EMPLOYING CONSENSUS-BASED REASONING WITH LOCALLY DEPLOYED LLMS FOR ENABLING STRUCTURED DATA EXTRACTION FROM SURGICAL PATHOLOGY REPORTS

Aaksh Tripathi*
Department of Machine Learning
H. Lee Moffitt Cancer Center & Research Institute

Asim Waqas, PhD*
Department of Cancer Epidemiology
H. Lee Moffitt Cancer Center & Research Institute

Kavya Venkatesan
Department of Machine Learning
H. Lee Moffitt Cancer Center & Research Institute

Ehsan Ullah, PhD, M.Phil., MBBS
Department of Surgery
Health New Zealand, Counties Manukau
Auckland, New Zealand

Asma Khan, MBBS
Armed Forces Institute of Pathology
Rawalpindi, Pakistan

Farah Khalil, MD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Wei-Shen Chen, MD, PhD
Department of Dermatology & Cutaneous Surgery
University of South Florida

Zarifa Gahramanli Ozturk, MD
Clinical Science Lab
H. Lee Moffitt Cancer Center & Research Institute

Daryoush Saeed-Vafa, MD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Marilyn M. Bui, MD, PhD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Matthew B. Schabath, PhD
Department of Cancer Epidemiology
H. Lee Moffitt Cancer Center & Research Institute

Ghulam Rasool, PhD
Department of Machine Learning
H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL
Ghulam.Rasool@moffitt.org

ABSTRACT

Surgical pathology reports contain essential diagnostic information, in free-text form, required for cancer staging, treatment planning, and cancer registry documentation. However, their unstructured nature and variability across tumor types and institutions pose challenges for automated data extraction. We present a consensus-driven, reasoning-based framework that uses multiple locally deployed large language models (LLMs) to extract six key diagnostic variables: site, laterality, histology, stage, grade, and behavior. Each LLM produces structured outputs with accompanying justifications, which are evaluated for accuracy and coherence by a separate reasoning model. Final consensus values are determined through aggregation, and expert validation is conducted by board-certified

*Equal first authors.

or equivalent pathologists. The framework was applied to over 4,000 pathology reports from The Cancer Genome Atlas (TCGA) and Moffitt Cancer Center. Expert review confirmed high agreement in the TCGA dataset for behavior (100.0%), histology (98.5%), site (95.2%), and grade (95.6%), with lower performance for stage (87.6%) and laterality (84.8%). In the pathology reports from Moffitt (brain, breast, and lung), accuracy remained high across variables, with histology (95.6%), behavior (98.3%), and stage (92.4%), achieving strong agreement. However, certain challenges emerged, such as inconsistent mention of sentinel lymph node details or anatomical ambiguity in biopsy site interpretations. Statistical analyses revealed significant main effects of model type, variable, and organ system, as well as model × variable × organ interactions, emphasizing the role of clinical context in model performance. These results highlight the importance of stratified, multi-organ evaluation frameworks in LLM benchmarking for clinical applications. Textual justifications enhanced interpretability and enabled human reviewers to audit model outputs. Overall, this consensus-based approach demonstrates that locally deployed LLMs can provide a transparent, accurate, and auditable solution for integrating AI-driven data extraction into real-world pathology workflows, including cancer registry abstraction and synoptic reporting.

***K*eywords** Large Language Models (LLMs) · Surgical Pathology Reports · Natural Language Processing (NLP) · Reasoning · Extraction · Cancer Registry

1 Introduction

Surgical pathology reports are the cornerstone of cancer diagnosis, staging, and histologic classification. Additionally, they provide valuable information to guide treatment selection, inform prognostic assessments, and determine eligibility for clinical trials [1, 2, 3]. The information contained in these reports also underpin national cancer surveillance efforts, such as the Surveillance, Epidemiology, and End Results (SEER) program and the National Program for Cancer Registries (NPCR), which require consistent capture of variables like tumor site, laterality, histological diagnosis, stage, grade, and behavior [4, 5, 6, 7, 8]. Despite their centrality, most pathology reports are composed as unstructured free-text narratives, often dictated in formats that vary widely within and across institutions, subspecialties, and pathologists [9]. This heterogeneity poses significant challenges for automated information extraction and downstream reuse. While rule-based natural language processing (NLP) systems have been applied to map free-text to discrete fields, they are time-consuming to develop, sensitive to linguistic variation, and poorly generalize across diverse cancer types and documentation styles [9, 10, 11, 12]. As the holistic cancer care and research increasingly depend on large-scale, structured datasets to power precision medicine and population health efforts, there is a critical and urgent need for extraction methods that are not only accurate and scalable but also adaptable to real-world documentation variability.

Recent advances in large language models (LLMs), including ChatGPT, LLaMA, and DeepSeek, have enabled automated extraction of structured information from free-text clinical documents, including radiology, discharge summaries, and pathology reports [13, 14, 15, 16, 17, 18, 8, 19, 20]. These models are capable of identifying key clinical entities and relationships with little or no task-specific training, making them candidates for applications that require scalability across cancer types and report formats. However, despite their linguistic fluency, LLMs suffer from key limitations that restrict their reliability in clinical workflows. A major concern is their tendency to produce plausible-sounding but factually incorrect or unsupported outputs, a phenomenon often referred to as hallucination. [21, 22, 23]. Hallucinations may result in fabricated stage values, incorrect site mappings, or invalid combinations of diagnostic attributes, all of which pose risks in settings like cancer registry abstraction or clinical trial matching. In addition, most LLMs do not express uncertainty and may provide definitive-sounding answers even when source evidence is weak or contradictory. These behaviors, combined with the opacity of current LLM decision-making processes, have raised concerns among clinicians and pathologists about their safety and trustworthiness [24, 25, 26, 27]. As a result, there is a growing consensus that LLM-based extraction systems must incorporate structured reasoning, interpretability, and safeguards that mitigate error propagation in high-stakes clinical contexts.

Despite their promise, the deployment of LLMs in clinical environments is constrained by stringent requirements around patient privacy, regulatory compliance, and institutional data governance. Cloud-based LLMs, such as those offered by commercial providers, often require transmitting protected health information to external servers, raising concerns about data breaches and non-compliance with regulations like HIPAA and GDPR [28]. These concerns are heightened by recent controversies around unauthorized model training on clinical data, leading to increased institutional scrutiny and legal risk. As a result, there is growing interest in deploying LLMs locally within secure hospital infrastructures, where data remains behind institutional firewalls and access can be tightly controlled [29, 30, 31]. Locally hosted models also offer greater transparency, auditability, and customization for clinical use cases, particularly when based on open-source architectures. However, practical challenges remain: most LLMs require substantial computational resources, and integration into clinical workflows demands interoperability with legacy systems, domain-specific prompt design,

and support for human oversight. Addressing these barriers is critical to translate LLMs from research prototypes to deployable clinical tools.

To bridge the gap between the capabilities of LLMs and the demands of clinical pathology, we developed a consensus-based framework that emphasizes accuracy, transparency, and clinical alignment. Rather than relying on a single model to interpret free-text reports, our approach leverages multiple locally deployed LLMs, each prompted to extract discrete diagnostic variables such as tumor site, histology, and behavior, alongside concise textual justifications for their selections. A separate reasoning model then evaluates these outputs, synthesizes inter-model agreement, and adjudicates conflicts based on justification strength and internal consistency. This multi-stage process mirrors how diagnostic consensus is achieved in clinical practice: through independent interpretation, reasoned comparison, and reconciliation of divergent opinions. By embedding structured reasoning into every stage of prediction and adjudication, the system produces results that are not only accurate but also explainable and auditable, which are the qualities essential for downstream integration into cancer registries, synoptic reporting systems, and clinical trial eligibility assessments.

This work distinguishes itself by combining locally deployed LLMs with a reasoning-based consensus architecture that prioritizes auditability, transparency, and clinical alignment. Rather than producing black-box outputs, the system generates structured predictions with accompanying justifications and resolves conflicts through an adjudication step based on coherence, agreement, and clinical plausibility. These features enable human review, support regulatory compliance, and foster trust in high-stakes environments like cancer registry abstraction and synoptic reporting. Designed to run entirely within institutional firewalls using quantized open-source models, the framework can be deployed on modest hardware, making it scalable and compatible with existing pathology workflows.

In this study, we apply a consensus-based reasoning framework to two large corpora of surgical pathology reports: a publicly available dataset from The Cancer Genome Atlas (TCGA) and a real-world institutional dataset from Moffitt Cancer Center, comprised of breast, brain, and lung cancers. The system operates entirely on locally deployed infrastructure, using multiple open-source LLMs to extract six clinically relevant variables along with model-generated justifications. A reasoning module adjudicates conflicting outputs, producing final consensus predictions that are auditable and suitable for human review. Model performance is evaluated using automated metrics and expert validation by board-certified or equivalent pathologists, and we conduct detailed statistical analyses to examine model accuracy across organ systems and diagnostic fields. Together, these findings demonstrate the clinical feasibility of deploying AI-driven extraction tools in routine pathology workflows while highlighting the importance of interpretability, adjudication, and dataset diversity in building trustworthy systems for structured data abstraction.

2 Methods

A locally deployable, three-stage framework was developed for extracting structured clinical variables from unstructured surgical pathology reports using LLMs. The system targets six key diagnostic variables, including site, laterality, histology, stage, grade, and behavior, and emphasizes interpretability by requiring each model to produce brief justifications alongside its predictions. The framework integrates outputs from multiple LLMs, evaluates their accuracy, and applies reasoning-based aggregation to produce a final consensus output [32]. To validate its reliability and clinical relevance, we applied the framework to two diverse pathology datasets and conducted both automated evaluation and expert pathologist review.

Prompting and Reasoning Framework

The prompting and reasoning framework consists of three sequential stages as presented in Fig. 1: multi-model extraction, automated evaluation, and consensus-based aggregation. In the first stage, each pathology report was independently processed by eleven locally deployed LLMs: LLaMA 3.1, LLaMA 3.2, LLaMA 3.3, Mistral, Mixtral, Qwen, Phi-4, Tulu3:8B, DeepSeek-R1:7B, DeepSeek-R1:32B, and DeepSeek-R1:70B. Each model received the full report text and was prompted to extract six structured clinical variables along with a brief justification for each value. If a variable was absent or could not be inferred, the model was instructed to return "Unknown" with an explanation. Outputs were returned in standardized JSON format to support downstream scoring and aggregation. For brain cancer reports, the "stage" variable was excluded due to the lack of clinical relevance.

In the second stage, each extracted variable and its corresponding justification were evaluated by a separate LLM acting as an automated adjudicator. DeepSeek-R1:70B was used as the evaluator model across all reports. For every extraction, the model was presented with the original pathology report text and the variable-specific justification provided by the LLM that produced the prediction. Based on internal consistency, textual evidence, and clinical plausibility, the evaluator assigned a binary score of 1 (correct) or 0 (incorrect) to each output. These evaluation scores were used to identify high-confidence predictions and to inform downstream aggregation in the final stage.

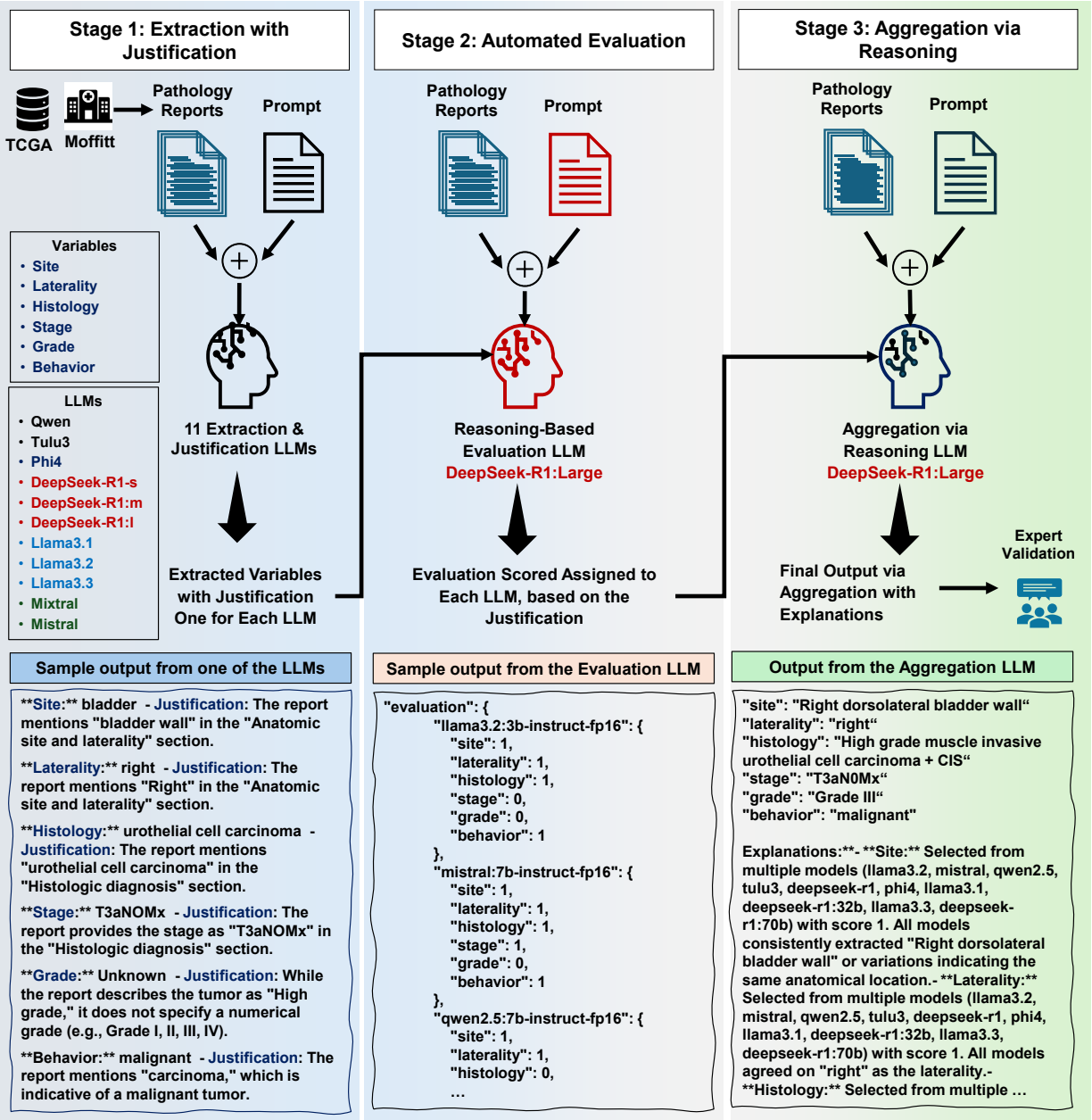

Figure 1: **Multi-stage reasoning framework for structured information extraction from surgical pathology reports.** This diagram illustrates the full three-stage prompting pipeline developed to extract six variables, site, laterality, histology, stage, grade, and behavior, from unstructured pathology reports. Reports were sourced from two datasets: TCGA and Moffitt Cancer Center, encompassing eight organ types across more than 4,000 reports. In **Stage 1**, eleven locally hosted LLMs (LLaMA 3.1, LLaMA 3.2, LLaMA 3.3, Mistral, Mixtral, Qwen, Phi-4, Tulu3:8B, DeepSeek-R1:7B, DeepSeek-R1:32B, and DeepSeek-R1:70B) were independently prompted to extract variable values and provide textual justifications. In **Stage 2**, DeepSeek-R1:70B was used as an automated evaluator to assign binary accuracy scores (1 = correct, 0 = incorrect) to each model's predictions by comparing them to the original report and corresponding justification. In **Stage 3**, a second prompt to DeepSeek-R1:70B served as an aggregation step, integrating all model predictions and their evaluation scores to select final consensus values. Outputs from this stage included not only the selected variable values but also concise rationales explaining the model's reasoning. A subset of outputs from both datasets was randomly sampled and independently reviewed by board-certified or equivalent pathologists, who scored each extracted variable and annotated common error patterns. This framework enables interpretable, accurate, and auditable structured data extraction suitable for integration into real-world clinical workflows such as cancer registry abstraction and synoptic reporting.

In the third stage, a reasoning-based aggregation process was used to determine final consensus values for each variable. DeepSeek-R1:70B was again used as the aggregation model, this time prompted to adjudicate among the outputs of all eleven LLMs using their corresponding evaluation scores from Stage 2. For each variable, the aggregation model synthesized the predictions, justifications, and evaluator scores to select a single consensus value. The model prioritized outputs with both inter-model agreement and high evaluation scores, while discounting hallucinated or unsupported responses. Final outputs were formatted as structured JSON objects that included not only the selected values but also concise rationales explaining the reasoning behind each choice. These justifications were designed to enable human-in-the-loop review and support traceability in clinical workflows.

Expert Review and Clinical Validation

To evaluate clinical validity and real-world usability, a subset of final outputs from the aggregation stage was independently reviewed by board-certified (or holding equivalent qualification) pathologists. Reviewers were presented with the original pathology report alongside the model-extracted values (from stage 3) and corresponding justifications. Each variable was scored on a three-point scale: 1 (correct), 0.5 (partially correct), or 0 (incorrect), based on consistency with the report content and accepted clinical interpretation. In addition to assigning scores, reviewers provided qualitative feedback to identify common error types, such as omission of laterality, stage misclassification due to missing contextual cues, and under-classification of tumor grade. These expert annotations helped validate model performance in nuanced clinical scenarios and informed recommendations for refining prompts, improving reasoning trace quality, and flagging low-confidence outputs for human review.

Datasets

The framework was evaluated using two datasets of surgical pathology reports: one derived from TCGA and the other from patient records at Moffitt Cancer Center. The TCGA dataset was sourced from the OCR-extracted corpus released by Kefeli and Tatonetti [33], from which we randomly selected 3,251 reports spanning seven organ sites: kidney (858), uterus (639), prostate (472), liver (421), bladder (413), cervix (309), and pancreas (139). A subset of 69 reports from five organs (bladder, prostate, liver, cervix, and kidney) was manually reviewed by three board-certified (or having equivalent qualification) pathologists (207 evaluations in total) to support expert evaluation.

The Moffitt dataset consisted of 510 de-identified surgical pathology reports from patients treated between 2009 and 2018. All reports were linked to institutional cancer registry records and restricted to three cancer types: breast (256), brain or central nervous system (171), and lung (156). For expert review, 30 reports were randomly selected from each organ group (90 total) and independently evaluated by three board-certified pathologists. All reports in both datasets were converted into structured JSON format containing a unique identifier, organ label, and full free-text report, allowing for consistent preprocessing, prompting, and scoring across the three-stage framework.

Statistical Analysis

Statistical analyses were conducted separately for the automated model evaluations and the expert-reviewed subsets. For each dataset, structured records containing model identity, variable type, organ site, and extraction score were compiled for analysis. One-way ANOVA was used to compare average extraction accuracy across the eleven LLMs, and a two-way ANOVA was used to assess interactions between model type and diagnostic variable. A three-way ANOVA was used to examine joint effects of model, variable, and organ site, including all pairwise and three-way interaction terms. When statistically significant main or interaction effects were observed, Tukey's Honestly Significant Difference (HSD) was used to identify pairwise differences between models or organs while controlling the family-wise error rate.

For the expert-reviewed subsets, the pathologist-assigned accuracy scores were analyzed using two-way ANOVA with fixed effects for variable and organ. This analysis was conducted separately for the TCGA and Moffitt datasets. As with the automated analysis, significant differences were followed by Tukey's HSD tests to assess specific pairwise contrasts. All statistical tests were performed using the statsmodels package in Python (v0.14.0), with significance defined as ($\alpha = 0.05$).

3 Results

The performance of eleven locally deployed LLMs was evaluated in extracting six clinically relevant variables, site, laterality, histology, stage, grade, and behavior, from unstructured surgical pathology reports. The evaluation was conducted across two distinct datasets: publicly available pathology reports from TCGA and patient reports from Moffitt Cancer Center. For each dataset, a two-tiered assessment strategy was applied: automated scoring by a reference

LLM based on the extracted values and justifications, and expert review by board-certified or equivalent pathologists on a randomly selected subset. Statistical comparisons were performed to assess variability in extraction accuracy across models, variables, and organ types. Results are presented in two parts: TCGA and Moffitt. Each part is organized by evaluation method: automated scoring followed by expert validation.

TCGA Dataset: Automated Evaluation

Figure 2 summarizes the automated evaluation of eleven LLMs across six diagnostic variables and seven organ types in the TCGA dataset. Each model-generated output was independently scored by a reasoning-based evaluator (DeepSeek-R1:70B) based on its alignment with the source report and accompanying justification. As shown in Fig. 2(A), accuracy varied by variable, with most models achieving high scores for behavior, histology, and site, and lower accuracy for laterality, stage, and grade. Panel (B) shows average model accuracy across organs, highlighting strong performance on kidney and pancreas reports, with greater variability in cervix and bladder. Heatmaps in panels (C) and (D) provide a more detailed view of model performance across variable and organ dimensions. Top-performing models such as Phi-4, DeepSeek-R1:32B, and LLaMA 3.3 showed consistent accuracy across variables and organs. In contrast, Mixtral and Mistral-7B underperformed across multiple categories, particularly on grade and stage extraction.

Statistical analysis confirmed significant performance differences across models, variables, and organ types. A one-way ANOVA revealed overall differences in extraction accuracy across the eleven LLMs ($F = 19.29$, $p < 0.001$). A two-way ANOVA showed significant main effects for model ($F = 55.41$, $p < 0.001$) and variable ($F = 83.38$, $p < 0.001$), along with a strong interaction ($F = 9.66$, $p < 0.001$), indicating that no single model performed uniformly across all variables. A three-way ANOVA incorporating organ type revealed additional variation, with significant effects for organ ($F = 8.91$, $p < 0.001$) and multiple interaction terms (model × organ, variable × organ, and model × variable × organ), highlighting the influence of organ-specific language patterns on model accuracy. Post-hoc Tukey HSD tests stratified by variable identified multiple pairwise differences between models. Mixtral underperformed significantly on histology, laterality, and stage compared to higher-performing models like DeepSeek and Phi-4, while behavior and site were more consistently extracted across models.

TCGA Dataset: Expert Evaluation

Figure 3 presents the results from the expert evaluations based on manual review of 138 randomly selected TCGA pathology reports spanning five organ types: bladder, prostate, liver, cervix, and kidney. Each variable extracted by the reasoning-based aggregation model was independently scored by board-certified or equivalent pathologists on a three-point scale: 1 (correct), 0.5 (partially correct), or 0 (incorrect). Panel (A) shows overall extraction accuracy by organ, with kidney achieving the highest average score and cervix the lowest. As shown in panel (B), behavior, histology, and site were extracted with the highest consistency across organs, while laterality and stage showed greater variability. Panel (C) provides a detailed heatmap of organ-variable combinations. Kidney reports achieved perfect accuracy across all variables (mean = 1.00), while laterality in cervix (mean = 0.64) and stage in liver (mean = 0.73) were among the most challenging. Panels (D) and (E) summarize the same data by variable and by organ, respectively, reinforcing consistent performance on behavior and histology and greater organ-specific variability in laterality and stage.

Statistical analysis confirmed that extraction accuracy was significantly influenced by both the type of variable and the organ. Two-way ANOVA revealed main effects for variable ($F(5, 1195) = 16.95$, $p < 0.0001$) and organ ($F(4, 1195) = 3.94$, $p = 0.0035$), as well as a strong interaction effect between the two ($F(20, 1195) = 4.46$, $p < 0.0001$). Post-hoc Tukey HSD tests identified specific organ–variable pairs with significantly different performance. Laterality extraction was significantly more accurate in kidney and liver reports compared to cervix ($p = 0.005$ and $p = 0.0025$, respectively). For Stage, liver reports had lower accuracy than bladder ($p = 0.0002$), cervix ($p = 0.0073$), and kidney ($p = 0.0003$). No significant differences were found across organs for Behavior, Histology, Site, or Grade, indicating stable extraction performance for these variables across tumor types.

Moffitt Dataset: Automated Evaluation

Figure 4 presents the results from the automated evaluations of 510 pathology reports in the Moffitt dataset, covering brain, breast, and lung cancers. Each report was processed by eleven locally hosted LLMs and scored by the DeepSeek-R1:70B evaluator. Panel (A) shows variable-level accuracy across all models. Behavior, stage, and laterality were extracted with relatively high consistency, while grade and histology showed more variation between models. Panel (B) displays average accuracy across all six variables by organ, revealing relatively stable performance across brain, breast, and lung cases. Panels (C) and (D) present heatmaps of model accuracy by variable and by organ, respectively. High-performing models such as Phi-4, Tulu3:8B, and DeepSeek-R1:70B maintained robust accuracy across most

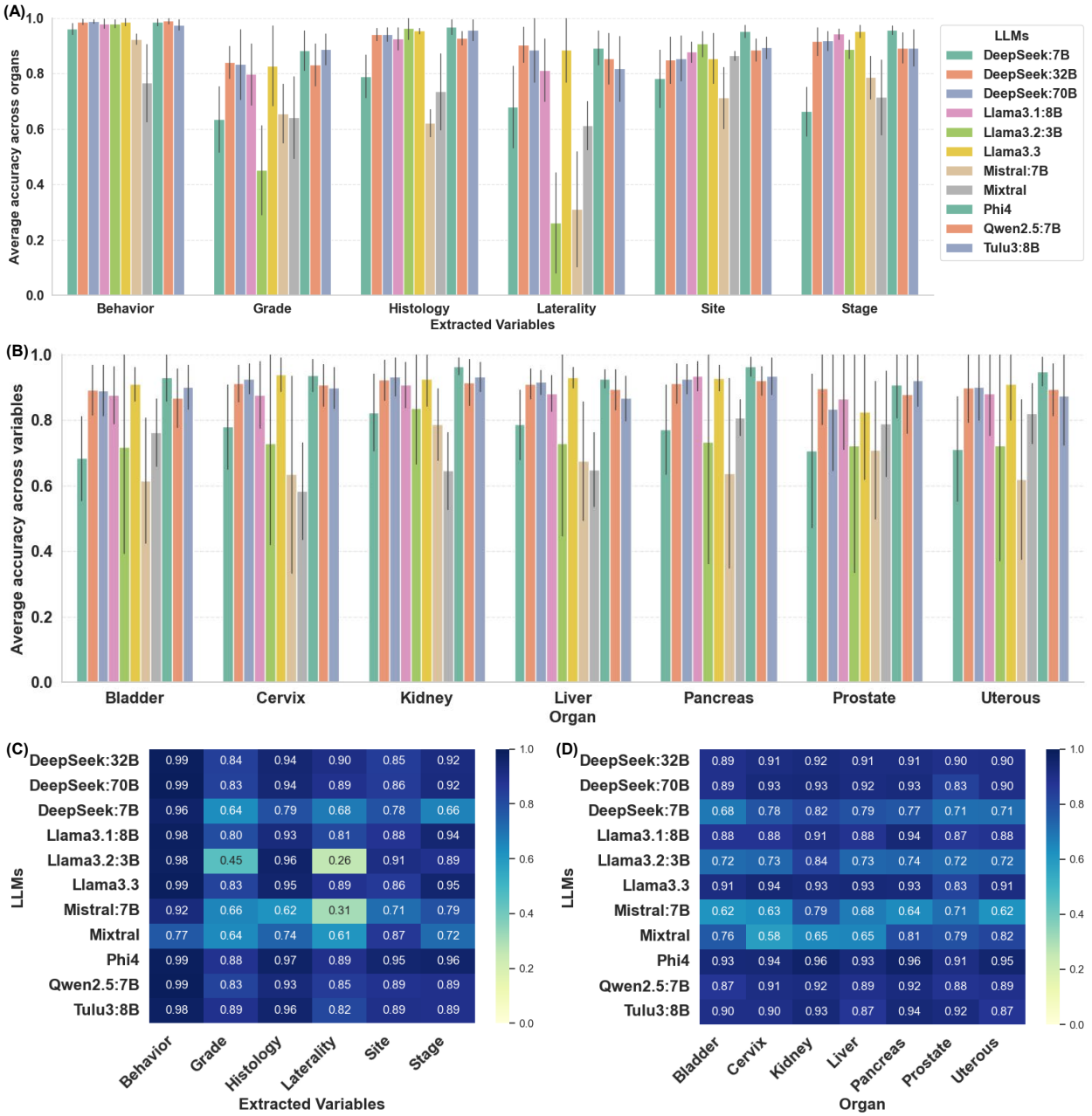

Figure 2: **Automated evaluation of variable extraction accuracy across 11 locally deployed LLMs on the TCGA dataset.** Each panel summarizes model performance across six diagnostic variables (site, laterality, histology, stage, grade, and behavior) and seven organ types (kidney, uterus, prostate, liver, bladder, cervix, and pancreas). Model outputs were scored by a reasoning-based evaluator (DeepSeek-R1:70B) based on alignment with report content and justification quality. (A) Mean accuracy per variable, averaged across organs and stratified by model. (B) Mean accuracy per organ, averaged across variables and stratified by model. (C) Heatmap of model-wise accuracy across diagnostic variables, highlighting consistent strengths and weaknesses across tasks. (D) Heatmap of model-wise accuracy across organ types, revealing performance differences attributable to organ-specific reporting conventions.

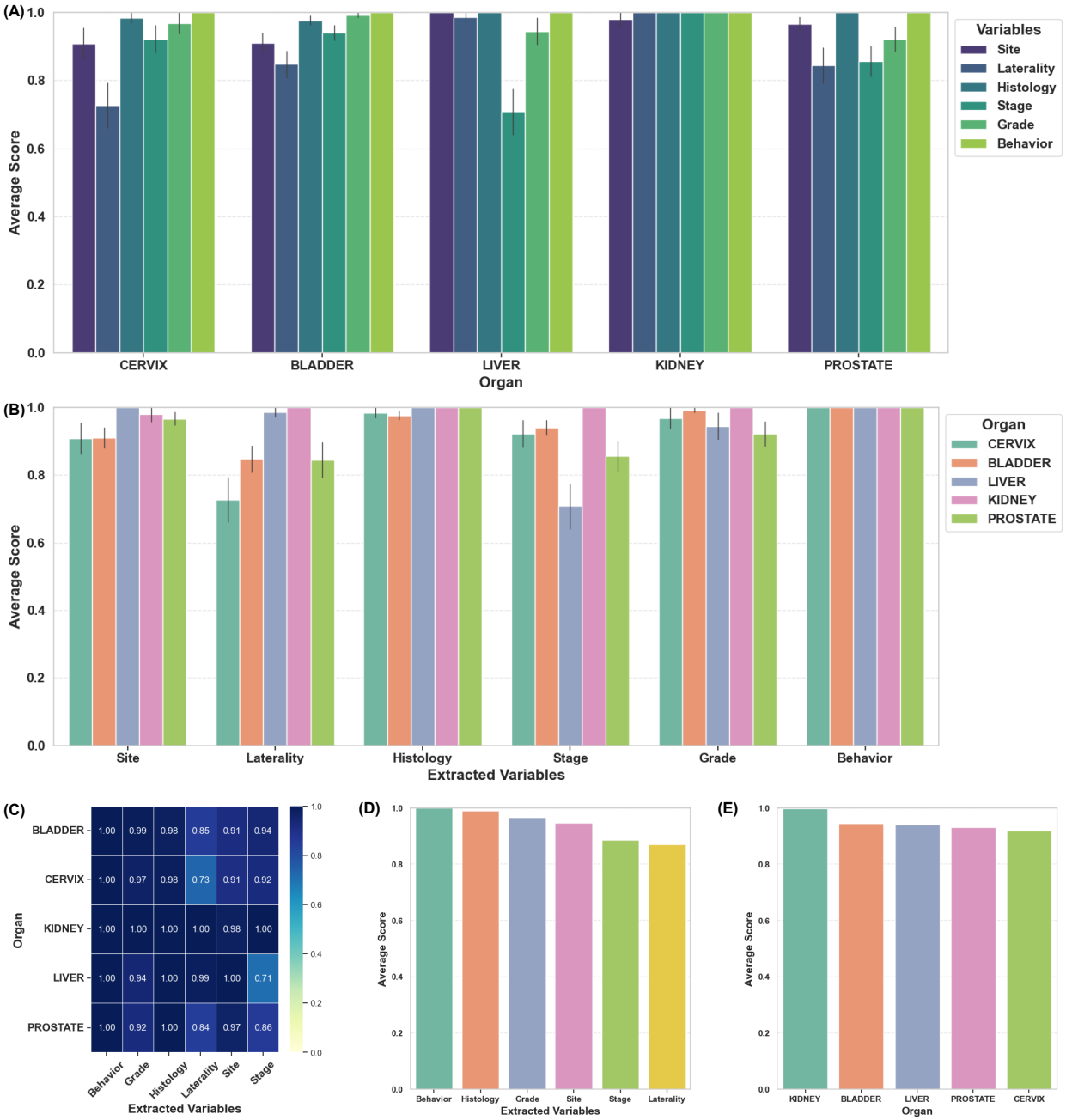

Figure 3: **Expert evaluation of structured variable extraction on TCGA surgical pathology reports.** A subset of 138 TCGA reports from five organ types (bladder, prostate, liver, cervix, and kidney) was reviewed by board-certified pathologists. Each extracted variable was scored on a 3-point scale: 1 (correct), 0.5 (partially correct), or 0 (incorrect), based on consistency with report content and clinical relevance. (B) Average accuracy by organ, stratified by variable. Kidney reports exhibited the highest overall accuracy across all variables, whereas cervix reports had the lowest, primarily due to reduced performance on laterality. (A) Average accuracy by variable, stratified by organ. Behavior, site, and histology were consistently extracted with high accuracy, while laterality and stage showed greater variability. (C) Heatmap showing average accuracy scores for each organ–variable pair. Perfect scores were observed across all variables in kidney reports. The lowest scores were seen in laterality for cervix (0.64) and stage for liver (0.73). (D) Variable-wise averages across all five organs, highlighting behavior and histology as the most reliably extracted fields. (E) Organ-wise averages across all six variables, with kidney and bladder showing the highest accuracy and cervix the lowest.

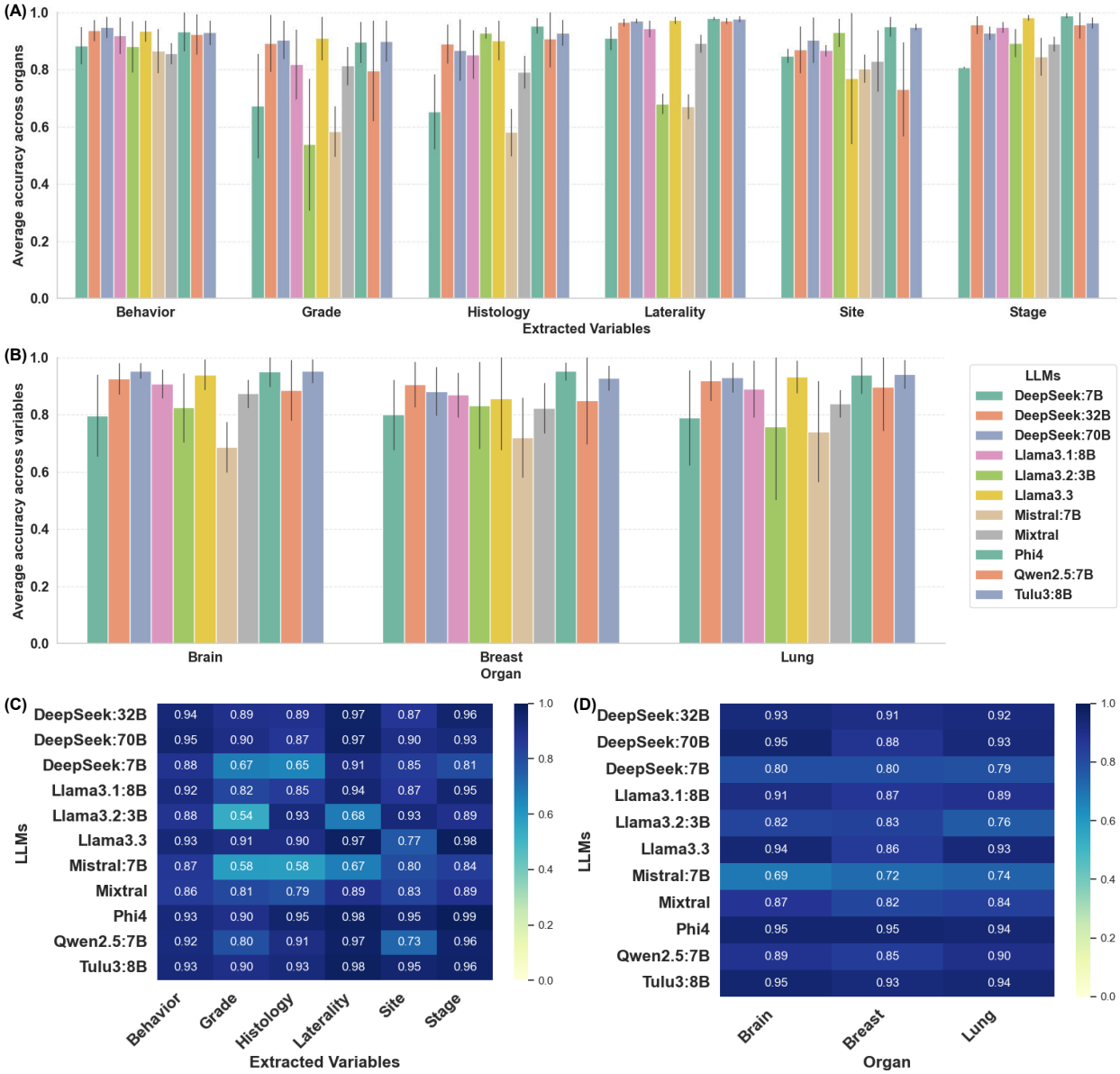

Figure 4: **Automated evaluation of structured variable extraction on the Moffitt pathology dataset.** This figure summarizes the accuracy of eleven locally deployed LLMs in extracting six diagnostic variables (site, laterality, histology, stage, grade, and behavior) from 510 surgical pathology reports covering brain, breast, and lung cancers. Outputs were evaluated by DeepSeek-R1:70B using the original report and model-provided justifications. (A) Average model accuracy per variable, aggregated across all organs. The highest consistency was observed in behavior and stage, while grade and histology exhibited greater inter-model variability. (B) Average model accuracy across all six variables, stratified by organ type. Accuracy was stable across brain, breast, and lung reports, with top-performing models maintaining high performance across organ sites. (C) Heatmap of model-wise accuracy by variable. (D) Heatmap of model-wise accuracy by organ.

variables and cancer types. In contrast, Mixtral and Mistral-7B showed reduced performance, particularly on grade and histology tasks. Compared to the TCGA dataset, organ-level accuracy was more uniform, suggesting reduced influence of reporting variation in this institutional cohort.

Statistical analysis confirmed significant differences in extraction accuracy across models and variables. A one-way ANOVA revealed a significant main effect of model ($F = 7.89$, $p < 0.001$), indicating performance heterogeneity among the eleven LLMs. Two-way ANOVA showed significant main effects for both model ($F = 13.21$, $p < 0.001$) and variable ($F = 11.43$, $p < 0.001$), along with a model × variable interaction ($F = 2.33$, $p < 0.001$), confirming that accuracy depended on the specific variable being extracted. When organ was added as a third factor, a main effect for organ was observed ($F = 7.84$, $p < 0.001$), as well as a significant interaction between variable and organ ($F = 17.69$, $p < 0.001$). However, no significant model × organ or three-way interaction effects were found, suggesting greater consistency in model performance across organs compared to the TCGA cohort. Post-hoc Tukey HSD tests showed that Mixtral underperformed significantly on histology, laterality, and stage compared to DeepSeek and other top models, while behavior and site remained consistently well-predicted across models.

Moffitt Dataset: Expert Evaluation

Figure 5 presents expert validation of final extracted outputs from 180 randomly selected pathology reports in the Moffitt dataset. Reports spanned three cancer types: brain, breast, and lung. Each variable extracted by the aggregation model was independently scored by three board-certified pathologists using a 3-point scale: 1 (correct), 0.5 (partially correct), or 0 (incorrect). As shown in panel (A), accuracy varied by organ and variable. Behavior, histology, and laterality were consistently extracted with high reliability, while site and stage showed greater variability, particularly in lung and breast cases. Panel (B) displays the same data stratified by variable, revealing that behavior had the highest overall score, followed by histology and laterality. Panel (C) shows a heatmap of mean accuracy for each organ–variable pair, highlighting two notably lower-performing combinations: site in lung (0.80) and stage in breast (0.87). Panels (D) and (E) summarize average scores by variable and organ, respectively. Breast reports received the highest overall scores across all variables, followed by brain and lung.

The results from the two-way ANOVA confirmed that extraction accuracy in the Moffitt dataset was statistically significantly affected by variable type ($F(5, 997) = 4.05$, $p = 0.003$) and by the interaction between variable and organ ($F(10, 997) = 3.05$, $p < 0.001$). The main effect of organ alone was not statistically significant ($F(2, 997) = 2.73$, $p = 0.066$), indicating that overall accuracy was comparable across brain, breast, and lung reports. Post-hoc Tukey HSD tests identified specific organ–variable pairs with reduced performance. Site extraction was significantly more accurate in brain and breast reports than in lung ($p = 0.017$ and $p = 0.039$, respectively). Behavior was also more accurately extracted in brain compared to lung ($p = 0.047$). No significant organ-level differences were observed for histology, laterality, stage, or grade, suggesting consistent performance for these variables across cancer types.

4 Discussion

This study demonstrates that reasoning-based LLMs, when locally deployed and guided through structured prompting, can accurately and transparently extract key diagnostic variables from free-text surgical pathology reports. Using a three-stage framework applied to over 4,000 reports from eight cancer sites, including a large institutional cohort from Moffitt Cancer Center, we achieved high overall accuracy in extracting six clinically relevant fields: site, laterality, histology, stage, grade, and tumor behavior. Automated evaluation using an independent LLM adjudicator identified consistent accuracy trends, with histology, site, and behavior emerging as the most reliably extracted variables. These findings were corroborated by expert review from board-certified pathologists (or pathologists holding specialist qualifications outside the US), who confirmed high fidelity across most variables and noted specific challenges for laterality and stage in select organ contexts. Together, these results support the technical robustness, clinical validity, and real-world adaptability of our consensus-based extraction framework for routine and research applications in surgical pathology.

Differences in extraction accuracy across diagnostic variables reflect both model-level challenges and variation in how clinical information is recorded across pathology reports. Variables such as behavior and histology were consistently extracted with high accuracy, likely due to their relatively standardized representation and clearer phrasing in most documents. In contrast, stage and laterality were more error-prone, particularly in cervix, liver, and lung cases, where these details are often embedded in narrative descriptions or distributed across sections of the report rather than summarized explicitly [34]. Expert reviewers noted that even for human abstractors, accurate extraction of these variables can require inferential reasoning or synthesis of scattered findings. These patterns underscore that model performance is shaped not only by algorithmic capabilities but also by the linguistic and documentation norms of specific organ systems and pathology subspecialties. To be clinically reliable, AI-based extraction frameworks must

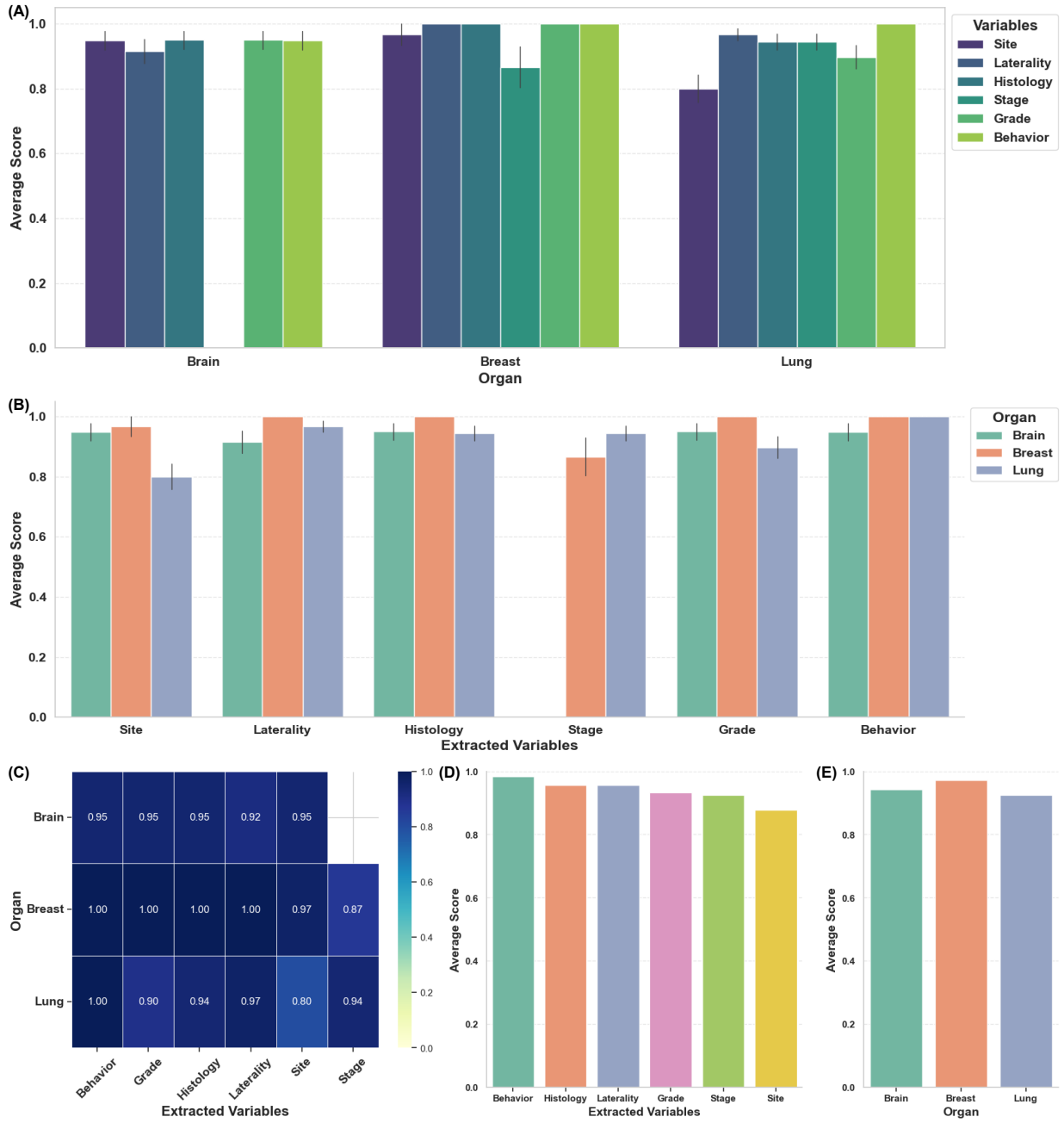

Figure 5: **Expert evaluation of structured variable extraction on Moffitt pathology reports.**. A total of 180 reports from brain, breast, and lung cancer cases were reviewed by board-certified pathologists. Each of the six extracted variables-site, laterality, histology, stage, grade, and behavior-was scored on a 3-point scale (1 = correct, 0.5 = partially correct, 0 = incorrect) based on alignment with the report content and clinical interpretation. (A) Average accuracy per variable stratified by organ. Behavior, histology, and laterality showed the highest consistency, while site and stage exhibited more variability, especially in lung and breast cases. (B) Average accuracy per organ stratified by variable. Breast reports achieved the highest average scores, followed by brain and lung. (C) Heatmap showing mean accuracy scores for each organ–variable pair. Site in lung (0.80) and stage in breast (0.87) were the lowest-performing combinations. (D) Variable-wise accuracy averaged across all organs. Behavior and histology were extracted with the highest overall accuracy. (E) Organ-wise accuracy averaged across all six variables, with breast performing best overall and lung showing slightly lower consistency.

therefore adapt to these real-world variations in reporting style, especially for use in applications such as cancer registry abstraction and structured synoptic reporting.

Performance on the internal Moffitt dataset, which included brain, breast, and lung pathology reports, further demonstrated the robustness of the framework across institutional settings and cancer types. This cohort presented unique challenges due to variability in documentation style, particularly in breast cancer reports, where stage and grade were often embedded in narrative assessments or ancillary sections rather than clearly labeled fields. Despite this complexity, the framework achieved high extraction accuracy, underscoring its adaptability to diverse real-world reporting. Statistical analyses revealed significant interactions between model, variable, and organ in the TCGA dataset, with laterality and stage showing the strongest variability across organ types. In contrast, model performance on the Moffitt dataset was more consistent, though specific models (e.g., Mixtral) still underperformed on behavior extraction. These findings emphasize the need for systematic, context-aware model validation, particularly in heterogeneous, multi-organ environments. The adjudication stage played a critical role in mitigating variability by synthesizing outputs from eleven LLMs and generating consensus predictions based on justification quality and inter-model agreement [35]. This consensus-driven process virtually mirrors the clinical practice of reconciling multiple expert interpretations and supports transparent, auditable decision-making, an essential feature for deploying AI in complex and institution-specific pathology workflows.

Expert review of extracted outputs offered critical insights into the clinical context underlying model performance. Across both the TCGA and Moffitt datasets, pathologists frequently observed that apparent "errors" by the model often stemmed from ambiguity or inconsistency in the source documentation, rather than algorithmic failure. For instance, laterality was flagged as incorrect in cases where it was either not clinically applicable (e.g., cervix, bladder) or where bilateral disease was noted without specifying which side the specimen represented. Stage misclassifications similarly arose from the use of various staging schemes or when reports lacked complete TNM information, such as missing nodal or metastatic status or omitted descriptors like "Mx" or "Nx", which also challenged expert interpretation. In many of these scenarios, reviewers noted that the model's output reflected a plausible clinical inference, even if it did not fully align with reporting conventions. These observations underscore that diagnostic accuracy in automated systems must be interpreted in the broader context of real-world variability in documentation practices.

Expert review of the Moffitt dataset (breast, brain, and lung cancers) uncovered organ-specific reasoning challenges that were not evident through automated evaluation alone. In breast cancer reports, pathologists frequently noted that sentinel lymph node status, denoted by markers such as "(sn)" or "(i-)", was omitted from predicted stage values and inadequately addressed in model justifications. Brain tumor cases revealed difficulty with accurate site and laterality assignment, especially when specimens were labeled differently from surgical targets or when multiple involved sites were present. In lung and metastatic cases, the most common issue was confusion between the biopsy site (e.g., lymph node or liver) and the presumed primary site (e.g., lung), occasionally resulting in under-classification of behavior or grade. Reviewers also observed that even when variable values were correct, model rationales often failed to highlight key clinical cues, limiting transparency and auditability. These findings reinforce the importance of human-in-the-loop validation and suggest that rationale filtering may improve model alignment with expert expectations in complex diagnostic contexts.

A central innovation of this framework lies in its integration of explicit reasoning throughout the extraction process. Rather than relying on black-box predictions, each LLM is prompted to generate not only a structured variable value but also a brief textual justification. These justifications are then evaluated by an independent reasoning model, enabling a second layer of validation that emphasizes coherence and clinical plausibility. This approach brings interpretability into a domain where opaque AI outputs are unacceptable and auditability is critical. The final aggregation stage synthesizes responses from eleven models, prioritizing inter-model agreement and justification quality, an approach that mirrors how clinical consensus is reached in pathology through the reconciliation of multiple expert viewpoints. By aligning technical design with established clinical reasoning practices, the system produces outputs that are both reliable and explainable. For users such as pathologists and cancer registrars, this design supports a shift from passive oversight to active validation, where the rationale behind each suggestion is transparent and subject to expert judgment [35].

Beyond model performance, the framework is designed with real-world clinical deployment in mind, particularly for environments with stringent data privacy and governance requirements. By utilizing locally hosted, open-source LLMs, the system enables HIPAA-compliant inference within institutional firewalls, avoiding the need to transmit protected health information to external servers [36]. This makes it well-suited for integration into hospital infrastructure, including pathology departments and cancer centers. The structured outputs generated by the adjudication model are readily compatible with downstream tasks such as cancer registry abstraction, synoptic report generation, and clinical trial screening. Each extracted variable is accompanied by a concise rationale, allowing pathology staff to efficiently review and approve outputs in human-in-the-loop workflows. This balance between automation and interpretability is especially critical in oncology, where diagnostic precision has direct implications for treatment, staging, and prognosis.

By embedding transparency and alignment with clinical practice into the system design, the framework supports safe, scalable adoption of LLMs in routine pathology workflows.

While this study demonstrates the reliability and clinical relevance of a reasoning-based LLM framework for structured pathology report extraction, several limitations warrant consideration. First, although strong performance was observed across multiple organ sites, the generalizability of the system to rare cancers, atypical documentation formats, or resource-constrained settings remains to be evaluated. Second, the adjudication process, though designed to promote transparency, relies on the reasoning model's assessment of justification quality, which may introduce its own biases, especially in cases involving ambiguous or incomplete language. Third, expert validation, while essential, was necessarily limited to a subset of reports due to the labor-intensive nature of manual review. Future work will explore the integration of uncertainty estimation, active learning for iterative refinement, and expansion to additional variables such as biomarkers, treatment response, or recurrence status. Prospective validation within live clinical environments, including deployment into laboratory information systems and registry pipelines, will be critical for assessing operational impact on accuracy, workflow efficiency, and downstream decision-making [29]. Addressing these challenges will be key to realizing the full potential of this framework as a scalable, trustworthy foundation for AI-enabled pathology practice.

5 Conclusion

This study introduces a robust, locally deployable, consensus-based framework for extracting structured diagnostic variables from surgical pathology reports using reasoning-augmented LLMs. By combining multi-model prompting, justification-driven evaluation, and transparent adjudication, the system achieves high accuracy and interpretability across diverse cancer types and reporting styles. Expert validation confirmed strong alignment with clinical expectations, particularly for reliably documented fields such as histology, site, and behavior. Importantly, the framework operates entirely within institutional infrastructure, supporting privacy-preserving deployment and seamless integration into workflows such as cancer registry abstraction and synoptic reporting. These findings highlight the promise of explainable, reasoning-based AI systems in transforming unstructured medical documentation into clinically actionable data at scale.

Acknowledgments

This research was supported in part by the Department of Pathology at the H. Lee Moffitt Cancer Center & Research Institute; the Cancer Center Support Grant (P30-CA76292) awarded by the NIH/NCI to the H. Lee Moffitt Cancer Center & Research Institute; a Florida Biomedical Research Grant (21B12); an NIH/NCI grant (U01-CA200464); NSF Awards 2234836 and 2234468; and NSF NAIRR pilot funding.

References

[1] James L Connolly, Stuart J Schnitt, Helen H Wang, Janina A Longtine, Ann Dvorak, and Harold F Dvorak. Role of the Surgical Pathologist in the Diagnosis and Management of the Cancer Patient. In *Holland-Frei Cancer Medicine. 6th edition*. BC Decker, 2003.

[2] NIH. Surgical Pathology Reports. `https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/pathology-reports-fact-sheet`. [Accessed 12-04-2025].

[3] Dr Jose Candido Caldeira Xavier-Junior. The Role of a Pathologist in Diagnosis and Treatment: What Does a Pathology Report Contain? - The Waiting Room. `https://thewaitingroom.karger.com/tell-me-about/the-role-of-a-pathologist-in-diagnosis-and-treatment-what-does-a-pathology-report-contain/`. [Accessed 12-04-2025].

[4] NIH. Surveillance, Epidemiology, and End Results Program. `https://seer.cancer.gov/`. [Accessed 12-04-2025].

[5] CDC. National Program of Cancer Registries. `https://www.cdc.gov/national-program-cancer-registries/index.html`. [Accessed 12-04-2025].

[6] Paul N. Valenstein. Formatting Pathology Reports: Applying Four Design Principles to Improve Communication and Patient Safety. *Archives of Pathology and Laboratory Medicine*, 132(1):84–94, January 2008.

[7] Ken G. Zeng, Tarun Dutt, Jan Witowski, Kranthi Kiran GV, Frank Yeung, Michelle Kim, Jesi Kim, Mitchell Pleasure, Christopher Moczulski, L. Julian Lechuga Lopez, Hao Zhang, Mariam Al Harbi, Farah E. Shamout, Vincent J. Major, Laura Heacock, Linda Moy, Freya Schnabel, Linda M. Pak, Yiqiu Shen, and Krzysztof J. Geras. Improving Information Extraction from Pathology Reports using Named Entity Recognition. *pre-print*, July 2023.

[8] Aakash Tripathi, Asim Waqas, Kavya Venkatesan, Ehsan Ullah, Marilyn Bui, and Ghulam Rasool. 1391 ai-driven extraction of key clinical data from pathology reports to enhance cancer registries. *Laboratory Investigation*, 105(3), 2025.

[9] Ryzen Benson, Candace Winterton, Maci Winn, Benjamin Krick, Mei Liu, Noor Abu-el rub, Mike Conway, Guilherme Del Fiol, Andrew Gawron, and Sheetal Hardikar. Leveraging natural language processing to extract features of colorectal polyps from pathology reports for epidemiologic study. *JCO Clinical Cancer Informatics*, 1(7), January 2023.

[10] Pilar López-Úbeda, Teodoro Martín-Noguerol, José Aneiros-Fernández, and Antonio Luna. Natural Language Processing in Pathology. *The American Journal of Pathology*, 192(11):1486–1495, November 2022.

[11] Phillip Park, Yeonho Choi, Nayoung Han, Ye-Lin Park, Juyeon Hwang, Heejung Chae, Chong Woo Yoo, Kui Son Choi, and Hyun-Jin Kim. Leveraging natural language processing for efficient information extraction from breast cancer pathology reports: Single-institution study. *PLOS ONE*, 20(2):e0318726, February 2025.

[12] Yoojoong Kim, Jeong Hyeon Lee, Sunho Choi, Jeong Moon Lee, Jong-Ho Kim, Junhee Seok, and Hyung Joon Joo. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. *Scientific Reports*, 10(1), November 2020.

[13] OpenAI. GPT-4 Technical Report, 2024.

[14] Mohamed Omar, Varun Ullanat, Massimo Loda, Luigi Marchionni, and Renato Umeton. ChatGPT for digital pathology research. *The Lancet Digital Health*, 6(8):e595–e600, 2024.

[15] Hooman H. Rashidi, Joshua Pantanowitz, Alireza Chamanzar, Brandon Fennell, Yanshan Wang, Rama R. Gullapalli, Ahmad Tafti, Mustafa Deebajah, Samer Albahra, Eric Glassy, Matthew G. Hanna, and Liron Pantanowitz. Generative Artificial Intelligence in Pathology and Medicine: A Deeper Dive. *Modern Pathology*, 38(4):100687, April 2025.

[16] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology*, 19(1), February 2024.

[17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[18] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025.

[19] Asim Waqas, Marilyn M. Bui, Eric F. Glassy, Issam El Naqa, Piotr Borkowski, Andrew A. Borkowski, and Ghulam Rasool. Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. *Laboratory Investigation*, 103(11):100255, November 2023.

[20] Asim Waqas, Aakash Tripathi, Ravi P Ramachandran, Paul A Stewart, and Ghulam Rasool. Multimodal data integration for oncology in the era of deep neural networks: A Review. *Frontiers in Artificial Intelligence*, 7:1408843, 2024.

[21] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024.

[22] Savyasachi V. Shah. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953, August 2024.

[23] Ken Cheligeer, Guosong Wu, Alison Laws, May Lynn Quan, Andrea Li, Anne-Marie Brisson, Jason Xie, and Yuan Xu. Validation of large language models for detecting pathologic complete response in breast cancer using population-based pathology reports. *BMC Medical Informatics and Decision Making*, 24(1):283, 2024.

[24] Ian E. Nielsen, Dimah Dera, Ghulam Rasool, Ravi P. Ramachandran, and Nidhal Carla Bouaynaya. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4):73–84, July 2022.

[25] Daniel Truhn, Chiara ML Loeffler, Gustav Müller-Franzes, Sven Nebelung, Katherine J Hewitt, Sebastian Brandner, Keno K Bressem, Sebastian Foersch, and Jakob Nikolas Kather. Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (gpt-4). *The Journal of Pathology*, 262(3):310–319, 2024.

[26] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, 7, November 2024.

[27] Victor Brodsky, Ehsan Ullah, Andrey Bychkov, Andrew H Song, Eric E Walk, Peter Louis, Ghulam Rasool, Rajendra S Singh, Faisal Mahmood, Marilyn M Bui, et al. Generative artificial intelligence in anatomic pathology. *Archives of Pathology & Laboratory Medicine*, 2025.

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

[29] Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, et al. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45, 2025.

[30] Lei Wang, Yinyao Ma, Wenshuai Bi, Hanlin Lv, and Yuxiang Li. An entity extraction pipeline for medical text records using large language models: Analytical study. *J Med Internet Res*, 26:e54580, Mar 2024.

[31] Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K. Keloth, Vincent J. Zhang, Ruey-Ling Weng, Qingyu Chen, Xiaoqian Jiang, Kirk E. Roberts, and Hua Xu. Information extraction from clinical notes: Are we ready to switch to large language models?, 2025.

[32] Asim Waqas, Ehsan Ullah, Asma Khan, Farah Khalil, Zarifa Gahramanli Ozturk, Vaibhav Chumbalkar, Daryoush Saeed-Vafa, Zena Jameel, Weishen Chen, Humberto Trejo Bittar, Jasreman Dhillon, Rajendra S. Singh, Andrey Bychkov, Anil V. Parwani, Marilyn M. Bui, Matthew B. Schabath, and Ghulam Rasool. Reasoning beyond accuracy: Expert evaluation of large language models in diagnostic pathology. *Preprint*, April 2025.

[33] Jenna Kefeli and Nicholas Tatonetti. TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models. *Patterns*, 5(3):100933, March 2024.

[34] Jingwei Huang, Donghan M Yang, Ruichen Rong, Kuroush Nezafati, Colin Treager, Zhikai Chi, Shidan Wang, Xian Cheng, Yujia Guo, Laura J Klesse, et al. A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj Digital Medicine*, 7(1):106, 2024.

[35] Gerald Gartlehner, Leila Kahwati, Barbara Nussbaumer-Streit, Karen Crotty, Rainer Hilscher, Shannon Kugley, Meera Viswanathan, Ian Thomas, Amanda Konet, Graham Booth, and Robert Chew. From promise to practice: challenges and pitfalls in the evaluation of large language models for data extraction in evidence synthesis. *BMJ Evidence-Based Medicine*, pages bmjebm–2024–113199, December 2024.

[36] Isabella Catharina Wiest, Fabian Wolf, Marie-Elisabeth Leßmann, Marko van Treeck, Dyke Ferber, Jiefu Zhu, Heiko Boehme, Keno K. Bressem, Hannes Ulrich, Matthias P. Ebert, and Jakob Nikolas Kather. Llm-aix: An open source pipeline for information extraction from unstructured medical text based on privacy preserving large language models. *medRxiv*, 2024.