Automating the cancer registry: An Autonomous, Resource-Efficient AI for Multi-Cancer Data Abstraction from Pathology Reports

Nan-Hua Chou, 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-Chu Hsu, Ph.D.[6], Yen-Wei Chu, Ph.D.[7], 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

[+]These authors contributed equally to this work

*** Corresponding Author:**

Kai-Po Chang

017179@tool.caaumed.org.tw

+886-4-2205-2121 #12678

2 Yude Rd., Taichung, Taiwan

Abstract

Background The manual abstraction of data from unstructured pathology reports for cancer registries is a major bottleneck in oncology informatics. While Large Language Models (LLMs) show promise, their adoption is hindered by computationally prohibitive resource requirements and data privacy concerns. There is an urgent need for an accessible, robust, and autonomous system deployable at the front lines of healthcare.

Methods We developed a fully autonomous, multi-stage AI workflow that operates on a single professional-grade GPU. The system first deploys a detector that triages raw clinical documents to identify cancer surgical reports requiring cancer registry entry, and a classifier that routes pathology reports to one of ten cancer-specific information extraction engines. This privacy-preserving pipeline is powered by a locally deployed open-weight LLM (gpt-oss:20b) and programmed via the DSPy framework, a self-optimizing framework for LLMs. The end-to-end system was validated on 893 uncurated, real-world cancer-surgical reports, with a gold standard generated by two board-certified pathologists.

Findings The triage detector achieved an accuracy of 96.6% and an F1 score of 97.9% on cancer surgical reports detection, and the cancer-type classifier routed reports with an accuracy of 97.8% and a macro average F1 score of 95.5%. Across ten distinct malignancies, the pipeline achieved a mean exact-match accuracy of 93.9%±8.6% for 196 registry fields. Critical oncologic information, including margin positivity and key

biomarkers for breast cancer, were extracted with acceptable fidelity (90–100%). The entire autonomous process was executed at an acceptable speed.

Interpretation This work provides a deployable, resource-efficient blueprint for a "digital registrar" robust enough for real-world clinical workflows. This marks a pivotal step toward transforming global cancer surveillance from a slow, retrospective process into a near real-time system.

Funding National Science and Technology Council (NSTC) of Taiwan

1 Introduction

Population-based cancer registries are the cornerstone of oncology surveillance and public health policy, yet their operational efficacy is fundamentally constrained by the manual abstraction of data from unstructured pathology reports(1). This process is notoriously laborious, error-prone, and unscalable, creating data availability delays and thereby limiting the timeliness of research and public health interventions. For decades, the bottleneck is in needs to be overcome by automated solutions. Early Natural Language Processing (NLP) methods, including rule-based systems and traditional machine learning models, offered partial solutions but struggled with the linguistic heterogeneity of clinical text and lack of generalizability required for diverse cancer types and reporting styles(2,3).

The advent of Large Language Models (LLMs) has catalyzed a paradigm shift, demonstrating a remarkable capacity for clinical information extraction. Since 2023, a surge of research has confirmed that LLMs can achieve performance comparable to that of human experts in extracting structured data from pathology reports(4). This leap in

capability is driven by the models' deep contextual understanding and their proficiency in zero-shot learning, which significantly reduces the need for task-specific annotated data(5). The period between 2024 and 2025 has seen a proliferation of studies validating the high performance of LLMs for pathology data extraction across various institutions and cancer types. For instance, a prior study from Institute of Cancer Research and the National Cancer Institute, UK demonstrated its capacity on breast cancer report extraction, OpenAI's GPT-4o and Meta's Llama 3.1 405B achieved accuracies of 96.1% and 94.7%, respectively—performance statistically comparable to a human annotator (95.4%)(4). Similarly, researchers from University of California San Diego, working with the nationwide Veterans Health Administration (VHA) database, used Google's Gemma-2 model to identify colorectal dysplasia and cancer at F1-scores ranging from 95% to over 99%(6). At Memorial Sloan Kettering Cancer Center, a model study using GPT-4 and GPT-3.5 to extract key elements from hepatocellular carcinoma reports yielded average accuracies between 84.1% and 94.8%(7). However, while these studies affirm the capability of LLMs, they also highlight a crucial limitation: the majority of current research remains focused on a single cancer type or a narrow set of features. In the context of multiple cancers, the scope is often limited. For example, a notable study from British Columbia Cancer Registry and University of British Columbia developed an ensemble model that achieved an average precision and recall of 0.94 across 19 different organs; but this was only for the specific classification task of assigning a "tumor group" rather than comprehensive, multi-field data extraction(8). To date, a truly comprehensive, multi-cancer, multi-feature extractor suitable for end-to-end registry automation has not been demonstrated.

This gap persists largely because transition of LLMs from research benchmarks to routine clinical practice is impeded by critical implementation barriers. A primary

challenge is the risk of factual inaccuracy, or "hallucination," where models generate plausible but incorrect information—a failure mode that poses a significant threat to patient safety and data integrity(9)(10). Equally significant are the immense computational costs and data privacy concerns associated with state-of-the-art models. The highest-performance LLMs are often proprietary and accessible only via cloud-based APIs, requiring transmission of sensitive patient data to third-party vendors and thus raising regulatory and privacy issues. Alternatively, deploying powerful open-source models locally has historically been bothered by expensive, multi-GPU hardware clusters, a configuration that is infeasible for most healthcare institutions(11)(12). These creates a critical gap between the theoretical capabilities of LLMs and their practical deployment in resource-constrained clinical environments.

This research is situated at the confluence of two groundbreaking developments that directly address these implementation challenges: the release of powerful, resource-efficient open-weight models and the maturation of programmatic LLM development frameworks. The recent introduction of OpenAI's gpt-oss-20b, a 21-billion parameter model released under a permissive Apache 2.0 license, marks a crucial moment in democratizing access to state-of-the-art AI. Optimized for deployment on a single GPU within 16GB of memory, it provides a privacy-preserving alternative to proprietary, API-gated models without sacrificing performance on complex structuring tasks(13). Concurrently, the development of frameworks like Stanford's DSPy has revolutionized how such models are integrated into reliable systems. DSPy shifts the paradigm from "prompt engineering" to robust "programming," allowing developers to define complex AI workflows with declarative Python modules. Its compiler then automatically optimizes the underlying prompts and model interactions to maximize performance against user-defined data validation models(14). The synergy of a locally

deployable, high-performance model like gpt-oss and a programmatic optimization framework like DSPy provides, for the first time, the necessary toolkit to build the comprehensive, multi-cancer, multi-feature extractor that this study proposes.

This study directly addresses this implementation gap by presenting a novel, fully autonomous information extraction pipeline that achieves state-of-the-art performance across ten solid-organ cancers and can operate entirely on a single, commercially available GPU. Our framework makes three key contributions. First, we demonstrate that a locally deployed, open-source LLM can achieve registry-grade accuracy across a comprehensive set of features for multiple cancer types, ensuring data privacy and security. Second, we introduce a robust, multi-stage workflow built using the DSPy framework, a programmatic approach that replaces brittle, manual prompt engineering with a more principled, self-optimizing, and declarative AI development process. Finally, by validating our system on a large cohort of uncurated, real-world pathology reports, we provide a deployable and economically viable blueprint that is truly comprehensive, breaking the resource barrier to enable the widespread automation of cancer surveillance.

2 Methods

2.1 Study cohort

A total of 893 de-identified surgical pathology reports (2023–2024) from China Medical University Hospital were analyzed under institutional review board approval (IRB #CMUH114-REC2-037). Reports describing resected, invasive primary malignancies (e.g., lobectomy, mastectomy, esophagectomy) with microscopic details on tumor, margins, and/or lymph-node status were included as

tumor excision cases (n = 694). Reports describing biopsies (n = 25), carcinoma in situ (n = 9), metastatic resections (n = 28), or benign/inflammatory conditions (n = 137) were used as non-tumor-excision controls (n = 199).

The tumor excision cohort (n = 694) comprised definitive surgical pathology reports describing invasive primary carcinomas across ten organ systems covered by the schema library. Each report described a resection specimen containing verified single invasive malignancy, except for a small group categorized as Others, which included reports describing more than one distinct primary tumor (e.g., synchronous double cancers) or malignancies outside the schema scope (e.g., skin, kidney, hematolymphoid, or sarcoma). Organ-site labels were determined from final diagnosis field and confirmed by manual review. Table 1 summarizes the composition of the tumor excision cohort.

All identifiers were removed prior to analysis.

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

Table 1. Composition of the cancer excision cohort (n = 694)

2.2 Data preparation and expert annotation

Each report was segmented by double newlines into a list of paragraphs. A rough JSON summary was generated from these paragraphs by single inference of gpt-oss:20b by DSPy.Predict module, preserving the original phrasing and mapping key entities such as specimen type, laterality, and gross dimensions. Two board-certified pathologists independently reviewed and corrected the generated outputs using a side-by-side web interface showing text and JSON concurrently.

Discrepancies were resolved by blinded consensus, producing an expert-verified gold standard. All final JSONs were version-controlled and stored on secure institutional servers.

2.3 Schema definition

All outputs conform to a typed, Pydantic-style JSON schema composed of two hierarchical layers: an universal core and ten organ-specific extensions. The core layer provides the general structural backbone that applies to all surgical pathology reports, while the organ-specific layer captures details unique to individual tumour sites. Structure of the hierarchical schema layers are described in Table 2.

Each schema is implemented as a combination of four to six DSPy Signatures using Python type hints equivalent to Pydantic models, ensuring prompt and validator synchronization. Categorical fields are implemented as enumerated string

literals (Literal[...]), booleans as bool, numeric values as exact int, and missing data as None. This explicit typing eliminates synonym drift, rounding ambiguity, and fuzzy similarity, allowing deterministic, exact-match evaluation essential for registry and ICD coding applications.

Layer	Purpose	Example fields
Core layer	Defines attributes shared by all reports, including whether the document represents a cancer excision and the primary tumour site.	cancer_excision_report, cancer_category
Organ-specific layer	Extends the core with histologic, staging, margin, lymph-node, and biomarker fields tailored to a particular organ system.	procedure, tumor_size, margins, regional_lymph_node, biomarkers

Table 2. Structure of the hierarchical schema layers.

2.3.1 Alignment with CAP Cancer Reporting Protocols

Each organ-specific schema was constructed to comply with the College of American Pathologists (CAP) Cancer Protocols, which define standardized elements required for pathology reporting and registry submission(15). For every organ site, the corresponding schema's field definitions, enumerated values, and

logical dependencies were mapped to the CAP checklist and AJCC staging framework valid at the time of implementation.

Table 3 summarizes the CAP protocol references corresponding to each schema.

Organ site	CAP Cancer Reporting Protocol (version / release year)
Esophagus	Esophagus, Version 4.2.0.1 (June 2022)
Stomach	Stomach, Version 4.4.0.0 (March 2023)
Colorectum	Colon and Rectum, Resection, Version 4.3.1.0 (June 2024)
Breast	Breast Invasive, Resection, Version 4.10.0.0 (June 2024)
Uterine cervix	Uterine Cervix, Resection, Version 5.1.1.0 (April 2023)
Pancreas	Pancreas (Exocrine), Version 4.2.0.2 (November 2021)
Prostate	Prostate, Resection, Version 4.3.0.0 (September 2023)
Lung	Lung, Resection, Version 4.3.0.1 (September 2022)
Liver	Hepatocellular Carcinoma, Version 4.3.0.0 (June 2022)
Thyroid	Thyroid, Version 4.4.0.0 (March 2023)
Others (dual / out-of-scope)	Not applicable — excluded from schema alignment

Table 3. CAP Cancer Reporting Protocols corresponding to each organ-specific schema

2.3.2 Typed field semantics (illustrative examples)

The schema contains numerous typed fields spanning tumour descriptors, staging parameters, margins, lymph-node groups, and biomarkers.

The following examples illustrate the typing and semantic constraints applied consistently across all field classes:

- margin_category: Literal[...] — The named margin under evaluation, drawn from an organ-specific fixed enumeration (e.g., proximal, distal, deep, radial, medial, lateral).
- margin_involved: bool — Indicates whether invasive tumour is present at the inked margin (True = involved; False = clear).

- distance: Optional[int] — Integer distance (mm) from tumour to the named margin (0 if involved; None if unreported).
- station_name: Literal[...] — Named anatomical lymph-node group or "station" (e.g., level II, station 7, perigastric).

(Additional fields follow the same typed pattern, covering histologic type, grade, tumour size, margin descriptors, nodal metrics, staging components, and biomarker profiles.)

2.3.3 Variable-length lists

Fields such as margins, regional_lymph_node, and biomarkers are defined as List[ElementModel] and may be empty ([]) when unreported. Each element inherits the same strict typing rules as its parent schema. Other organ-specific fields—such as histological_type, tumor_size, or stage_group—are expressed as direct (non-

nested) key–value pairs within cancer_data. The resulting hierarchical organization of the schema is illustrated in Figure 1. Complete organ-specific field definitions, enumerations, and validation constraints are provided in the Supplementary Tables S1–S10, corresponding to the ten CAP-aligned schema modules (one per organ site). Pre-defined lymph node stations and surgical margin sites are also explained in the Supplementary Tables S1-S10.

Breast

```
{
  "cancer_excision_report": true,
  "cancer_category": "breast",
  "cancer_category_others_description": null,
  "cancer_data": {
    "procedure": "modified_radical_mastectomy",
    "histology": "other_special_types",
    ......
    "margins": [
      {
        "margin_category": "base",
        "margin_involved": false,
        "distance": 2,
        "description": "Free of tumor"
      },
      ......
    "regional_lymph_node":
    ......
```

Figure 1. Illustrative JSON structure of a breast cancer surgery report.

2.4 Extraction pipeline

The deterministic extraction workflow was implemented in DSPy 2.1 and 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.

2.4.1 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”.

2.4.2 Organ detection

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

2.4.3 Organ-specific extraction

Ten sets of organ-specific DSPy modules (for example, for breastcancer, for modules, BreastCancerNonnested, BreastCancerMargins, BreastCancerLN, BreastCancerBiomarkers, was used) receive both original pathology report and the json summary, previously generated by gpt-oss:20b, as input and output a validated JSON matching their schema.

Modules will attempt to extract tumor characteristics, AJCC/TNM staging, surgical margins and lymph-node details for each cancer type. For breast and colorectal cancer, biomarker statuses are extracted as well. An overview of the

deterministic extraction pipeline is shown in Figure 2 for a breast cancer report example, highlighting each computational stage and its data outputs.

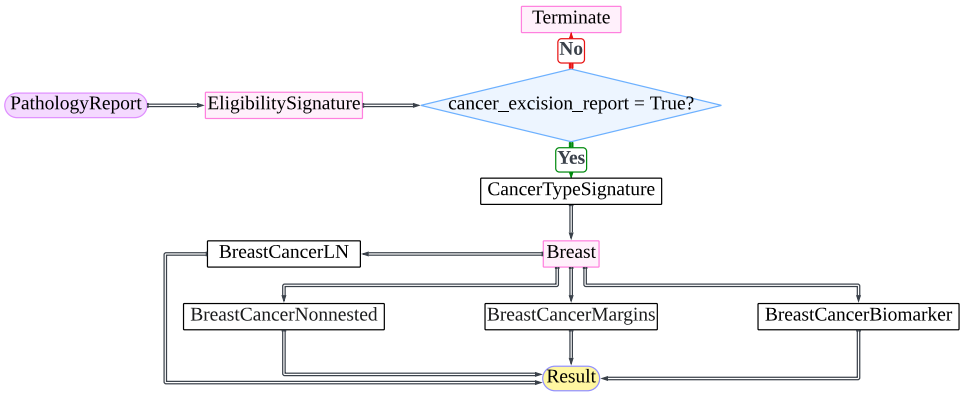


Figure 2. Extraction pipeline for structured cancer data from surgical pathology reports (breast as an example).

2.5 Evaluation metrics

Model performance was evaluated through exact-match field accuracy, comparing each generated JSON output against the expert-verified reference. Because all variables are strictly typed (Literal, bool, int, or None), evaluation was performed using string-equivalent, positionally aligned comparison rather than fuzzy or similarity-based metrics. Fields marked "others" were excluded from analysis.

The evaluation comprised four organ-level field groups and a fifth aggregate category encompassing all remaining variables. The metrics are explained in table 4.

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.

Table 4. Evaluation metrics and definitions

Each metric represents a binary exact-match comparison between model output and expert-annotated reference. No partial credit or token-level similarity was used. Overall accuracy for each category equals the ratio of correctly matched fields to the total number of evaluated fields.

2.6 Computational environment and reproducibility

All experiments were conducted on a Windows 11 Pro workstation equipped with an Intel Xeon W3 CPU, 64 GB RAM, and a NVIDIA RTX A6000 Ada Generation GPU (48 GB VRAM). The pipeline was implemented in Python 3.11

using DSPy 2.1 for declarative program synthesis and the open-source gpt-oss:20b model (Apache 2.0 license) for rough JSON generation and LLM inference.

All downstream schema logic and validation modules were proprietary, implemented in-house using typed Pydantic models. Model inference was executed at *temperature = 0.7* and *top-p = 0.7* to balance completeness and determinism, as lower sampling temperatures occasionally produced empty outputs. The complete extraction of a cancer excision report—including preprocessing, classification, organ-specific extraction, and validation—required approximately 60 seconds per case on average.

Although proprietary modules are not open-source, full reproducibility can be achieved by re-implementing equivalent schema definitions using typed field specifications and CAP protocol mappings provided in the Supplementary Tables S1–S10, under identical inference settings.

2.7 Statistical analysis

All performance metrics were computed using the scikit-learn library (version 1.5.0) under Python 3.11.

Each extracted field was compared with its expert-verified reference using exact string matching, consistent with the schema's typed definitions (Literal, bool, int, None).For list-type variables such as margins and lymph-node stations, corresponding elements were aligned by field name before scoring.

Per-field precision, recall, and F1 scores were calculated using sklearn.metrics.precision_score, recall_score, and f1_score, employing the "micro"

averaging method to reflect field-level correctness. Macro F1 was obtained as the unweighted mean of class-level F1 scores, giving equal weight to each field type. Overall accuracy was defined as the proportion of exactly matched fields across all evaluated reports.

The CancerTypeSignature multiclass classifier, which assigns reports to one of eleven organ categories, was evaluated using per-class precision, recall, F1, and accuracy, together with overall accuracy and macro F1 across all classes. The EligibilitySignature binary classifier, used to distinguish cancer-excision from non–cancer-excision reports, was evaluated across multiple independent test runs using the same scikit-learn metrics framework.

2.8 Use of generative AI during manuscript production

This manuscript was prepared with partial drafting assistance from large language models. OpenAI GPT-5 was used for non-confidential sections to improve structure and clarity; gpt-oss:20b was used to draft and revise paragraphs describing proprietary schema design and extraction logic under a controlled, offline environment; and Gemini Pro 2.5 with Deep Research was used to identify and summarize peer-reviewed literature cited in the Introduction. All analytical workflows, data interpretation, and final text approval were performed by the authors. No LLM generated, altered, or fabricated experimental data.

2.9 Ethical oversight

This study was approved by the research ethics committee of China Medical University & Hospital, under REC number of CMUH114-REC2-037. All reports

were fully de-identified and processed on encrypted institutional servers. No re-identification was attempted.

3 Results

3.1 Eligibility and organ classification

The eligibility classifier reliably distinguished cancer surgery reports from non-cancer-surgery documents across three independent test runs (N = 694). Accuracy consistently exceeded 96 % (0.966–0.969), with average accuracy of 96.6% and a macro average F1 score of 97.9%, confirming near-perfect sensitivity and minimal false-negative risk. Only a single breast-excision report was misclassified in Run 1, while remaining false positives (28–30 per run) were confined to the "others" category. These results indicate that the model can safely operate as the first triage stage in an automated pipeline, ensuring that no eligible cancer surgery report is missed.

The subsequent organ-labeling sub-model demonstrated similarly strong performance, achieving precision 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 mostly attributable to cases with double primary malignancies—for example, reports describing independent tumors in both lung lobes, each staged as a separate primary cancer. Such cases, which were intentionally excluded from automated staging logic and assigned to the heterogeneous "others" category, were occasionally interpreted by the model as

belonging to the "lung" class. Collectively, these results confirm that foundational classification layers of the DSPy pipeline operate with registry-grade reliability, providing clean input for downstream extraction of tumor characteristics, margins, and lymph-node (LN) data. Detailed confusion matrices for each of the three test runs are provided in the Supplementary Tables S11–S13.

Building on the model's demonstrated stability in both eligibility determination and organ classification—even under stochastic sampling (temperature = 0.7)—subsequent evaluations focused on the extraction components. Accordingly, accuracy assessments for per-organ extraction, lymph-node (LN) extraction, margin extraction, and biomarker extraction were conducted using results from Run 3 exclusively.

3.2 Per-organ extraction accuracy for various registry item

After classification, 893 cancer surgery reports were parsed through organ-specific extraction modules, each aligned to a JSON schema of variable fields according to different organ system (procedure, histology, grade, AJCC T/N/M categories, and vice versa, please see **Supplementary Tables S14–S23** for detail). Exact-match accuracies across all ten organ systems (totally 193 fields) ranged between 88.9 % and 96.2 %, with an overall mean of 94.3%±8.32%. Table 5 summarizes the cohort size, field count, and mean performance for each organ. Surgical margins, lymph nodes and biomarkers are extracted by separate modules and discussed in later section.

Table 5. Per-organ extraction accuracy for various registry item

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**

High-performing fields (≥ 95 %) were consistent across organs expect thyroid (where frequent referrance to other reports caused some model difficulty) and included procedure histology, grade, and 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 cases labeled “Nx” but inferred by the model as “N0”—also contributed to minor discrepancies. 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 or tumor percentage and prostate weight in prostate reports—where textual variability is high.

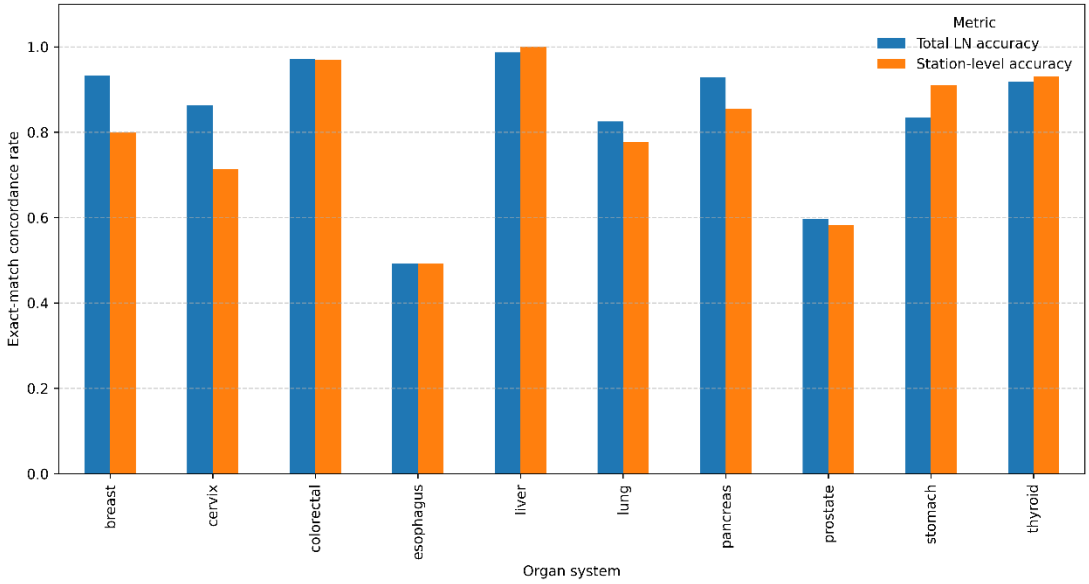
In summary, the organ-specific modules sustain ≥ 90 % mean exact-match accuracy across varying schema sizes, reaffirming the scalability of the one-shot LLM pipeline for comprehensive extraction of structured oncologic data. Full per-field extraction results for each organ, including exact-match accuracy for all schema items, are provided in the Supplementary Tables S14–S23.

3.3 Lymph node status extraction Performance

The lymph-node (LN) module captures both global and station-level nodal information.

For each report, the extractor returns (i) the total number of nodes examined and involved, and (ii) a structured list of anatomical stations, each with its own lymph node category label (lymph_node_category) and numeric counts (examined, involved). Exact-match concordance with the expert-annotated reference was computed separately for these two granularities—total-node accuracy and station-level accuracy—as summarized in Figure 3.

Figure 3. Lymph node status extraction accuracy by organ.



Across the ten target organs, the mean total-node concordance exceeded 80 %, with high consistency in breast, cervix, liver, and thyroid, each achieving nearly complete agreement with the reference dataset. Performance was comparatively lower in organs with numerous and anatomically complex nodal compartments—particularly lung, colorectal, and pancreas—where model output occasionally contained empty or missing entries rather than incorrect classifications.

Station-level accuracy followed a similar trend, with breast, cervix, liver, and thyroid again demonstrating near-perfect concordance and colorectal, lung, and pancreas showing moderate values corresponding to station density. Detailed per-organ LN metrics, including both total and station-level analyses, are provided in Supplementary Table S24.

3.4 Surgical margin extraction performance

The surgical margin module extracted a binary positivity flag (margin_involved), a categorical descriptor (margin_category), and a numeric distance (mm) for each margin item in margin lists for each report.Two evaluations were performed for each margin category: (i) margin positivity, requiring concordant margin_category + margin_involved, and (ii) margin distance, requiring exact match of margin_category, margin_involved, and distance.

Across organs, margin positivity showed a mean concordance of 91.2 %, highest in breast, colorectal, stomach, lung, and esophagus (≥ 95 %), and lowest in pancreas (67 %), cervix (84 %), and thyroid (89 %). Discrepancies primarily reflected category-mapping failures in organs with multiple, variably named margins rather than incorrect tumor-involvement classification. Margin distance concordance ranged 66–89 %, limited by inequality notation and unit variation. Detailed data are summarized in Figure 4.

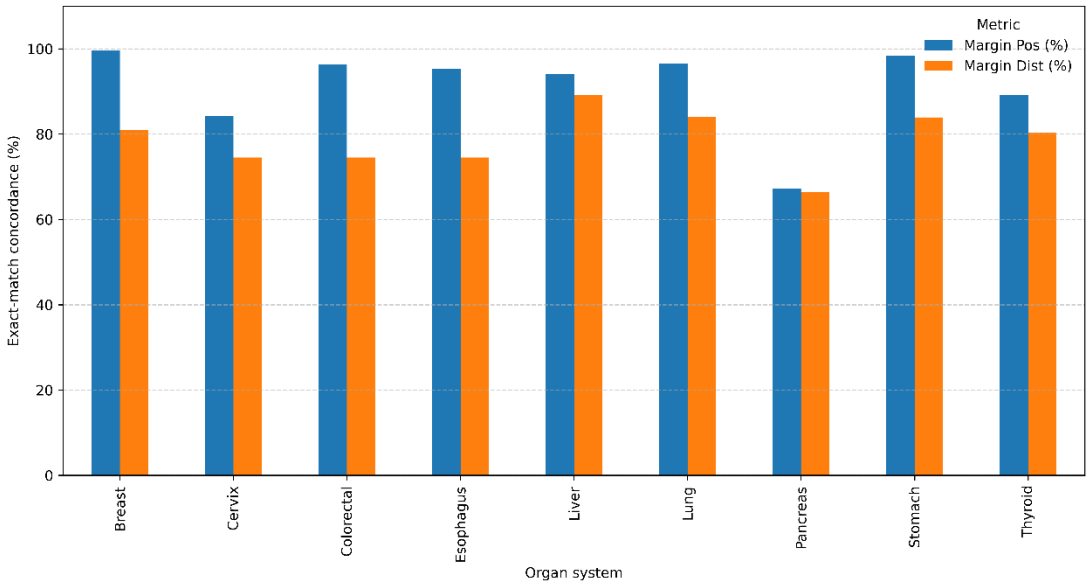


Figure 4. Surgical margin status extraction accuracy by organ.

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.

Biomarker	Field type	Correct (n)	Accuracy (%)
Estrogen receptor (ER)	Positivity	75 / 75	100.0
Estrogen receptor (ER)	Percentage	73 / 75	97.3
Progesterone receptor (PR)	Positivity	74 / 75	98.7
Progesterone receptor (PR)	Percentage	73 / 75	97.3
HER2	Score	75 / 75	100.0
Ki-67	Percentage	74 / 75	98.7

Table 6. Accuracy of breast cancer biomarker extraction (N = 75)

3.6 Overall Performance Summary

The DSPy pipeline achieved registry-grade reliability: >96 % accuracy for classification, >94 % for per-organ extraction, >80 % for lymph-node concordance, and >90 % for margin positivity. Breast biomarker extraction exceeded 97 %. Overall, the system demonstrated robust, scalable, and clinically interpretable performance across diverse organ systems.

4 Discussion

4.1 Description and discussion of the context

While large language models (LLMs) now demonstrate human-level accuracy for information extraction from clinical texts(4), their clinical deployment is hindered by a dilemma: proprietary models raise data privacy and cost concerns, while open-source alternatives have historically required extensive computational resources(16). This landscape is changing, as top-tier open-source models are now matching the performance of proprietary systems on complex clinical reasoning tasks(17). Concurrently, new solutions are emerging to address resource barriers, such as retrieval-augmented generation (RAG) for efficiently processing long documents(18), and the rise of powerful Small Language Models (SLMs) suitable for local deployment(19). Together, these advancements offer a path to develop systems that can overcome the significant technological and organizational barriers to AI adoption in healthcare(20).

4.2 Limitations of this study

The principal limitation of this study is its exclusive reliance on internal institutional data for both model development and validation. The pipeline has not yet been evaluated on external or public datasets, leaving its cross-institutional generalizability unverified. Report conventions, formatting heterogeneity, and linguistic variability across centers may influence extraction accuracy.

Future work should therefore include independent external benchmarking to confirm robustness across institutions and reporting templates.

4.3 Further research direction and intrinsic challenges

Further refinement of prompt structures and model parameters may reduce current failure modes. Blank outputs occasionally observed in lengthy lymph-node or margin lists, as well as inconsistencies triggered by ambiguous shorthand (e.g., N0, Nx), could likely be mitigated through context-aware prompting or adaptive post-processing heuristics. Incremental tuning along these lines would enhance completeness without compromising precision.

Certain pathology report types remain inherently complex and difficult to standardize. Examples include colorectal resections involving multiple adjacent organs or specimens containing synchronous esophageal and gastric primaries. In such cases, the definition of a “standard margin” or “regional node” schema becomes ambiguous, limiting performance even under ideal model configurations. Addressing these cases may require dynamic schema generation or hybrid reasoning frameworks that accommodate overlapping anatomical and diagnostic contexts.

5 Conclusion

This study demonstrates that large-language-model–based extraction can achieve registry-grade fidelity across diverse organ systems when trained on structured institutional data. However, the findings are restricted to internal validation, and broader applicability requires external testing and schema generalization. Future work integrating multi-institutional datasets and open

benchmarking will be essential to establish scalability and clinical deployment readiness.

6 Ethical Statement

This study was approved by the research ethics committee of China Medical University & Hospital, under REC number of CMUH114-REC2-037. All reports were fully de-identified and processed on encrypted institutional servers. No re-identification was attempted.

7 Data Sharing

The de-identified patient data that support the findings of this study will not be made publicly available due to the sensitive nature of the clinical information contained within the free-text pathology reports and to protect patient privacy in accordance with the institutional review board (IRB) protocol under which the data were collected.

The source code for the AI extraction pipeline and the complete data dictionary are not deposited in a public repository. The authors will consider making these materials available only upon reasonable request under specific and compelling circumstances, such as for non-commercial scientific investigation that aims to directly replicate, validate, or challenge the study's core methodological findings. All such requests must be directed to the corresponding author and include a detailed proposal outlining the precise purpose for which the materials are required. Access, if granted, will be at the sole discretion of the

authors and will require the execution of a formal material transfer and use agreement.

8 Declaration of interests

We declare no competing interests.

9 Funding

This work was supported by the National Science and Technology Council (NSTC) of Taiwan, under grant number NSTC 113-2221-E-039-017, and the NSTC College Student Research Grant, number 114-2813-C-039-132-E.

10 Acknowledgement

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.

11 Reference

1. Dahl S, Bøgsted M, Sagi T, Vesteghem C. Performance of Natural Language Processing for Information Extraction From Electronic Health Records Within Cancer: Systematic Review. JMIR Med Inform. 2025 Sept 12;13:e68707.

2. Park P, Choi Y, Han N, Park YL, Hwang J, Chae H, et al. Leveraging natural language processing for efficient information extraction from breast cancer pathology reports: Single-institution study. PLOS ONE. 2025 Feb 18;20(2):e0318726.

3. Abedian S, Sholle ET, Adekkanattu PM, Cusick MM, Weiner SE, Shoag JE, et al. Automated Extraction of Tumor Staging and Diagnosis Information From Surgical Pathology Reports. JCO Clinical Cancer Informatics. 2021 Dec;(5):1054–61.

4. Balasubramanian JB, Adams D, Roxanis I, de Gonzalez AB, Coulson P, Almeida JS, et al. Leveraging large language models for structured information extraction from pathology reports. Journal of Pathology Informatics. 2025 Oct 10;100521.

5. Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, Papadopoulos N, Odavic D, Risteski P, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform. 2025 Jan 19;32(1):e101139.

6. Johnson B, Bath T, Huang X, Lamm M, Earles A, Eddington H, et al. Large language models for extracting histopathologic diagnoses of colorectal cancer and dysplasia from electronic health records. BMJ Open Gastroenterol. 2025 Sept 18;12(1):e001896.

7. Geevarghese R, Sigel C, Cadley J, Chatterjee S, Jain P, Hollingsworth A, et al. Extraction and classification of structured data from unstructured hepatobiliary pathology reports using large language models: a feasibility study compared with rules-based natural language processing. J Clin Pathol. 2025 Jan 17;78(2):135–8.

8. Gondara L, Simkin J, Devji S, Arbour G, Ng R. ELM: Ensemble of Language Models for Predicting Tumor Group from Pathology Reports [Internet]. arXiv; 2025 [cited 2025 Oct 19]. Available from: http://arxiv.org/abs/2503.21800

9. Cheng J. Applications of Large Language Models in Pathology. Bioengineering. 2024 Apr;11(4):342.

10. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). 2025 Jan 21;5:26.

11. Fime AA, Hossain MZ, Zaman S, Shahid ARB, Imteaj A. Benchmarking Large-Language Models for Resource-Efficient Medical AI for Edge Deployment. In: Proceedings of the AAAI Symposium Series [Internet]. 2025 [cited 2025 Oct 19]. p. 137–42. Available from: https://ojs.aaai.org/index.php/AAAI-SS/article/view/35580
12. Lorencin I, Tankovic N, Etinger D. Optimizing Healthcare Efficiency with Local Large Language Models. In: Intelligent Human Systems Integration (IHSI 2025): Integrating People and Intelligent Systems [Internet]. AHFE Open Acces; 2025 [cited 2025 Oct 19]. Available from: https://openaccess.cms-conferences.org/publications/book/978-1-964867-36-6/article/978-1-964867-36-6_58
13. Kumar D, Yadav D, Patel Y. GPT-OSS-20B: A Comprehensive Deployment-Centric Analysis of OpenAI's Open-Weight Mixture of Experts Model [Internet]. arXiv; 2025 [cited 2025 Oct 19]. Available from: http://arxiv.org/abs/2508.16700
14. Khattab O, Singhvi A, Maheshwari P, Zhang Z, Santhanam K, Haq S, et al. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In: The Twelfth International Conference on Learning Representations [Internet]. 2024 [cited 2025 Oct 19]. Available from: https://openreview.net/forum?id=sY5N0zY5Od
15. Torous VF, Simpson RW, Balani JP, Baras AS, Berman MA, Birdsong GG, et al. College of American Pathologists Cancer Protocols: From Optimizing Cancer Patient Care to Facilitating Interoperable Reporting and Downstream Data Use. JCO Clinical Cancer Informatics. 2021 Dec;(5):47–55.
16. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med. 2025 Mar 6;8:143.
17. Comparison of Frontier Open-Source and Proprietary Large Language Models for Complex Diagnoses | Health Policy | JAMA Health Forum | JAMA Network

[Internet]. [cited 2025 Oct 21]. Available from: https://jamanetwork.com/journals/jama-health-forum/fullarticle/2831206

18. Cheetirala SN, Raut G, Patel D, Sanatana F, Freeman R, Levin MA, et al. Less Context, Same Performance: A RAG Framework for Resource-Efficient LLM-Based Clinical NLP [Internet]. arXiv; 2025 [cited 2025 Oct 21]. Available from: http://arxiv.org/abs/2505.20320

19. Garg M, Raza S, Rayana S, Liu X, Sohn S. The Rise of Small Language Models in Healthcare: A Comprehensive Survey [Internet]. arXiv; 2025 [cited 2025 Oct 21]. Available from: http://arxiv.org/abs/2504.17119

20. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare. Cureus. 15(10):e46454.