Type of the Paper: Article

Digital Registrar: A Schema-First Framework for Multi-Cancer Privacy-Preserving Pathology Abstraction via Local LLMs

Nan-Haw Chow, M.D., Ph.D.[1+], Han Chang, M.D., Ph.D.[2,3+], Hung-Kai Chen[2], Chen-Yuan Lin, B.S.[4], Ying-Lung Liu, B.S.[5], Po-Yen Tseng[2], Li-Ju Shiu, Ph.D.[6], Yen-Wei Chu, Ph.D.[7], Pau-Choo Chung, Ph.D.[8], Kai-Po Chang, M.D., Ph.D.[2,3]*

1 Center for Precision Medicine, China Medical University Hospital, Taichung, Taiwan
2 School of Medicine, China Medical University, Taichung, Taiwan
3 Department of Pathology, China Medical University Hospital, Taichung, Taiwan
4 Signal1, Toronto, Canada
5 Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan
6 Department of Foreign Languages and Literature, National Chi Nan University, Nantou, Taiwan
7 Graduate Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung, Taiwan
8 Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

† These authors contributed equally to this work.

* Correspondence: 017179@tool.caaumed.org.tw; Tel.: +886-4-2205-2121 ext. 12678; 2 Yude Rd., Taichung, Taiwan

Abstract

Background/Objectives: Surgical pathology reports contain the most granular diagnostic data for cancer, yet their predominant free-text format creates a "translational gap" that hinders automated registry entry and secondary analytics. While current large language model (LLM) research often focuses on narrow extraction tasks, this study emphasizes the clinically governed schema layer as the more durable scientific contribution for long-term interoperability and reproducibility. Methods: We developed a College of American Pathologists (CAP)-aligned clinical ontology implemented as strictly typed, hierarchical schemas and encoded as DSPy signatures. The system covers 10 major cancer types across 193 registry fields, including complex variable-length structures such as lymph-node groups and surgical margins. An extraction pipeline was built using the DSPy framework, enabling a model-agnostic architecture in which LLMs serve as interchangeable inference engines. Performance was benchmarked on 893 internal pathology reports (2023–2024) and an external validation cohort of 150 TCGA reports. Hardware feasibility was tested on a single 48 GB GPU to ensure suitability for on-premise, privacy-preserving medical workstation deployment. Results: Using the gpt-oss:20b model, the framework achieved a mean exact-match accuracy of 94.3% across all internal registry fields. External validation accuracy remained high at 92.4% in the TCGA cohort, demonstrating robust generalizability across diverse institutional reporting styles. High fidelity was maintained for critical prognostic indicators, including breast cancer biomarkers (near-perfect accuracy) and surgical margin positivity (91.2% mean accuracy). Operationally, gpt-oss:20b provided the best balance of speed (40–70 s/report) and accuracy compared with denser or

more complex architectures. Conclusions: The primary contribution is a schema-first abstraction layer that decouples clinical logic from specific AI models. By transforming narrative reports into machine-readable structures with registry-grade reliability, this framework provides a portable foundation for automated cancer surveillance, privacy-preserving institutional data reuse, and future multimodal clinical systems.

Keywords: clinical ontology; cancer registry; CAP protocol; large language model; structured data extraction; DSPy; interoperability; privacy-preserving AI

1. Introduction

Surgical pathology reports contain the most granular descriptions of cancer diagnosis, staging, margin status, lymph-node involvement, and biomarker findings, making them indispensable for cancer surveillance, quality assessment, and secondary clinical research [1-13]. Yet these data are still documented predominantly as narrative free text. This creates a translational gap between what pathologists report and what registries, analytics pipelines, and downstream clinical systems can computationally reuse. As a result, high-value pathological detail often must be manually re-entered, simplified, or discarded before it becomes available for structured surveillance.

The bottleneck is therefore not only extraction accuracy, but also representation. A durable abstraction system must preserve organ-specific semantics, nested relationships, and variable-length structures such as lymph-node groups, specimen margins, and biomarker panels. Flat field lists or ad hoc prompt outputs are insufficient because they frequently collapse clinically meaningful context, making results harder to validate, compare across institutions, and integrate into longitudinal data infrastructures.

Recent large language model (LLM) studies have shown encouraging performance for oncology information extraction, but most have focused on narrow tasks, limited variable sets, or model-specific demonstrations [14-28]. These advances are important, yet model capabilities evolve rapidly and implementation details can age quickly. In contrast, the clinical ontology that governs what should be abstracted, how values are typed, and how repeating structures are organized is the more durable scientific contribution. A schema-first design can therefore outlast any single generation of models while improving reproducibility and interoperability.

The College of American Pathologists (CAP) protocols provide an appropriate foundation for such a design because they define stable, clinically governed data elements for cancer reporting. Translating these standards into strictly typed hierarchical schemas makes it possible to encode registry logic explicitly, constrain outputs programmatically, and separate clinical representation from the inference engine itself. This separation is especially important for privacy-preserving deployment, because institutions can retain on-premise control of protected health information while updating local models over time without rewriting the clinical abstraction layer.

Accordingly, we present Digital Registrar, a schema-first framework for privacy-preserving pathology abstraction using local LLMs. Rather than treating structured extraction solely as a model-performance problem, the framework implements CAP-aligned clinical ontologies across ten major cancer types and 193 registry fields, including complex variable-length structures, within a model-agnostic DSPy pipeline executed on a single 48 GB GPU. We evaluate the framework on 893 internal pathology reports and an

external TCGA cohort, demonstrating registry-grade accuracy while establishing a portable foundation for automated cancer surveillance, institutional data reuse, and future multimodal clinical systems [29-33].

2. Materials and Methods

2.1. Data Source and Annotation

This study was conducted under research ethics committee approval (REC #CMUH114-REC2-037) from China Medical University Hospital and in accordance with all relevant guidelines and regulations. The dataset comprised a total of 893 de-identified surgical pathology reports from the years 2023–2024. The final validation cohort consisted of 694 reports describing resected, invasive primary malignancies across ten major organ systems, with the detailed composition summarized in Table 4. An additional 199 reports describing biopsies, non-invasive tumors, or benign conditions served as negative controls for the initial classification stages of the pipeline.

Table 4. Composition of the cancer excision cohort (n = 694).

Organ site	Number of reports (n)
Esophagus	75
Stomach	80
Colorectum	71
Breast	76
Uterine cervix	30
Pancreas	29
Prostate	95
Lung	64
Liver	80
Thyroid	75
Others (dual or out-of-scope cancers)	30
Total	694

To establish an expert-verified gold standard for evaluation, a semi-automated annotation process was employed. First, an initial structured JSON summary was generated for each report using a single inference pass of the gpt-oss:20b model. Subsequently, two board-certified pathologists independently reviewed and corrected these preliminary outputs in a side-by-side interface displaying the original report text and the structured data. All discrepancies between the two reviewers were resolved through a final blinded consensus, producing the definitive reference dataset used for all performance evaluations.

2.2. CAP-Aligned Clinical Ontology and Schema Definition

All data extracted by the system conformed to a CAP-aligned clinical ontology implemented as a strictly typed, hierarchical JSON schema designed for clinical precision and computational reliability. The ontology comprised two layers: (1) a universal core layer defining attributes common to all surgical pathology reports, such as report eligibility and primary tumor site, and (2) ten organ-specific extension layers capturing the detailed variables unique to each malignancy.

Each organ-specific ontology module was developed in strict alignment with the College of American Pathologists (CAP) Cancer Protocol data elements and organized into reusable semantic groups, including procedure, histology, grade, stage, lymph-node structures, margin structures, and biomarker submodules where applicable. This schema design preserved nesting, cardinality, and field-level typing so that LLM outputs could be validated deterministically and compared directly with expert reference annotations.

2.3. Ontology-Guided Extraction Workflow

The deterministic extraction workflow was implemented in DSPy 2.1 using dspy.Predict() modules with predefined dspy.Signature specifications that encoded essential clinical instructions and ontology constraints. The pipeline was executed on-premises using the gpt-oss:20b model served via Ollama.

No model training, fine-tuning or few-shot prompting was applied; each report was processed in a single forward pass.

2.3.1 Eligibility classification

The EligibilitySignature predicts whether a report represents a cancer surgery (cancer_excision_report: bool), the organ category (cancer_category), and a free-text descriptor when “others”.

2.3.2 Organ detection

If cancer_excision_report = True, a secondary signature (CancerTypeSignature) verifies or refines the organ label prior to extraction.

2.3.3 Organ-specific extraction

Ten sets of organ-specific DSPy modules were used for organ-specific extraction. For example, the breast cancer pipeline included BreastCancerNonnested, BreastCancerMargins, BreastCancerLN, and BreastCancerBiomarkers. These modules receive both the original pathology report and the JSON summary previously generated by gpt-oss:20b as input and output validated JSON matching their respective schemas. The pipeline is illustrated in Figure 3. The modules extract tumor characteristics, AJCC/TNM staging, surgical margins, and lymph-node details for each cancer type. For breast and colorectal cancer, biomarker status is also extracted. Readers are referred to Supplementary Appendix S25 for examples of model output.

2.3.4 Validation of Model Agnosticism and Hardware Feasibility

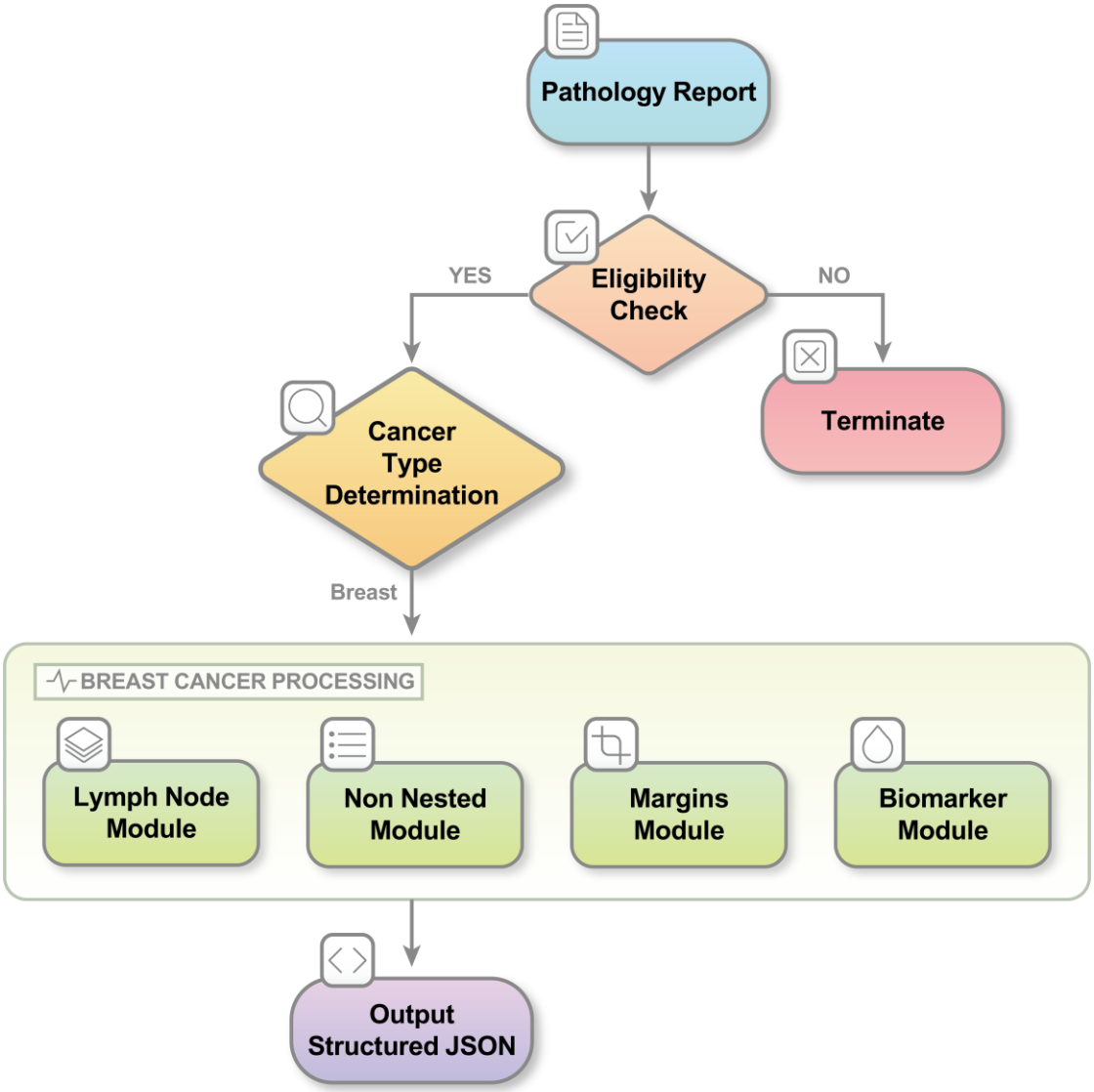
To validate the architectural independence of the extraction framework and assess the feasibility of deployment on standard workstation hardware, the identical pipeline structure was instantiated and executed using two alternative open-weight transformer architectures: qwen3-30b (Sparse Mixture-of-

Experts) and gemma3:27b (Dense). The foundational design of the AI pipeline explicitly embraces principles of model agnosticism, utilizing the DSPy 2.1 framework to decouple the extraction logic from specific model weights.

All comparative models were executed on-premises using the same dedicated NVIDIA RTX A6000 Ada Generation GPU (48 GB VRAM). This hardware constraint was strictly enforced to evaluate whether models could operate efficiently within the memory envelope of a standard high-end medical workstation, without requiring multi-GPU datacenter configurations.

The overall ontology-guided extraction workflow is summarized in Figure 3.

Figure 3. Ontology-guided extraction pipeline for structured cancer data from surgical pathology reports.



2.4. Evaluation Metrics

Model performance was evaluated by comparing each generated JSON output against the expert-verified reference standard. Given that all schema variables are strictly typed (i.e., Literal, bool, int, or None),

evaluation was conducted using deterministic, exact-string matching rather than similarity-based metrics. For list-type variables, such as surgical margins and lymph node stations, corresponding elements were aligned by field name before comparison. All statistical analyses were performed using the scikit-learn library (version 1.5.0) in Python 3.11.

The primary metrics included accuracy, precision, recall, and the F1 score. The F1 score is calculated as the harmonic mean of precision and recall, providing a single, balanced measure of a test's accuracy. For multiclass classification tasks, such as the organ-type identification performed by the CancerTypeSignature classifier, the macro-average F1 score was computed as the unweighted mean of the F1 scores for each class, thereby giving equal weight to each category regardless of its frequency. Overall accuracy was defined as the proportion of all extracted fields that exactly matched the reference standard across the validation cohort. The specific definitions for composite metrics applied to the extraction modules, such as margin status correctness and total lymph node accuracy, are detailed in Table 5.

Table 5. Evaluation metrics and definitions.

Metric	Definition
Margin status correctness	Exact match of both margin_category and margin_involved for every reported margin element.
Margin distance correctness	Exact match of margin_category, margin_involved, and distance (integer, mm) for each margin element.
Total lymph-node accuracy	Exact match of examined and involved integer counts within each organ's regional_lymph_node list.
Lymph-node station accuracy	Exact match of station_name, examined, and involved values for each lymph-node group ("station").
Accuracy of other fields	Proportion of all remaining scalar fields (e.g., histological_type, tumor_size, grade, biomarker results, and stage-group entries) that exactly match the expert reference.

2.5. Independent External Validation

To assess the robustness and transportability of the extraction model to external, non-institutional data, we utilized the TCGA-Reports dataset [33] from The Cancer Genome Atlas (TCGA) repository. This dataset serves as a rigorous gold standard for external validation, comprising machine-readable pathology reports from 32 tissues/cancer types generated by a wide variety of institutions, thereby ensuring high structural and linguistic heterogeneity.

The model was applied in an "off-the-shelf" capacity, without any further training or fine-tuning, to a validation subset of 150 reports spanning Breast, Colorectal, Esophagus, Stomach, and "Others" categories. We evaluated both the classification (triage) performance and the extraction fidelity for 53 specific pathological fields across the Breast (22 fields), Colorectal (17 fields), and Esophagus (14 fields) cohorts.

3. Results

3.1. Assessment of Model Agnosticism and Computational Feasibility

To determine the optimal engine for clinical deployment, we benchmarked the extraction framework across three distinct open-weight architectures (gpt-oss:20b, qwen3-30b-A3B, and gemma3:27b). The objective was to validate the model-agnostic nature of the pipeline while identifying the configuration

that best balances registry-grade accuracy with the hardware constraints of a standard medical workstation (single GPU, 48GB VRAM).

As shown in Table 1, the framework achieved high exact-match accuracy across all tested architectures, confirming its agnosticism. However, a decisive divergence emerged in operational efficiency. Although all three models operated within the workstation-class envelope (e.g., a 48 GB VRAM workstation GPU), the fastest model by a clear margin was gpt-oss:20b, which maintained the highest accuracy (94.3%) while processing complex reports in roughly 40–70 seconds, about 2–3× faster than its alternatives.

The performance gap is not primarily due to simple VRAM capacity exhaustion; rather, it reflects the interplay of model architecture, total parameter state, expert routing overhead (in the case of mixture-of-experts (MoE) models), and memory/bandwidth traffic. For instance, the qwen3-30b-A3B model, despite activating only ~3.3 billion parameters per token, has a total parameter size of ~30.5 billion and uses 128 experts with 8 activated per token [31]. These aspects increase system overhead and latency, positioning it behind gpt-oss:20b in our single-GPU workstation tests.

Consequently, based on this efficiency and accuracy analysis, gpt-oss:20b was identified as the optimal candidate for the "Digital Registrar" framework, delivering registry-grade accuracy (94.3%) with high throughput on standard workstation hardware. For this reason, all subsequent results in this study—including eligibility triage, organ classification, and granular field extraction (Sections 3.2–3.6)—are derived exclusively from the gpt-oss:20b pipeline.

Table 1. Comparative performance and efficiency of locally deployed LLMs.

Model	Architecture Type	Total Parameters	Mean Exact-Match Accuracy (193 Registry Fields)	Typical Processing Speed(Per Complex Report)
gpt-oss:20b	Sparse MoE	20B	94.30%	40 – 70 seconds
qwen3-30b-A3B	Sparse MoE	30B	92.90%	140 – >200 seconds
gemma3:27b	Dense	27B	89.80%	50 – 130 seconds

3.2. Automated Report Triage and Organ Classification

To ensure that the pipeline selectively processes valid cancer registry cases, the system was first evaluated on its ability to triage unstructured pathology reports, distinguishing eligible cancer excision cases from a background of biopsies and non-malignant specimens. As detailed in Table 2, the eligibility classifier demonstrated robust reliability across the full validation cohort (N = 893). Crucially for a screening tool, the model exhibited near-perfect sensitivity; out of 694 eligible cancer excision reports, only a single breast excision case was missed (false negative), confirming its safety as an automated filter. Specificity was similarly high, with most discrepancies (29 false positives) restricted to clinically ambiguous cases, such as large metastatic deposits or extensive biopsies, that mimic the linguistic complexity of primary resections. These results confirm that the model operates with the sensitivity required to ensure that virtually no eligible cancer surgery report is overlooked during the initial screening process.

The subsequent organ-labeling sub-model demonstrated similarly strong performance, achieving F1 scores ≥ 0.955 across all ten target organs, with an average accuracy of 0.978 and a macro-average F1 score of 0.955. Eight organ classes—breast, cervix, colorectal, liver, pancreas, prostate, stomach, and thyroid—achieved perfect recall (1.000), while both lung and esophagus reached F1 ≥ 0.955. Misclassifications were rare and distinctively patterned: qualitative review revealed these were often cases with double primary malignancies—such as reports describing independent tumors in both lung lobes—which were intentionally excluded from the automated staging logic (ground truth "Others") but were occasionally interpreted by the model as belonging to the specific organ class (e.g., "Lung"). Collectively, these results confirm that the foundational classification layers of the DSPy pipeline operate with registry-grade reliability, providing clean input for downstream extraction. Detailed confusion matrices for all three test runs are provided in Supplementary Tables S12–S14.

Table 2. Confusion matrix of the eligibility triage classifier (run 1).

	Predicted Eligible (Cancer Excision)	Predicted Ineligible (Biopsy / Benign / Other)	Total
Actual Eligible	693 (TP)	1 (FN)	694
Actual Ineligible	29 (FP)	170 (TN)	199
Total	722	171	893

3.3. Per-Organ Extraction Accuracy across Ontology Fields

After classification, cancer surgery reports were parsed through organ-specific extraction modules, each aligned to a JSON schema tailored to its organ system (including procedure, histology, grade, and AJCC T/N/M categories; see Supplementary Tables S14–S23 for details). Exact-match accuracies across all ten organ systems (193 fields in total) ranged from 88.9% to 96.2%, with an overall mean of 94.3% ± 8.32%. Table 3 summarizes the cohort size, field count, and mean performance for each organ. Surgical margins, lymph nodes, and biomarkers were extracted by separate modules and are discussed in later sections.

High-performing fields (≥95%) were consistent across organs, except for thyroid, where frequent references to other reports caused some model difficulty. These high-performing fields included procedure, histology, grade, and the presence of vascular invasion, indicating that the pipeline captures core oncologic variables with near-perfect fidelity even in organs with extensive schema complexity (≥24 fields).

In contrast, performance on TNM staging fields showed greater variability across organs, reflecting the inherent complexity and inter-organ heterogeneity of staging schemas. Occasional ambiguities in clinical phrasing—such as “anatomic staging” and “pathologic staging” in breast—also contributed to minor discrepancies. A representative example is provided in Supplementary Appendix S25.3 (Sample S25.3.1–S25.3.2). Nevertheless, overall accuracy for T, N, and M fields remained high, typically exceeding 90% across test runs.

Lower-accuracy fields (60–90%) were limited to poorly documented or quantitative attributes, such as surgical technique in stomach and cervix reports or tumor percentage and prostate weight in prostate

reports, where textual variability is high. A representative example and additional discussion are provided in Supplementary Appendix S25.4 (Samples S25.4.1–S25.4.2).

In summary, the organ-specific modules sustained ≥90% mean exact-match accuracy across varying schema sizes, reaffirming the scalability of the one-shot LLM pipeline for comprehensive extraction of structured oncologic data. Full per-field extraction results for each organ, including exact-match accuracy for all schema items, are provided in Supplementary Tables S15–S24.

Table 3. Per-organ extraction accuracy across registry fields.

Organ	n Reports	# Fields Evaluated	Mean Accuracy ± SD (%)
Breast	75	26	93·5 ± 9·7
Uterine cervix	29	16	90.9 ± 9.7
Colon and rectum	70	20	96.6 ± 5.3
Esophagus	73	16	96.2 ± 4.8
Liver – HCC	79	17	94.1 ± 9.0
Lung	63	20	92.5 ± 7.1
Pancreas	28	15	91.7 ± 12.9
Prostate	94	25	97.0 ± 3.4
Stomach	79	18	93.2 ± 11.6
Thyroid	74	20	92.3 ± 9.5

3.4. Lymph-Node and Surgical Margin Status

The pipeline demonstrated strong performance in extracting prognostically critical data on lymph node (LN) and surgical margin status, as summarized in Figures 1 and 2. For LN evaluation, the system achieved a mean total-node concordance of over 80% across all ten organ systems. Performance was highest in organs with well-defined nodal basins such as the breast and liver, while it was more variable in anatomically complex sites like the lung and esophagus, where occasional omissions were the primary source of error rather than misclassification. Detailed per-organ LN metrics, including both total and station-level analyses, are provided in Supplementary Table S25. Examples of LN extraction performance are provided in Supplementary Appendix S25.5 (Samples S25.5.1.1–S25.5.2.2).

For surgical margins, the model achieved a mean accuracy of 91.2% for determining margin positivity. Concordance was excellent (≥95%) for several major cancers, including breast, colorectal, and lung. In cases with lower accuracy, such as pancreas resections, discrepancies were primarily attributable to challenges in mapping variably named margin categories rather than incorrect identification of tumor involvement. As expected, accuracy for extracting the exact margin distance was lower, reflecting the inherent ambiguity of inequality notations and unit variations in the source reports. Representative margin-status extraction results that show the complexity of pancreatic cancer are provided in Supplementary Appendix S25.6 (Samples S25.6.1–S25.6.2). Because prostate cancer reports do not define multiple margin stations and instead document margin involvement using only overall positivity, extent of involvement, and the specific anatomic sites involved, prostate margins were excluded from this station-based margin accuracy evaluation.

Figure 1. Lymph-node extraction accuracy across organ systems. Bar plots show total lymph-node accuracy and station-level accuracy for each organ system.

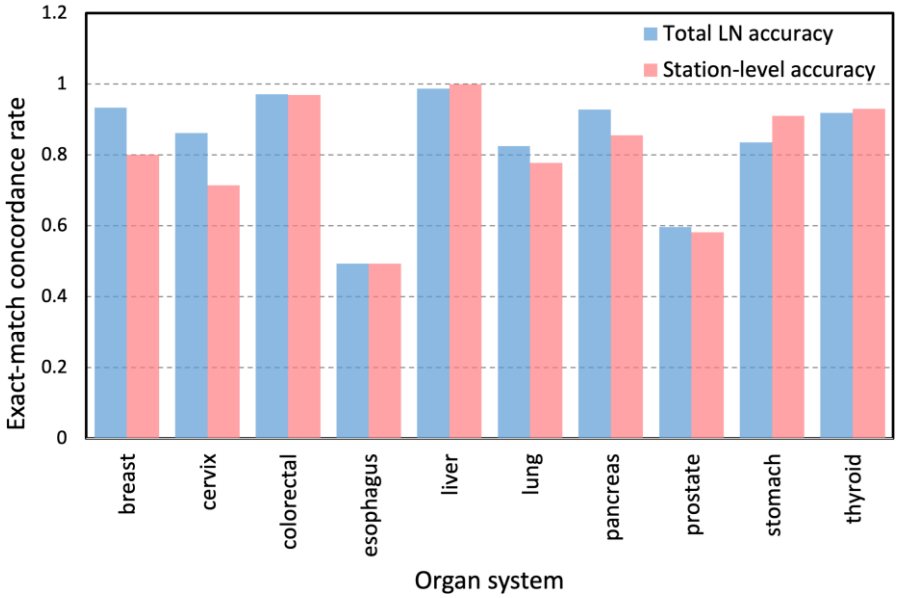
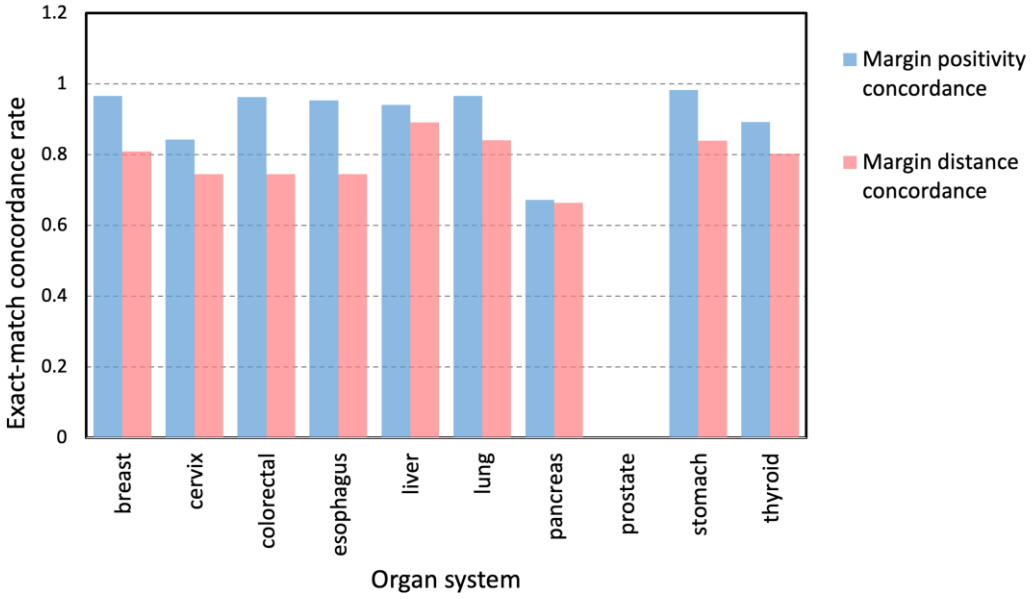


Figure 2. Surgical margin extraction accuracy across organ systems. Bar plots compare margin-positivity concordance and margin-distance concordance across organ systems.



3.5. Breast Biomarker Extraction Performance

Breast cancer biomarker fields were extracted with near-perfect accuracy across 75 reports. ER positivity and percentage were correct in 75/75 and 73/75 cases, respectively. PR positivity and percentage were correct in 74/75 and 73/75 cases. HER2 score was correctly extracted in all 75 cases, and Ki-67 percentage

in 74/75 cases. These results confirm the robustness of field-level extraction and validation under exact string matching.

3.6. External Validation Performance (TCGA Cohort)

In the external TCGA cohort, the model maintained performance comparable to that in the internal validation set, confirming strong cross-domain robustness. Across 150 reports, triage and organ classification achieved 97.3% accuracy, with near-perfect discrimination for major cancers (F1 = 1.000 for colorectal; 0.990 for esophagus). The few discrepancies were administrative rather than interpretative: three reports labeled as "Others" in the ground truth were predicted as breast cancer but were subsequently verified as synchronous bilateral breast cancers—cases intentionally excluded from organ-level staging due to the system's single-primary design. Field-level extraction across 53 pathological variables reached a mean exact-match accuracy of 92.4% ± 11.0%, closely mirroring internal-set performance (93.9% ± 8.6%) and demonstrating that the extraction logic generalizes effectively beyond the institution where it was developed.

4. Discussion

This study demonstrates that a clinically governed ontology, paired with schema-bound LLM execution, can convert free-text surgical pathology reports into registry-grade structured data at scale. Importantly, the core contribution is not only the use of local LLMs for extraction, but the schema-first abstraction layer that preserves pathology-specific semantics across cancers. The pipeline achieved high exact-match accuracy across ten malignancies and 193 fields while remaining feasible on a single workstation-class GPU, supporting practical deployment for document triage, organ classification, and fine-grained abstraction.

Although LLMs now show strong performance on clinical extraction tasks, the field has faced both a computational barrier and a representational barrier [24,34]. The computational barrier concerns privacy-preserving local deployment; the representational barrier concerns the lack of a durable schema layer capable of preserving nested pathology semantics while remaining portable across models. Our CAP-aligned ontology addresses the latter problem by defining a reusable clinical target that can outlast rapid changes in model architecture.

The framework's demonstrated capacity for model agnosticism—validated by the seamless interchange of three models (gpt-oss:20b, qwen3-30b-A3B, and gemma3:27b)—provides crucial flexibility for healthcare deployment. All three operate within a workstation-class envelope (e.g., a 48 GB GPU such as the NVIDIA RTX A6000) rather than requiring datacenter-scale infrastructure. Nonetheless, our comparative analysis shows that even within this workstation context, deployment efficiency diverges sharply. While all models are technically viable, gpt-oss:20b achieved the fastest inference (≈40–70 seconds) in our single-GPU tests. In contrast, qwen3-30b-A3B recorded substantially higher latency (≈140 to >200 seconds) despite its low active-parameter count (~3 billion), because its 30 billion-parameter total state plus MoE routing and memory-traffic overhead pushed the system toward bandwidth and interconnect limits. gemma3:27b (≈27 billion parameters) fell in between (≈50–130 seconds)—still deployable on workstation-class hardware, but slower than gpt-oss:20b and with lower accuracy. The key takeaway is that hardware capacity is not simply a matter of "sufficient" or "insufficient"; architecture

(MoE vs. dense), active-versus-total parameters, memory traffic, and routing overhead all combine to determine practical performance.

In the use case of extracting structured fields from complex reports, as in our healthcare setting, the superior accuracy of gpt-oss:20b matters materially: it maintained the highest exact-match accuracy in our benchmarks. Meanwhile, although gemma3:27b's latency was comparable in some cases, its accuracy lagged behind. Notably, when extraction tasks demand high fidelity, MoE-based models (gpt-oss:20b and qwen3-30b-A3B) may offer advantages, as their smaller active-parameter footprint can reduce memory overhead and potentially support more targeted inference. Thus, while gpt-oss:20b currently presents the best balance of speed and accuracy, qwen3-30b-A3B's MoE design may still merit consideration in contexts where higher model capacity might pay off and latency constraints are less stringent.

The independent external validation results confirm the model's stability when applied to a structurally and institutionally diverse patient cohort. The total mean extraction accuracy of 92.4% in the external TCGA cohort is quantitatively similar to the 94.3% reported for the internal development cohort. This finding is significant as it indicates that the model has learned fundamental pathological principles rather than mimicking institutional reporting artifacts, retaining high performance without the need for site-specific fine-tuning.

While overall performance was high, error analysis highlighted challenges inherent to multi-institutional data curation. The highest variability was observed in fields such as surgical_technique in colorectal cancer (48.0% accuracy) and pm_category (distant metastasis) in esophageal cancer (62.5% accuracy). These discrepancies largely reflect known limitations in the ground-truth definitions within the TCGA metadata (e.g., inconsistent M1 status annotations) rather than model failure. Furthermore, classification errors were largely attributable to complex bilateral breast cancer cases which, although correctly identified by the model as breast tissue, fell outside the current single-primary schema scope. On core oncologic variables such as histology and grade, the model maintained 100% accuracy, supporting its utility as a reliable "digital registrar."

A key future direction is the extension of this framework into a multimodal system. While our system excels at abstracting data from textual reports, the "digital registrar" of the future must integrate information from various data streams to build a comprehensive patient model. The integration of genomic data and digital pathology images, as highlighted in recent reviews of multimodal AI in oncology, represents the next frontier in automated cancer surveillance [35]. The development of pathology-specific multimodal LLMs, such as PathChat+, which can reason directly over high-resolution pathology images, represents a pivotal technological enabler for this vision [36]. A future system built on our framework could leverage such models to correlate histologic patterns with molecular markers and clinical outcomes at an unprecedented scale, moving beyond mere data extraction to true integrated diagnostics and prognostication [37].

5. Conclusions

By combining a CAP-aligned clinical ontology with schema-bound large language model execution, this work provides a practical route from narrative pathology reports to interoperable, machine-readable cancer data. The ontology—not any single model family—is the durable contribution: it preserves

clinically meaningful structure, enables privacy-preserving local deployment, and provides a portable foundation for future multimodal and real-time cancer data infrastructures.

Supplementary Materials

Supplementary materials referenced in the manuscript include Supplementary Tables S12–S25 and Appendix S25, which contain detailed per-field results, confusion matrices, and representative extraction examples.

Author Contributions

Conceptualization, K.-P.C.; methodology, K.-P.C., H.-K.C. and P.-C.C.; software, H.-K.C., C.-Y.L. and Y.-L.L.; validation, H.-K.C., P.-Y.T., N.-H.C. and H.C.; formal analysis, H.-K.C. and Y.-W.C.; investigation, N.-H.C., H.C. and K.-P.C.; resources, N.-H.C. and H.C.; data curation, P.-Y.T.; writing—original draft preparation, K.-P.C.; writing—review and editing, K.-P.C., P.-C.C., Y.-W.C. and L.-J.S.; supervision, N.-H.C., H.C. and K.-P.C.; funding acquisition, K.-P.C. and H.-K.C. All authors have read and agreed to the published version of the manuscript. Nan-Haw Chow and Han Chang contributed equally to this work.

Funding

This research was funded by the National Science and Technology Council (NSTC) of Taiwan, grant numbers NSTC 113-2221-E-039-017 and NSTC 114-2813-C-039-132-E, and by China Medical University Hospital, grant number DMR-115-078. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of China Medical University & Hospital (protocol code CMUH114-REC2-037).

Informed Consent Statement

Patient consent was waived by the Research Ethics Committee of China Medical University & Hospital due to the retrospective nature of the study and the use of de-identified datasets.

Data Availability Statement

The internal development cohort consists of de-identified surgical pathology reports from China Medical University Hospital and is not publicly available because of patient privacy protections, but may be made available to qualified researchers upon reasonable request to the corresponding author and subject to an appropriate data use agreement. The external validation cohort, including the 150 selected TCGA pathology reports, expert-annotated ground-truth labels, and model outputs, has been deposited in Zenodo (https://doi.org/10.5281/zenodo.17689361); the original source data remain accessible via the NCI Genomic Data Commons Data Portal. The source code for the Digital Registrar extraction pipeline, including DSPy signatures, organ-specific schema definitions, and orchestration logic, is openly available at https://github.com/kblab2024/digitalregistrar; a permanent Zenodo DOI will be minted upon acceptance.

Acknowledgments

We are deeply grateful to Dr. Ting-Sung Hsieh of UT Southwestern Medical Center, as well as Dr. Ching-Hang Chien and Dr. Yi-Chung Dzeng of the Research Center for Applied Sciences, Academia Sinica. Their insightful feedback and critical suggestions were invaluable to the intellectual development and final refinement of this manuscript. During the preparation of this manuscript, the authors used OpenAI GPT-5 for non-confidential language editing and drafting support. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Piñeros, M.; et al. The role and utility of population-based cancer registries in cervical cancer surveillance and control. Prev. Med. 2021, 144, 106237, doi:10.1016/j.ypmed.2020.106237.
2. The Lancet. Cancer registries: the bedrock of global cancer care. Lancet 2025, 405, 353, doi:10.1016/S0140-6736(25)00189-8.
3. Nair, S.; Ngwa, W.; Addai, B.W.; Addai, A.O.; Oti, B.A. The role of a national cancer registry in effective implementation and monitoring of the Global Breast Cancer Initiative (GBCI) framework in Ghana: a narrative review. Transl. Breast Cancer Res. 2025, 6, 20, doi:10.21037/tbcr-24-60.
4. Gondara, L.; Simkin, J.; Arbour, G.; Devji, S.; Ng, R. Classifying tumor reportability status from unstructured electronic pathology reports using language models in a population-based cancer registry setting. JCO Clin. Cancer Inform. 2024, 8, e2400110, doi:10.1200/CCI.24.00110.
5. Hands, I.; Kavuluru, R. A survey of NLP methods for oncology in the past decade with a focus on cancer registry applications. Artif. Intell. Rev. 2025, 58, 314, doi:10.1007/s10462-025-11316-5.
6. Hailer, L.; Miller, J.; Chapman, S. Cancer Registrar Workload and Staffing Study: Guidelines for Hospital Cancer Registry Programs. J. Registry Manag. 2024, 51, 175–176.
7. Penberthy, L.T.; Rivera, D.R.; Lund, J.L.; Bruno, M.A.; Meyer, A.-M. An overview of real-world data sources for oncology and considerations for research. CA Cancer J. Clin. 2022, 72, 287–300, doi:10.3322/caac.21714.
8. Moor, M.; et al. Foundation models for generalist medical artificial intelligence. Nature 2023, 616, 259–265, doi:10.1038/s41586-023-05881-4.
9. Singhal, K.; et al. Large language models encode clinical knowledge. Nature 2023, 620, 172–180, doi:10.1038/s41586-023-06291-2.
10. Tertulino, R.; Antunes, N.; Morais, H. Privacy in electronic health records: a systematic mapping study. J. Public Health (Berl.) 2024, 32, 435–454, doi:10.1007/s10389-022-01795-z.
11. Hernandez-Boussard, T.; Blayney, D.W.; Brooks, J.D. Leveraging digital data to inform and improve quality cancer care. Cancer Epidemiol. Biomarkers Prev. 2020, 29, 816–822, doi:10.1158/1055-9965.EPI-19-0873.
12. Bagherzadeh, P.; Sultanem, K.; Batist, G.; Abbasinejad Enger, S. An automatic pipeline for temporal monitoring of radiotherapy-induced toxicities in head and neck cancer patients. npj Precis. Oncol. 2025, 9, 40, doi:10.1038/s41698-025-00824-w.

13. Paiva, B.; et al. A new natural language processing-inspired methodology (detection, initial characterization, and semantic characterization) to investigate temporal shifts (drifts) in health care data: Quantitative study. JMIR Med. Inform. 2024, 12, e54246, doi:10.2196/54246.
14. Saluja, R.; et al. Cancer type, stage and prognosis assessment from pathology reports using LLMs. Sci. Rep. 2025, 15, 27300, doi:10.1038/s41598-025-10709-4.
15. Chen, D.; Alnassar, S.A.; Avison, K.E.; Huang, R.S.; Raman, S. Large language model applications for health information extraction in oncology: Scoping review. JMIR Cancer 2025, 11, e65984, doi:10.2196/65984.
16. Hung, T.K.W.; et al. Human vs. AI: A comparative effectiveness study of large language models for automated biomarker extraction. J. Clin. Oncol. 2025, 43, 16_suppl, e13605, doi:10.1200/JCO.2025.43.16_suppl.e13605.
17. Wiest, I.C.; et al. A software pipeline for medical information extraction with large language models, open source and suitable for oncology. npj Precis. Oncol. 2025, 9, 313, doi:10.1038/s41698-025-01103-4.
18. Bartels, S.; Carus, J. From text to data: Open-source large language models in extracting cancer related medical attributes from German pathology reports. Int. J. Med. Inform. 2025, 203, 106022, doi:10.1016/j.ijmedinf.2025.106022.
19. Wiest, I.C.; et al. Privacy-preserving large language models for structured medical information retrieval. npj Digit. Med. 2024, 7, 257, doi:10.1038/s41746-024-01233-2.
20. Bednarczyk, L.; et al. Scientific evidence for clinical text summarization using large language models: Scoping review. J. Med. Internet Res. 2025, 27, e68998, doi:10.2196/68998.
21. Tripathi, A.; et al. Employing consensus-based reasoning with locally deployed LLMs for enabling structured data extraction from surgical pathology reports. medRxiv 2025, 2025.04.22.25326217, doi:10.1101/2025.04.22.25326217.
22. Bumgardner, V.K.C.; Mullen, A.; Armstrong, S.; Hickey, C.; Talbert, J. Local large language models for complex structured medical tasks. arXiv 2023, arXiv:2308.01727, doi:10.48550/arXiv.2308.01727.
23. Lee, D.; et al. Using large language models to automate data extraction from surgical pathology reports: Retrospective cohort study. JMIR Form. Res. 2025, 9, e64544, doi:10.2196/64544.
24. Balasubramanian, J.B.; Adams, D.; Roxanis, I.; Berrington de Gonzalez, A.; Coulson, P.; Almeida, J.S.; García-Closas, M. Leveraging large language models for structured information extraction from pathology reports. J. Pathol. Inform. 2025, 19, 100521, doi:10.1016/j.jpi.2025.100521.
25. Grothey, B.; et al. Comprehensive testing of large language models for extraction of structured data in pathology. Commun. Med. 2025, 5, 96, doi:10.1038/s43856-025-00808-8.
26. Rajaganapathy, S.; et al. Synoptic reporting by summarizing cancer pathology reports using large language models. npj Health Syst. 2025, 2, 11, doi:10.1038/s44401-025-00013-8.
27. Hao, Y.; et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. npj Digit. Med. 2025, 8, 450, doi:10.1038/s41746-025-01824-7.
28. Atil, B.; et al. Non-determinism of "deterministic" LLM settings. arXiv 2024, arXiv:2408.04667, doi:10.48550/arXiv.2408.04667.
29. Khattab, O.; et al. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
30. Kumar, D.; Yadav, D.; Patel, Y. GPT-OSS-20B: A comprehensive deployment-centric analysis of OpenAI's open-weight mixture of experts model. arXiv 2025, arXiv:2508.16700, doi:10.48550/arXiv.2508.16700.

31. Qwen Team. Qwen3 technical report. arXiv 2025, arXiv:2505.09388, doi:10.48550/arXiv.2505.09388.
32. Gemma Team. Gemma 3 technical report. arXiv 2025, arXiv:2503.19786, doi:10.48550/arXiv.2503.19786.
33. Kefeli, J.; Tatonetti, N. TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models. Patterns 2024, 5, 100933, doi:10.1016/j.patter.2024.100933.
34. Dennstädt, F.; Hastings, J.; Putora, P.M.; Schmerder, M.; Cihoric, N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. npj Digit. Med. 2025, 8, 143, doi:10.1038/s41746-025-01476-7.
35. Waqas, A.; Tripathi, A.; Ramachandran, R.P.; Stewart, P.A.; Rasool, G. Multimodal data integration for oncology in the era of deep neural networks: a review. Front. Artif. Intell. 2024, 7, 1408843, doi:10.3389/frai.2024.1408843.
36. Chen, C.; et al. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. arXiv 2025, arXiv:2506.20964, doi:10.48550/arXiv.2506.20964.
37. Witowski, J.; et al. Multi-modal AI for comprehensive breast cancer prognostication. arXiv 2024, arXiv:2410.21256, doi:10.48550/arXiv.2410.21256.
38. Torous, V.F.; et al. College of American Pathologists Cancer Protocols: From optimizing cancer patient care to facilitating interoperable reporting and downstream data use. JCO Clin. Cancer Inform. 2021, 5, 47–55, doi:10.1200/CCI.20.00104.