remains ambiguous, infer it from contextual clues. [Input 2:] Given the patient's full medical record history, extract the tumor grade, grading system, and individual component scores, if available, from the most recent pathology report that discusses tumor grading. Ignore irrelevant or outdated reports without pathology-based grading. Extract: (i) the highest relevant tumor grade in the most recent pathology document, (ii) the grading system used, and (iii) component scores such as tubule formation, pleomorphism, and mitotic count, when available.

C.2 Morphology Extraction Prompts

Table 9: Prompt designs for morphology extraction.

Prompt Type	Prompt Text
Prefix (p1)	Extract the most relevant morphology from the latest pathology report in the patient's medical history (e.g., pathology or biopsy). If available, prioritize the final morphology stated in the *Impression* or *Diagnosis* section. Select the final answer from the predefined label list: Adenocarcinoma, NOS; Carcinoma, NOS; Follicular adenocarcinoma, NOS; Infiltrating duct carcinoma, NOS; Lobular carcinoma, NOS; Medullary carcinoma, NOS; Mucinous adenocarcinoma; Neoplasm, malignant; Signet ring cell carcinoma. Ignore non-pathology reports. Return only the result in the specified JSON format without additional details.
Cloze (p2)	From the following concatenated clinical notes, extract the most relevant morphology from the latest pathology report. Ignore unrelated or irrelevant sections. The morphology mentioned in the text is ___, and the specific carcinoma type (if available) is ___. Select the final morphology from the predefined label list and return only the result in the specified JSON format.
Anticipatory (p3)	[Input 1:] How can morphology information be identified in a medical report history? [Input 2:] For each tumor mentioned in the patient's medical history, extract the most relevant morphology reported in the most recent pathology report. If multiple morphology types are mentioned, select the most relevant one. Prioritize the final morphology stated in the *Impression* or equivalent summary section. Ignore irrelevant or outdated reports. Choose the final answer from the predefined label list and return only the result in JSON format.
Chain-of-Thought (p4)	**EXAMPLE:** Identify and extract the morphology from the most recent pathology report. If multiple morphology types are mentioned, select the most relevant one, prioritizing the final diagnosis. **TEXT:** "The pathology report reveals features consistent with invasive ductal carcinoma. The diagnosis concludes invasive ductal carcinoma, NOS." **QUESTION 1:** What morphology is mentioned? **ANSWER 1:** Infiltrating duct carcinoma, NOS. **QUESTION 2:** Are there additional morphology types? **ANSWER 2:** No additional morphology types are introduced. **QUESTION 3:** What is the final morphology? **ANSWER 3:** Infiltrating duct carcinoma, NOS. **TEXT:** "Microscopic examination reveals mucinous adenocarcinoma with signet ring features. The final diagnosis is mucinous adenocarcinoma." **QUESTION 1:** What morphology is mentioned? **ANSWER 1:** Mucinous adenocarcinoma. **QUESTION 2:** Are additional patterns mentioned? **ANSWER 2:** Yes, signet ring features are mentioned, but not as the final diagnosis. **QUESTION 3:** What is the final morphology? **ANSWER 3:** Mucinous adenocarcinoma. Choose the final morphology from the predefined label list.
Heuristic (p5)	[Input 1:] First, apply the following rules: If morphology is explicitly stated, use it directly. If descriptors such as "ductal," "lobular," or "invasive" appear, map them to the corresponding morphology category. Terms such as "mucinous" map to Mucinous adenocarcinoma, and "signet ring" maps to Signet ring cell carcinoma. If only "adenocarcinoma" is mentioned without specification, classify as Adenocarcinoma, NOS. If only "carcinoma" is mentioned, classify as Carcinoma, NOS. Ambiguous descriptions should be mapped to Neoplasm, malignant. [Input 2:] From the patient's full medical history, extract the morphology from the most recent relevant pathology report. Ignore irrelevant or outdated reports. Select one label from the predefined list and return only the result in JSON format.

C.3 TNM Extraction Prompts

Table 10: Prompt designs for TNM staging extraction.

Prompt Type	Prompt Text
Prefix (p1)	Extract the most relevant TNM staging information from the most recent pathology or oncology report that mentions tumor staging. For each component (T, N, and M), extract the appropriate stage based on the findings. If substages (like T2bN1cM0) are mentioned, only extract the letter and numerical stage associated with each. Ignore non-pathology reports and irrelevant sections. Return the final TNM staging in the JSON template provided. Use the gold standard labels below to fill in each category: T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx
Cloze (p2)	From the following concatenated clinical notes, extract the TNM staging information based on the most relevant pathology report. Ignore unrelated or irrelevant sections. The T stage in the text is ___, the N stage in the text is ___, and the M stage in the text is ___. Ensure that the stages are chosen from the following gold standard labels: T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx If substages like T2bN1cM0 are mentioned, extract only the letter and numerical stage for each component. Your final response should be in the given JSON format.
Anticipatory (p3)	[Input 1:] How to identify TNM staging information in a medical report history? [Input 2:] For each tumor mentioned in the patient's medical history, extract the TNM staging information reported in the most recent pathology report. The TNM staging consists of the primary tumor stage (T), the regional lymph node stage (N), and the distant metastasis stage (M). If any substages are mentioned, extract only the letter and numerical stage associated with each component (e.g., T2bN1cM0 would be T2, N1, M0). Ignore irrelevant or outdated reports. Return the TNM stages in the provided JSON format. T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx
Chain-of-Thought (p4)	EXAMPLE: Identify and extract the TNM staging from the most recent relevant pathology report. If the staging includes substages (e.g., T2b, N1c), extract only the main stage component (e.g., T2, N1). Use only the values from the gold standard label list. TEXT: "The pathology report states: Tumor shows invasion into muscularis propria consistent with T2b staging. There is evidence of one positive regional lymph node (N1c). No distant metastasis identified (M0)." QUESTION 1: What is the T stage mentioned in the report? ANSWER 1: The report mentions T2b, so the T stage is T2. QUESTION 2: What is the N stage mentioned in the report? ANSWER 2: The report mentions N1c, so the N stage is N1. QUESTION 3: What is the M stage mentioned in the report? ANSWER 3: The report states M0 explicitly, so the M stage is M0. FINAL RESPONSE: T: "T2", N: "N1", M: "M0" Valid Label Options: T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx
Heuristic (p5)	[Input 1:] First, store these rules in memory: - If the T, N, or M stage is explicitly mentioned (e.g., "T2b", "N1c", "M0"), extract only the main stage (e.g., T2, N1, M0). - If a range of stages is given (e.g., "T1–T2", "N0–N1"), choose the highest one. - If only descriptive terms are used (e.g., "localized invasion," "no nodal involvement," "distant spread detected"), infer the TNM stage from context: "Localized tumor" → likely T1 or T2 "No lymph node involvement" → N0 "Metastatic disease" → M1 Prioritize the most recent relevant pathology report in the patient's medical history. Ignore outdated or irrelevant notes (e.g., preoperative assessments, non-pathology documentation). Use only the following label options for the final output: T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx [Input 2:] Given the patient's full medical record history, extract the TNM staging by identifying the most recent pathology report that provides clear staging information. Apply the rules above to extract: - T stage (primary tumor) - N stage (regional lymph nodes) - M stage (distant metastasis) Return your final answer in the given JSON format. Only use the values from the predefined label lists. Do not generate your own stages.

C.4 Laterality Extraction Prompts

Table 11: Prompt designs for laterality extraction.

Prompt Type	Prompt Text
Prefix (p1)	Extract the tumor laterality from the latest pathology report in the patient's medical history (e.g., pathology, biopsy). If available, prioritize the final laterality mentioned in the 'Impression' or 'Diagnosis' section. Choose your final answer from the following list: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un Ignore non-pathology reports. Return ONLY in the given JSON format, with a single key-value pair. Do NOT include additional qualifiers, sites, modifiers, or nested fields.
Cloze (p2)	From the following concatenated clinical notes, extract the tumor laterality from the latest pathology report. Ignore unrelated or irrelevant sections. The tumor laterality mentioned in the text is ___. Choose the laterality from the following list, and return ONLY in the given JSON format: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un
Anticipatory (p3)	[Input 1:] How to identify tumor laterality in a medical report history? [Input 2:] For each tumor mentioned in the patient's medical history, extract the laterality reported in the latest pathology report. If laterality is mentioned in multiple places within the same report, prioritize the clearest and most complete expression, especially from the 'Impression', 'Diagnosis', or equivalent summary sections. Anticipate ambiguities such as paired organs without a side specified, and distinguish between bilateral, unilateral, and unknown cases. Ignore outdated or irrelevant reports. Choose the laterality from the following options, and return ONLY in the given JSON format: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un
Chain-of-Thought (p4)	INSTRUCTIONS: Identify and extract the laterality from the most recent pathology report. If laterality is mentioned multiple times, prioritize the clearest and most definitive expression, especially from the 'Impression', 'Diagnosis', or equivalent summary sections. TEXT: "The tumor is located in the right breast, with no evidence of contralateral involvement. The final diagnosis section confirms malignancy in the right breast." QUESTION 1: What laterality is mentioned in the medical report? ANSWER 1: The laterality mentioned is right, primary organ (explicitly stated as right breast). QUESTION 2: Are there any ambiguous or conflicting laterality statements elsewhere in the report? ANSWER 2: No conflicting laterality is reported; all references are to the right side. QUESTION 3: What is the final laterality you would assign to this tumor based on the report? ANSWER 3: The final laterality assigned is right, primary organ. TEXT: "Pathology notes indicate bilateral lesions, with suspicious masses observed in both lobes. The diagnosis section confirms bilateral malignancy." QUESTION 1: What laterality is mentioned in the medical report? ANSWER 1: The laterality mentioned is bilateral involvement. QUESTION 2: Are there any ambiguous or conflicting laterality statements elsewhere in the report? ANSWER 2: No, the statement about bilateral malignancy is consistent. QUESTION 3: What is the final laterality you would assign to this tumor based on the report? ANSWER 3: The final laterality assigned is bilateral involvement. Choose the laterality from the following options: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un
Heuristic (p5)	[Input 1:] First, store these rules in memory: - If laterality is explicitly stated as right, left, or bilateral, use that information as the laterality type. - If the tumor is located in a paired organ (e.g., breast, kidney, lung), and the report mentions both sides, classify it as bilateral involvement. - If only one side is referenced (e.g., "mass in the right lobe"), assign left, primary organ or right, primary organ accordingly. - If the site is a paired organ but the laterality is not specified, classify it as paired, lat. unknown. - If the tumor occurs in an organ that is not paired (e.g., stomach, pancreas), classify it as not a paired site/un. [Input 2:] Given the patient's full medical report history, extract the laterality, ensuring you choose one of the gold standard labels. Identify the most recent relevant pathology report that includes a laterality mention. Ignore reports that are outdated or do not include laterality cues relevant to a paired primary tumor site. Use the rules mentioned above to extract: - The laterality of the tumor described in the most recent relevant pathology report. Make sure to choose the laterality from the following list: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un

D Multi-Agent Prompt Designs

D.1 Chunking Agent Prompt

Table 12: Prompt design for the chunking agent.

Component	Prompt Text

Prompt	You are a medical NLP system that segments ANY clinical text (even if it has no pathology-style headers) into semantically meaningful chunks.
	Output requirements - Return ONLY a single JSON object (no markdown, no explanations). - Schema (strict):
	`{`
	`  "chunk": [`
	`  {`
	`  "section": "string",`
	`  "text": "string",`
	`  "focus": "grade | morphology | tnm | laterality | treatment | other"`
	`  }`
	`  ]`
	`}`
	- The "chunk" array MUST be non-empty (never return an empty list). - Each item MUST have all three fields. - "focus" must be exactly one of: grade, morphology, tnm, laterality, treatment, other.
	Coverage & missing categories (very important) - Never fabricate text. - If a category (e.g., grade) is completely absent in the input, simply do NOT emit a chunk for that category. Downstream will treat missing categories as None. - The overall output MUST still contain at least one chunk. If no focal info is found, emit a single fallback chunk: - section: "other" - text: a short verbatim excerpt from the input (e.g., the first 1–3 sentences or the most informative line) - focus: "other"
	General rules (works for messy text) - The input may be free text, clinic letters, summaries, MDT notes, radiology-pathology blends, bullet lists, tables, or copy-paste artifacts. - If there are no formal headers, infer a best-fit "section" from this shortlist: ["diagnosis","gross","microscopy","immuno","comment","staging","clinical","treatment","history","other"]. - If the text mixes topics, split into small focused chunks (~1–4 sentences) and assign the primary focus: • grade → tumor grading phrases (e.g., "grade 2", "moderately differentiated", "Gleason group 3") • morphology → histologic type/cell type (e.g., "invasive ductal carcinoma", "adenocarcinoma", "papillary carcinoma") • tnm → any T/N/M mentions, stage group, pathologic vs clinical, even partial like "pT3", "N1", "M0", or "indeterminate" • laterality → left/right/bilateral/unifocal/multifocal; single-site organs can be "not applicable" only if clearly stated • treatment → surgery, chemo, radio, hormonal, targeted, immuno; past or planned • other → demographics, dates, admin, unrelated text - Keep each "text" verbatim (trim only leading/trailing whitespace and bullets). Do NOT invent placeholders like "none" or "[absent]".
	Normalization hints BEFORE deciding focus - TNM appears as "pT3 N1 M0", "T3N1M0", "T3 N1 (Mx)", embedded in sentences, or on separate lines. - Grades: prose ("moderately differentiated"), numerals ("grade II/2"), Gleason groups/scores. - Laterality: "Lt", "Rt", "B/L", "bilat." map to left/right/bilateral. - Morphology often near "diagnosis", "type", "histology", "consistent with".
	Edge cases & guarantees - If you cannot find any obvious clinical focus, still emit the fallback "other" chunk (see Coverage). - If TNM elements are scattered (e.g., T far from N), use multiple tnm chunks. - Prefer short, precise chunks over one giant block.
	Pathology/clinical text: `{report}`

D.2 Retriever Agent Prompt

Table 13: Prompt design for the retriever agent.

Component	Prompt Text

Prompt	You are a medical AI assistant helping extract information from pathology report chunks.
	Goal: Find the chunks most relevant to the field: **{field}**
	Chunks: `{chunks_text}`
	Instructions: - Return only relevant chunks for the field - Return ONLY JSON array of strings - If none → return []
	Example:
	`[`
	`  "Grade 2 tumor with moderate atypia."`
	`]`
	or
	`[]`
	No explanation.

D.3 Grade Extraction Agent Prompt

Table 14: Prompt design for the grade extraction agent.

Component	Prompt Text

Prompt	You are an expert medical assistant specialized in cancer pathology information extraction.
	Task: extract tumor grade.
	Return JSON only:
	1. "grade stated" 2. "grade estimated" (must be from valid values) 3. "grade_CD" 4. "grade_evidence"
	VALID VALUES: `{valid_values_bulleted}`
	Rules: - If no info → Unknown - Confidence 0–1 - Evidence = short quote
	Input: `{context}`
	Output:
	`{`
	`  "grade stated": "...",`
	`  "grade estimated": "...",`
	`  "grade_CD": 0.0,`
	`  "grade_evidence": "..."`
	`}`

D.4 Morphology Extraction Agent Prompt

Table 15: Prompt design for the morphology extraction agent.

Component	Prompt Text
Prompt	You are a medical language model specialized in extracting tumor morphology from pathology reports. Your task is to extract and estimate the tumor **morphology** from the given report sections below. Please output the following four key-value pairs: 1. "morphology stated": <The morphology as written in the pathology report, even if not standardized. If no morphology is directly stated, return "Unknown">, 2. "morphology estimated": <The closest standardized morphology based on the list below. Use your clinical knowledge to resolve synonymy or near matches. If unsure, return "Unknown">, 3. "morphology_CD": <Certainty degree of the estimation between 0.00 and 1.00>, 4. "morphology_evidence": <Key text or rationale supporting your estimation> Valid standardized morphologies: - Adenocarcinoma, NOS - Carcinoma, NOS - Follicular adenocarcinoma, NOS - Infiltrating duct carcinoma, NOS - Lobular carcinoma, NOS - Medullary carcinoma, NOS - Mucinous adenocarcinoma - Neoplasm, malignant - Signet ring cell carcinoma - Unknown Report Excerpt: `{context}` Output Format (JSON Only): `{ "morphology stated": "...", "morphology estimated": "...", "morphology_CD": ..., "morphology_evidence": "..." }`

D.5 TNM Extraction Agent Prompt

Table 16: Prompt design for the TNM extraction agent.

Component	Prompt Text
Prompt	You are a clinical information extraction assistant. Your task is to extract the patient's **TNM stage** from the pathology report excerpt below. Please extract the following key-value pairs for each TNM component: **T (Primary Tumor), N (Lymph Nodes), and M (Metastasis)**. Each component should have the following 4 fields: 1. "<component> stated": <As written in the pathology report — e.g., "T2", "N1", "M0", or "Unknown">, 2. "<component> estimated": <Your best estimate based on AJCC 7th edition, if the stated value is missing or ambiguous. Use "Unknown" if not inferable>, 3. "<component>_CD": <Certainty degree of the estimation from 0.00 to 1.00>, 4. "<component>_evidence": <Supporting text or reasoning from the report> Allowed values: - **T (Primary Tumor)**: ["T0", "Tis", "Tmic", "T1", "T2", "T3", "T4", "Tx", "Unknown"] - **N (Regional Lymph Nodes)**: ["N0", "N1", "N2", "N3", "Nx", "Unknown"] - **M (Distant Metastasis)**: ["M0", "M1", "Mx", "Unknown"] Report Sections: `{context}` Output Format (JSON Only): `{ "T stated": "...", "T estimated": "...", "T_CD": ..., "T_evidence": "...", "N stated": "...", "N estimated": "...", "N_CD": ..., "N_evidence": "...", "M stated": "...", "M estimated": "...", "M_CD": ..., "M_evidence": "..." }`

D.6 Laterality Extraction Agent Prompt

Table 17: Prompt design for the laterality extraction agent.

Component	Prompt Text

Prompt	You are an expert medical assistant specialized in cancer pathology information extraction.
	Your task is to extract and estimate the **tumor laterality** from the provided pathology report excerpt below.
	Please extract the following four key-value pairs related to the **tumor laterality**:
	1. "laterality stated": <The laterality as explicitly mentioned in the pathology report: "left", "right", "bilateral involvement", or "not primary site/unpaired">,
	2. "laterality estimated": <Your best estimate of laterality based on clinical knowledge. Use "not primary site/unpaired" if unclear or unpaired organ>,
	3. "laterality_CD": <Certainty Degree between 0.00 and 1.00 reflecting confidence in the estimation>,
	4. "laterality_evidence": <A short explanation or quote from the input text that supports your estimation>
	Valid laterality values: ["left", "right", "bilateral involvement", "not primary site/unpaired"]
	Input: `{context}`
	Output Format (JSON Only):
	`{`
	`  "laterality stated": "...",`
	`  "laterality estimated": "...",`
	`  "laterality_CD": ...,`
	`  "laterality_evidence": "..."`
	`}`

D.7 Reviewer Agent Prompt

Table 18: Prompt design for the reviewer agent.

Component	Prompt Text

Return specification	`{`
	`  "grade_valid": true or false,`
	`  "morphology_valid": true or false,`
	`  "morphology_mapped": "..." or null,`
	`  "tnm_valid": true or false,`
	`  "laterality_valid": true or false or null,`
	`  "treatment_valid": true or false or null,`
	`  "comment": "<summarize inconsistencies or mapping decisions>"`
	`}`

Prompt	You are a clinical reviewer. Validate and map the extracted fields from a pathology report. Output **ONLY** a single JSON object; no prose, no markdown, no code fences. Input values: - Tumor Grade: `{grade}` - Morphology: `{morphology}` - TNM Stage: `{tnm}` - Laterality: `{laterality}` - Treatment: `{treatment}` Accepted Grade Values: ["Grade I", "Grade II", "Grade III", "Unknown"] Accepted Morphology Values: ["Adenocarcinoma, NOS", "Carcinoma, NOS", "Follicular adenocarcinoma, NOS", "Infiltrating duct carcinoma, NOS", "Lobular carcinoma, NOS", "Medullary carcinoma, NOS", "Mucinous adenocarcinoma", "Neoplasm, malignant", "Signet ring cell carcinoma", "Unknown"] Accepted TNM Components: - T: ["T0", "Tis", "Tmic", "T1", "T2", "T3", "T4", "Tx", "Unknown"] - N: ["N0", "N1", "N2", "N3", "Nx", "Unknown"] - M: ["M0", "M1", "Mx", "Unknown"] Accepted Laterality Values: ["left, primary organ", "not a paired site/un", "paired, lat. unknown", "right, primary organ", "total colon", "unknown"] Accepted Treatment Values: ["chemotherapy", "chemotherapy + radiotherapy", "chemotherapy + surgery", "chemotherapy + surgery + radiotherapy", "chemotherapy+hormonal", "chemotherapy+radiotherapy", "hormonal", "invalid code.", "neoadjuvant chemotherapy+surgery+radiotherapy+hormonal", "radical prostatectomy", "radiotherapy + surgery", "surgery", "surgery+chemotherapy", "surgery+chemotherapy+hormonal", "surgery+chemotherapy+radiotherapy", "surgery+chemotherapy+radiotherapy+hormonal", "surgery+hormonal", "surgery+radiotherapy+hormonal", "total thyroidectomy", "unknown"] Rules: - If you are unsure, use null for $*_v alid fields that can be unknown, or false when it clearly doesn't match. - DO NOT include any explanations; return JSON only.$ Return JSON exactly in this shape: `{return_spec}`

Component	Prompt Text
Repair prompt	You will receive text that SHOULD be a JSON object with the following keys and types: {return_spec} Fix it if needed and return ONLY a valid JSON object. No comments, no markdown, no code fences. Text: {bad}

D.8 Evaluator Agent Prompt

Table 19: Prompt design for the evaluator agent.

Component	Prompt Text

Prompt	You are a clinical NLP evaluator.
	Your task is to estimate the confidence score (from 0 to 100) for the accuracy of each extracted field from a pathology report: - Tumor **grade** - Tumor **morphology** - **TNM stage** - **Laterality** - **Treatment**
	Base your estimate on: - Whether the extracted value is complete and clinically valid - Whether it follows accepted terminology - Whether it is likely justified based on context
	Input Fields: - Grade: {grade} - Morphology: {morphology} - TNM: {tnm} - Laterality: {laterality} - Treatment: {treatment}
	Output Format (JSON Only):
	`{`
	`  "grade_confidence": integer (0-100),`
	`  "morphology_confidence": integer (0-100),`
	`  "tnm_confidence": integer (0-100),`
	`  "laterality_confidence": integer (0-100),`
	`  "treatment_confidence": integer (0-100)`
	`}`

Role in pipeline	The evaluator agent performs two complementary functions: 1. **Ground-truth comparison (deterministic)**: It compares extracted values against reference annotations (if available) and assigns correctness flags: `grade_correct, morphology_correct, tnm_correct, laterality_correct, treatment_correct` 2. **Confidence estimation (LLM-based)**: If ground truth is partially or fully unavailable, the agent invokes the LLM to assign confidence scores (0–100) for each field. This dual strategy ensures: - objective evaluation when labels exist - robust fallback when labels are missing

D.9 Aggregation Agent Prompt

Table 20: Prompt design for the aggregation agent.

Component	Prompt Text
Prompt	You are a clinical summarization assistant. Your task is to compile a structured report using STRICT JSON format. Important: - DO NOT return markdown or explanation. - DO NOT wrap the JSON in "'`json` or any formatting. - Return ONLY valid JSON exactly matching this structure: {json_spec} Input variables: Grade: {grade}, Stated: {grade_stated}, Estimated: {grade_estimated}, Confidence: {grade_confidence}, Valid: {grade_valid}, Correct: {grade_correct}, Evidence: {grade_evidence} Morphology: {morphology}, Stated: {morphology_stated}, Estimated: {morphology_estimated}, Confidence: {morphology_confidence}, Valid: {morphology_valid}, Correct: {morphology_correct}, Evidence: {morphology_evidence} TNM: {tnm}, Confidence: {tnm_confidence}, Valid: {tnm_valid}, Correct: {tnm_correct} T: Stated: {t_stated}, Estimated: {t_estimated}, Evidence: {t_evidence} N: Stated: {n_stated}, Estimated: {n_estimated}, Evidence: {n_evidence} M: Stated: {m_stated}, Estimated: {m_estimated}, Evidence: {m_evidence} Laterality: {laterality}, Stated: {laterality_stated}, Estimated: {laterality_estimated}, Confidence: {laterality_confidence}, Valid: {laterality_valid}, Correct: {laterality_correct}, Evidence: {laterality_evidence} Treatment: {treatment}, Stated: {treatment_stated}, Estimated: {treatment_estimated}, Confidence: {treatment_confidence}, Valid: {treatment_valid}, Correct: {treatment_correct}, Evidence: {treatment_evidence}

Component	Prompt Text
Repair prompt	You will receive a model output that SHOULD be valid JSON matching this schema: `{json_spec}` The text may contain extra words, code fences, or invalid JSON. Return ONLY a corrected JSON object that strictly conforms to the schema above. No markdown, no explanation. Text to fix: `{bad}`

D.10 Matrix Calculator Agent Prompt

Table 21: Prompt design for the matrix calculator agent.

Component	Prompt Text
Role	LLM-only, ground-truth-driven evaluation utility for the Matrix Calculator app. The closed set is built from ground-truth unique labels for the current site × category. The LLM must choose **one** option from a numbered list and return only the corresponding number. For TNM categories, substages are collapsed to parent labels for both ground truth and valid labels. MRN alignment is preserved without duplicate collapsing.
LLM mapping prompt	You are a medical coding normalizer. Category: `{category}` Select the single BEST match for the raw prediction from the options below. OPTIONS (answer with the number only): `{numbered}` RAW PREDICTION: "`{raw_value}`" Rules: - Output ONLY the number (e.g., 2). No words. - Choose the closest semantic match. - Handle synonyms, abbreviations, Roman numerals, punctuation, and casing. - If meaning is unknown or unclear, choose the option labeled `unknown` when available.
Strict retry prompt	You are a medical coding normalizer. Category: `{category}` Select the single BEST match for the raw prediction from the options below. OPTIONS (answer with the number only): `{numbered}` RAW PREDICTION: "`{raw_value}`" Rules: - Output ONLY the number (e.g., 2). No words. - Choose the closest semantic match. - Handle synonyms, abbreviations, Roman numerals, punctuation, and casing. - If meaning is unknown or unclear, choose the option labeled `unknown` when available. IMPORTANT: Respond with ONLY the number of the chosen option. No other text.
Evaluation logic	For each category (`grade`, `morphology`, `t`, `n`, `m`, `laterality`), the agent: - resolves the corresponding prediction and ground-truth columns, - builds the valid closed set from site-specific ground-truth labels, - maps predictions into that closed set using the LLM, - computes classification metrics using `classification_report`, - computes a confusion matrix using `confusion_matrix`.
TNM parent collapse	For TNM categories, substages are collapsed to parent labels before evaluation. Examples include: - `pT1a` → `t1` - `Tis` → `tis` - `N1c` → `n1` - `M0` → `m0` - unknown or empty values → `unknown`