Large language models for extracting histopathologic diagnoses from electronic health records

Brian Johnson[1], Tyler Bath[1], Xinyi Huang[2], Mark Lamm[2], Ashley Earles[2], Hyrum Eddington[1], Lily J. Jih[3,4], Samir Gupta[3,5,6], Shailja C. Shah[3,5,6], Kit Curtius[1,3,6*]

1. Division of Biomedical Informatics, Department of Medicine, University of California San Diego, La Jolla, CA, USA
2. Veterans Medical Research Foundation, San Diego, CA, USA
3. VA San Diego Healthcare System, San Diego, CA, USA
4. Department of Pathology, University of California, La Jolla, CA, USA
5. Division of Gastroenterology, University of California, La Jolla, CA, USA
6. Moores Cancer Center, University of California San Diego, La Jolla, CA, USA

Correspondence to: kcurtius@health.ucsd.edu

Author contributions
Study conception and design: KC, BJ
Data collection and curation: BJ, TB, XH, ML, AE, HE, LJJ
Model development and formal analyses: BJ, KC, SG, SCS
Paper writing: BJ, KC, SG, SCS

Conflict of interest statement
KC has an investigator-led research grant from Phathom Pharmaceuticals. SCS is a paid ad hoc consultant for RedHill Biopharma and Phathom Pharmaceuticals, and unpaid scientific advisory board member for Ilico Genetics, Inc.

Data availability
A CSV (all_results.csv) with results from all validated runs, including additional details such as prevalence of model positives, exact numbers validated, and the full confusion matrix, will be made available in the supplementary. This CSV is the source for all three main text tables and Supplementary Tables S3-S5. Raw data access is reserved for VA investigators with appropriate research approvals.

Code availability
Llama.cpp fork and our specific implementation will be made available on Github, which includes all prompts and custom parameters required to reproduce our work. Large language models, in gguf format, are stored and publicly available at Hugging Face.

Abstract

Background & Aims

Accurate data resources are essential for impactful medical research. To date, most large-scale studies have relied on structured sources, such as International Classification of Diseases codes, to determine patient diagnoses and outcomes. However, these structured datasets are often incomplete or inaccurate. Recent advances in natural language processing, specifically the introduction of open-weight large language models (LLMs), enable more accurate data extraction from unstructured text in electronic health records (EHRs).

Methods

We created an approach using LLMs for identifying histopathologic diagnoses, including presence of dysplasia and cancer, in pathology reports from the Department of Veterans Affairs Healthcare System, including those patients with genotype data within the Million Veteran Program (MVP) biobank. Our approach requires no additional training and utilizes a simple 'yes/no' question prompt to obtain an answer. We validated the method on 3 diagnostic tasks by applying the same prompts to reports from patients with vs without diagnoses of inflammatory bowel disease (IBD) and calculating F-1 scores as a balanced accuracy measure.

Results

In patients without IBD in MVP, we achieved F1-scores of 99.3% for identifying any dysplasia, 98.2% for identifying high-grade dysplasia and/or colorectal adenocarcinoma (HGD/CRC), and 96.2% for identifying CRC using LLM Gemma-2. In IBD patients in MVP, we achieved F1-scores of 97.1% for identifying dysplasia, 96.4% for identifying HGD/CRC, and 97.1% for identifying CRC.

Conclusion

LLMs provide excellent accuracy in extracting diagnoses from EHRs and can be applied to a variety of tasks with no additional human-led development required. Our validated methods generalized to unstructured pathology notes, even withstanding challenges of resource-limited computing environments.

Keywords: artificial intelligence, large language models, natural language processing, biomedical informatics, colorectal cancer, inflammatory bowel disease

Introduction

The expected breakthroughs in personalized treatments and improved medical outcomes have yet to fully materialize despite the exponential increase in volume of healthcare data available for research. One obstacle impeding these advances is the quality and accessibility of the vast data generated and stored as part of usual healthcare.

As an example use case, tailoring colonoscopy screening ages and surveillance intervals based on accurate risk stratification informed by large, high-quality datasets has real potential to reduce both the incidence of colorectal cancer, as well as the number of unnecessary colonoscopies that add burden to both patients and the healthcare system[1]. Current risk stratification approaches in both the general population and those with inflammatory bowel disease (IBD) are based on few clinical variables, which are often associated with widely varied published estimates of risk[2,3]. For example, when low-grade dysplastic lesions are diagnosed, current guidelines for all-comers recommend surveillance colonoscopy every 1-10 years (or 1-5 years in patients with IBD) based on clinical risk stratification[3,4]. As such, screening guidelines essentially serve as heuristics, representing the best approach in a climate of such limited data[5]. Our overall goal is to improve the quality of available large-scale data resources—an essential prerequisite for accurate downstream analyses to improve personalized medicine—by leveraging artificial intelligence to extract clinical information, with a focus on histopathologic diagnoses in the present study.

Traditionally, “rules-based” natural language processing (NLP) algorithms have dominated structured data extraction in this area. Briefly, NLP translates natural language into data formats that are easier for computers to process and analyze. Performance of these algorithms is excellent, with F1-scores frequently above 99% for identifying adenomas in data from the general population[6–13]. Alternatively, published deep-learning or embedding-based models are less common for classifying pathology findings, though work by Syed et al. reported an F1-score of 95% for identifying neoplastic (dysplastic) polyps[14]. However, there are many drawbacks to current approaches. Development requires extensive human effort to refine algorithms: creating concepts (e.g., enumerating all possible ways that each diagnosis could be written), identifying negation (ensuring that expressions of the absence or uncertainty of a diagnosis are captured and related to the correct concept), associating terms with their respective anatomical locations, and modifying the algorithm to address “edge” cases. Adapting these algorithms to new use cases or different databases presents similar challenges, as development is often tailored to the formatting and style of a specific hospital system, patient cohort, and/or time period[15]. For example, pathology reports from patients with IBD are often slightly different from reports from the non-IBD population. In the IBD patient cohort, historical terminology may be present (e.g., “DALMs”), non-targeted biopsies make up a large fraction of the biopsy specimens[4], and there are more instances of negation where pathologists explicitly rule out dysplasia and carcinoma than in patients without IBD. To our knowledge, previous NLP approaches have only been tested in the general population, where patients with IBD diagnoses represent a small fraction of the cohort or are removed from the cohort. Additionally, few of these approaches have been rigorously tested in identifying varying severities of dysplasia (e.g., low-grade, high-grade) and

adenocarcinoma. Where they have been tested rigorously, performance for adenocarcinoma is lower (F1 score < 90%) compared to performance for adenoma or dysplasia[8].

Large language models (LLMs) are capable of many tasks “out of the box” without additional tedious human-led development. Because of this, LLMs should be less susceptible to differences in formatting and style, working better across settings. For example, applying LLMs to determine colonoscopy follow-up time recommendations that are in line with established guidelines has been shown to be feasible without task-specific training[16]. Recently, there have also been remarkable advances in the quality of open-weight LLMs released under permissive licenses. Open-weight models avoid critical legal and regulatory issues, allowing researchers to conduct inference without significant privacy risks. Specifically, these models can be uploaded to the same computing environment where the data is stored, enabling researchers to structure data without it leaving this secured space. Because these models do not undergo training or fine-tuning within the computing environment, there is no risk of them 'remembering' patient information or accidental data breaches. Moreover, using these models does not involve sending data to third parties like OpenAI, avoiding the associated bureaucratic and privacy challenges. While current LLMs require significant computational bandwidth, they are rapidly becoming viable alternatives for large-scale applications as their efficiency improves and computational costs decline.

Here, we test and compare the performance of LLMs, without any task-specific training, on their ability to extract and characterize the presence vs. absence of dysplasia and adenocarcinoma from unstructured colonoscopy-associated pathology reports. We show that LLMs, even in resource-limited environments, are accurate at identifying features from pathology reports in a way that is easily reproducible.

Materials and Methods

Patient databases

We applied our methods to data from the Veterans Health Administration (VHA), one of the largest integrated health systems in the US. The Corporate Data Warehouse (CDW) in the Veterans Affairs (VA) contains all electronic health record (EHR) data from Veteran healthcare encounters, including notes, International Classification of Diseases (ICD) codes, and other registries such as the National Death Index (NDI) that can all be used and is intended for research purposes. In total, our provisioned CDW database contains the EHRs from 15.2 million current and former patients cared for through the VHA. This consists of roughly 6.2 billion notes, with a mean of more than 400 notes per patient. The earliest notes and other data relevant for our purposes, such as ICD codes, date back to around the year 2000, when records began to be consistently stored digitally.

The Million Veteran Program (MVP) is a research initiative where Veterans volunteer to have additional health, survey, and complete genetic data collected and made available for research in an anonymized way. To date, over one million Veterans have volunteered to become a part of this initiative. Our provisioned dataset contains 913,318 patients (v22 data used in Results).

Veteran volunteers in MVP are demographically similar to patients in CDW and are a representative subsample[17], though re-identification or any linking of clinical data is strictly disallowed to protect the privacy of MVP participants. Therefore, cohort building must be done in each dataset separately. Further, any overlap of patients and/or notes between MVP and CDW datasets is due to random chance and is not known to researchers.

VINCI workspaces

The VA Informatics and Computing Infrastructure (VINCI) is a platform where researchers can access clinical data from both CDW and MVP. VINCI allows researchers to analyze these data sources using various computational resources in a secured environment. Within VINCI, structured and unstructured free text data are organized in SQL tables. R and Rstudio were used to process text from SQL tables into ".txt" files for reading into compiled C++ software, which was a fork we created of the open source llama.cpp GitHub project[18]. Without access to a Graphics Processing Unit (GPU) for accelerated LLM inference, we used the standard 4-core CPUs available on VINCI development workspaces.

VA Pathology Domain

The VINCI team has created a dataset domain called the Pathology Domain, which takes full pathology reports and extracts certain sections based on their appropriate section headers. The resulting table contains columns representing each section header, with the associated text for each note. Our method uses the columns "Specimen", to determine if the report described tissue from the colon or rectum, and "Microscopic Exam", to extract the diagnosis.

The Pathology domain is available in CDW to approved VINCI researchers. In MVP, the Pathology Domain data must be requested. We requested all Microscopic Exam and Specimen sections where Specimen had one of the following terms: "rectum", "colorectal", "rectal", "cecum", "colon", "hepatic flexure", "ileocecal valve", "rectosigmoid", "splenic flexure", or "colonic flexure". This term search was not case-sensitive and used word boundaries to identify terms.

Large Language Models

The main model used is Gemma-2-9B-It-SPPO[19,20], referred to herein as Gemma-2 (9 billion parameters). We also applied Llama-3-8B-Instruct[21], referred to herein as Llama-3 (8 billion parameters), to all tasks/cohorts. All models used have licenses that allow commercial and research use, as required by VA policy. All models were run as ".gguf" files, the all-in-one file format used by llama.cpp[18]. For details, see Supplementary Information section "*Large Language Model (LLM) selection*".

Identifying colonoscopy pathology reports in the Pathology Domain and full pathology notes
As mentioned in *VA Pathology Domain*, the partitioned data in the MVP pathology domain was filtered to only include those where the Specimen matches colon or rectum terms. In CDW, we apply the same terms matching: "rectum", "colorectal", "rectal", "cecum", "colon", "hepatic

flexure", "ileocecal valve", "rectosigmoid", or "splenic flexure". As in MVP, this term search is not case-sensitive. Differing slightly from MVP, we did not use word boundaries, instead using wildcards before and after each term. This led to a corpus of relevant notes for our study totaling n=2,899,321 reports from 1,834,930 unique patients in CDW and n=279,964 reports from 170,806 unique patients in MVP. For identifying the full pathology reports, we linked the Pathology Domain table entry to the TIU (Text Integration Utilities) note containing the full text of the corresponding pathology report, where available.

Tasks

Our approach identified the presence or absence of the following three concepts (clinical conditions) on colonoscopy-associated pathology reports: any dysplasia, high-grade dysplasia and/or any adenocarcinoma (HGD/CRC), and invasive adenocarcinoma (CRC). Our definitions for each of these three concepts is as follows:

Any dysplasia: Presence of any dysplasia in the colon or rectum explicitly stated in the report (e.g., low-grade, mild, moderate, high-grade etc.). Presence of any adenoma or adenomatous lesions in the colon or rectum, excluding sessile serrated adenoma unless there is an explicit statement of sessile serrated adenoma with dysplasia. The current application excludes “indefinite for dysplasia” and other uncertain phrases. Adenomas were counted in definition of any dysplasia because all adenomas contain at least low-grade dysplasia.

HGD/CRC: Presence of HGD or any adenocarcinoma in the colon or rectum. Includes adenocarcinoma in-situ and intramucosal adenocarcinoma. Excludes uncertain phrases such as “bordering on high-grade dysplasia”.

CRC: Presence of invasive adenocarcinoma of the colon or rectum. Invasive is defined as T stage of 1 or greater, or equivalent language (e.g., “invades into submucosa”). Excludes metastatic adenocarcinoma suspected or known to be from a different primary location (i.e. primary is not colon or rectum). Excludes uncertain phrases such as “cannot rule out invasive adenocarcinoma” and “suspicious for invasion”.

CRC plausible set

We then used simple search terms to reduce the number of colonoscopy pathology reports to only include those potentially diagnostic of CRC, where the Microscopic Exam section text matches “%carcinoma%”, “%tumor%”, or “%invasi%”, where “%” represents a wildcard. The search is not case-sensitive. The pathology reports matching this search are considered a part of the “plausible set” for CRC identification. The plausible set for dysplasia and HGD/CRC was expanded to include more search terms identifying those diagnoses. See Supplementary sections “*HGD/CRC ascertainment*” and “*Dysplasia ascertainment*” for details. Supplementary Table S1 contains the numbers of reports considered for all patient cohorts and filtering steps across tasks.

LLM prompt development

LLMs require a ‘prompt’ to perform a given task. A ‘prompt’ is defined as the input text given to the model. The model then evaluates the prompt and generates additional text. The prompt we

provide to the model consists of some text that defines the task and the question to be answered. Additionally, the prompt includes the text from the pathology report or section to be evaluated. We developed the prompt using 48 pathology report Microscopic Exam sections where the Specimen section matched colon or rectum terms. These were drawn without consideration for the IBD diagnosis status of the patient. All 48 of these Microscopic Exam sections contained the term ‘carcinoma’. From manual chart review, 16 had invasive colorectal adenocarcinoma, 16 had high-grade dysplasia or adenocarcinoma in the colon or rectum, and 16 had neither. Some of the 48 also had dysplasia that was not high-grade in the colon or rectum. Prompt iteration and model selection occurred mainly in these 48, which were excluded from all future validation sets. Minor changes were made iteratively to the prompts to correct any obvious errors related to the prompt (e.g., missing tubular adenomas before dysplasia prompt included 'any adenoma') using additional development sets for each task, which consisted of 818 reports for CRC, 200 reports for HGD/CRC, and 572 reports for any dysplasia. For details on the evolution of the prompts, see Supplementary Methods section “*Lessons learned from applying LLMs in structured and unstructured pathology report data*”. No a priori performance targets were applied after this iteration. Then for each task, the final validation sets to evaluate the LLMs excluded all reports previously chart reviewed that were considered part of the development sets for a given task.

CRC identification using LLM

For each report in the plausible set of CRC pathology reports, we feed either the Microscopic Exam section (or the full text pathology note) to an LLM which determines if an individual has a pathological diagnosis of invasive adenocarcinoma. The input text is integrated into the prompt as shown below:

*“The text provided is a pathology report, with samples originating from the colon or rectum unless specified otherwise. We are interested in identifying whether invasive adenocarcinoma (stage greater than or equal to 1) is present in *any* colon or rectal sample. Without definite invasion identified, conditions such as 'high-grade dysplasia', 'in-situ [adeno]carcinoma', or 'intramucosal [adeno]carcinoma' are not typically classified as invasive adenocarcinoma. If the sample is classified as having adenocarcinoma without further specification, this typically implies invasive adenocarcinoma. Answer yes or no to the following question, matching the format 'Answer: Yes' or 'Answer: No'. Then, explain your reasoning. Does the pathology report indicate that the patient has an invasive adenocarcinoma in any colon or rectal sample?*
<<<
Pathology report:
{Insert Microscopic Exam text or full-text pathology note}
>>>
Does the pathology report indicate that the patient has an invasive adenocarcinoma in any colon or rectal sample?
Answer:”

The model then responds “Yes” or “No”. This response is recorded as an output “.txt” file with the corresponding ID of the pathology domain entry. Llama.cpp[18] is used for model inference. The

prompt is changed accordingly for the tasks of any dysplasia and HGD/CRC (see Supplementary Methods sections "*HGD/CRC ascertainment*" and "*Dysplasia* ascertainment" for details).

IBD colitis and non-IBD colitis cohorts

To split the cohort between patients with versus without IBD colitis, we use a modified version of a previously validated IBD ascertainment algorithm in VHA data[22]. Here we focused on identifying patients with IBD colitis specifically, who are at risk of colitis-associated dysplasia and colorectal cancer: ulcerative colitis, IBD-unclassified, and Crohn's colitis. Our ascertainment algorithm thus required at least 2 ICD codes matching to ones in the following list: ICD-10 codes K51.x (excluding K51.4x), K50.1x, K50.8x, K52.3, ICD-9 codes 555.1x, 555.2x, 556.x. These codes must be present on at least 2 encounters (dates) with at least one in an outpatient setting.

The non-IBD cohort comprised patients without *any* of the above listed ICD codes. Therefore, patients with only one of the above-listed ICD codes (MVP: n = 13,158; 1.44%) were excluded from both cohorts due to their uncertain history of IBD colitis based on ICD codes.

Model validation

Validation was performed independently in IBD and non-IBD populations using the same models and prompts. For each validation set, either N=100 (CDW) or N=150 (MVP) randomly selected putative positive cases and the same number of putative negative cases were selected for review. Putative positive (negative) cases were defined as cases where Llama-3 responded "Yes" ("No"). Validation was performed by a blinded single reviewer (BJ). Difficult cases were discussed at meetings with gastroenterologists Drs. S Shah and S Gupta. Validation was done at the level of the pathology report, consistent with the LLM prompt asking if the given features are present in any colon or rectal sample. Validation was performed independently for each of the three tasks, even if notes overlapped by chance in validation sets across tasks.

We performed validation only in the "plausible set" of notes that passed our search term filters (Supplementary Table S1) and recorded run-times with CPUs (Supplementary Table S2). Considering the very low expected prevalence (potentially zero) outside of these filters, this approach provides a more informative assessment of the LLMs' performance, as we are considerably more likely to include some false negative cases in our validation (see Supplementary methods for additional details on model validation). For testing generalizability, we also evaluated performance using full pathology reports as LLM prompt input, as described above. Note, full-text pathology reports were considered in validation analyses only and were not used in prompt development.

Performance metrics

We provide numbers of true positive, true negative, false positive and false negatives for all tasks. Because we use stratified sampling, selecting N model positive cases and N model negative cases for validation, the prevalence of model positive cases in the plausible set is required to

calculate downstream metrics. We provide an estimate of the prevalence of cases in the reports from the plausible set. We calculate the sensitivity (recall), specificity, F1-score, and Matthew's Correlation Coefficient (MCC) which are functions of the confusion matrix and the prevalence.

Where w = prevalence of model positives in the plausible set, we calculated

$$Sensitivity = PPV * w/(PPV * w + (1 - NPV) * (1 - w))$$

$$Specificity = NPV * (1 - w)/(NPV * (1 - w) + (1 - PPV) * w)$$

$$F1 = 2 * PPV * Sensitivity/(PPV + Sensitivity)$$

$$MCC = \sqrt{PPV * Sensitivity * Specificity * NPV} - \sqrt{(1 - PPV) * (1 - Sensitivity) * (1 - Specificity) * (1 - NPV)}$$

This approach helps minimize the number of cases needed for validation, especially when prevalence is imbalanced, and has been implemented previously[23].

Results

We applied LLMs to extract pathologic diagnoses from text in the VA Pathology Domain (Fig. 1) and free-text pathology reports. We tested two LLMs (Gemma-2 and Llama-3) for each classification task. After prompt development, we validated our methods by comparing model predictions to chart review of randomly chosen sets of reports for each task (dysplasia, HGD/CRC and CRC) in each patient cohort (IBD and non-IBD) and dataset (MVP and CDW).

Large language models extract pathologic diagnoses with high accuracy in patients with IBD

In model validation using strictly distinct reports from those used for prompt development (see Materials and Methods), all tasks achieved excellent performance (ranges for 3 tasks: PPV = 0.928 - 0.987, NPV = 0.961-1.00) using LLM Gemma-2 (Table 1). The F1 score, which combines precision and recall, was >91% in all cases. We found slightly lower performance when using LLM Llama-3 (Supplementary Table S3). As expected, smaller LLMs with fewer parameters were less accurate in the 3 tasks (Supplementary Table S4).

Validation of large language model approach in non-IBD colorectal dysplasia and cancer

We then applied the same approach with no changes to prompts to records from patients without IBD (no IBD colitis ICD code found in patient clinical history) and again achieved highly accurate ascertainment in all 3 tasks, as shown in Table 2. Specifically, we found that the F1 score for identifying dysplasia in patients without IBD was above 99% using both Gemma-2 and Llama-3 (Table 2, Supplementary Table S3). F1 scores were slightly lower but still excellent (>98%) for HGD/CRC using Gemma-2.

Accuracy of applying LLM methods to full text pathology report

To evaluate the generalizability of our model to environments that do not contain semi-structured resources such as the VA Pathology Domain, we applied our LLM approach to the full pathology

report to evaluate performance and found excellent measures using Gemma-2 (Table 3). While both Gemma-2 and Llama-3 were trained with context lengths up to 8,192 tokens, and the full notes never exceeded these thresholds, performance decreases slightly when using Llama-3 (Supplementary Table S3). We also found similar performance results to Gemma-2 alone when requiring either or both models to answer 'Yes' for a report to be deemed a positive case (Supplementary Table S5).

LLM approach for general population in the context of previous NLP approaches

While previous NLP approaches show excellent performance in identifying common features like adenoma[9,10,13], few have maintained excellent performance with thorough testing of their approaches to identify rarer advanced features such as HGD, carcinoma in situ, and invasive adenocarcinoma. Additionally, few have tested approaches in differing contexts, such as differing geographical locations, practice types (e.g., academic vs. private practice), and compensation structures (e.g., salary vs. fee-for-service)[15]. When attempted across 4 practice sites, Carrell et al. report an F1-score of 95% for identifying adenoma and highlight the considerable time-consuming challenges they encountered in adapting the NLP system[15]. The most comparable analysis in our study to previously published algorithms was the task of identifying any dysplasia in the non-IBD colitis cohort, where Gemma-2 had an average F1-score of 99.4% and Llama-3 had an average F1-score of 99.6%. Published analyses in similar cohorts have similarly high F1-scores, such as Bae et al., who report an F1 score of 99% for identifying the presence of "conventional adenoma"[13], and Nayor et al., who report a perfect F1 score of 100% for identifying "adenoma"[10]. Supplementary Table S7 shows a comparison of performance across comparable tasks.

Discussion

We have shown that LLMs are powerful, generalizable tools for accurately extracting important information from clinical semi-structured and unstructured text. With validated performance also shown in the Million Veteran Program (MVP), we expect this approach will enable large-scale health research studies that can incorporate patient genomics in disease risk assessment and prediction. Another strength of this work is that the methods are relatively simple. While not explicitly tested, we expect our findings to adapt relatively easily to other pathological diagnoses, healthcare systems, patient populations, and time periods. No aspect of the prompt or models used were specific to the VA or our cohorts, and no additional model fine-tuning was performed. The barriers to implementation are minimal; any researcher can clone the llama.cpp[18] GitHub, add their desired prompt, compile, and begin development. Our forked repository is available on GitHub for the community. Due to the ease of model implementation, with results as accurate as more complicated rule-based approaches, we suggest an LLM approach for many free-text classification tasks in biomedical research going forward.

While LLMs remain computationally expensive, the size and associated compute cost of proficient models has reduced drastically, with the best small (9 billion parameter), open-weight model of today (Gemma-2-9b-it[19], released June 27, 2024) generally performing better than the largest

proprietary models from a year prior (GPT-4-0613[24], released June 13, 2023)[25]. If such improvements in efficiency continue, boosted by potential advancements in the underlying transformer architecture[26], LLMs will become more attractive in domains where the current computational expense makes their use unfeasible. Even without further improvements in the models themselves, the increasing availability of GPU and the throughput of new chip architectures[27–29] may make current models a viable alternative to data structuring at scale.

Our work has some limitations. First, we had a single reviewer validating the pathology report sections, although difficult cases with any questions were reviewed with two experienced gastroenterologists (SG, SCS). Without access to Graphical Processing Units (GPUs), we could not feasibly test larger models which may overcome some of the shortcomings seen in smaller models; this addition can be expected to increase performance above what we find herein. Finally, we could not rule out overlap between MVP and CDW reports, though our results in either cohort considered alone are sufficient validation compared to previously published work.

Ongoing work includes adapting our approach to detect stage and location of cancers, identifying features of dysplasia (size, shape, type, location, inflammation level, etc.), and determining the impact of the quality of colonoscopy exams. Some tasks, such as identifying IBD sub-types and dates of diagnosis, may require larger models that are more capable of handling longer input text. Nonetheless, the general framework lends itself to many applications beyond the use cases analyzed here, including the potential for real-time data integration in models used to aid in shared decision-making (so-called medical digital twins)[30].

Accurate clinical data is essential for understanding trends in patient disease risk and for predictive models to be clinically useful. In an era of increasing opportunities for personalized medicine, we show that large language models offer a very useful tool for quickly and accurately obtaining relevant patient data to potentially inform medical decisions in real time.

Code availability

Llama.cpp fork will be made available at https://github.com/bdj34/llama.cpp_data_extraction, which includes the main inference code from llama.cpp. Our specific implementation can be found at https://github.com/bdj34/llama.cpp_data_extraction/tree/brian-features/examples/data-extraction which includes all prompts and custom parameters required to reproduce our work. Large language models, in gguf format, used in this work are stored at https://huggingface.co/briandj97/models_used.

Data availability

A CSV (all_results.csv) with results from all validated runs, including additional details such as prevalence of model positives, exact numbers validated, and the full confusion matrix, will be available in the supplementary information. This CSV is the source for all three main text tables and Supplementary Tables S3-S5. Raw data access is reserved for VA investigators with appropriate research approvals.

References

1. Choi CHR, Rutter MD, Askari A, et al. Forty-Year Analysis of Colonoscopic Surveillance Program for Neoplasia in Ulcerative Colitis: An Updated Overview. *Am J Gastroenterol*. 2015;110(7):1022-1034. doi:10.1038/ajg.2015.65

2. Shah SC, Itzkowitz SH. Colorectal Cancer in Inflammatory Bowel Disease: Mechanisms and Management. *Gastroenterology*. 2022;162(3):715-730.e3. doi:10.1053/j.gastro.2021.10.035

3. Gupta S, Lieberman D, Anderson JC, et al. Recommendations for Follow-Up After Colonoscopy and Polypectomy: A Consensus Update by the US Multi-Society Task Force on Colorectal Cancer. *Gastroenterology*. 2020;158(4):1131-1153.e5. doi:10.1053/j.gastro.2019.10.026

4. Murthy SK, Feuerstein JD, Nguyen GC, Velayos FS. AGA Clinical Practice Update on Endoscopic Surveillance and Management of Colorectal Dysplasia in Inflammatory Bowel Diseases: Expert Review. *Gastroenterology*. 2021;161(3):1043-1051.e4. doi:10.1053/j.gastro.2021.05.063

5. Rubin DT, Ananthakrishnan AN, Siegel CA, Sauer BG, Long MD. ACG Clinical Guideline: Ulcerative Colitis in Adults. *Am J Gastroenterol*. 2019;114(3):384-413. doi:10.14309/ajg.0000000000000152

6. Benson R, Winterton C, Winn M, et al. Leveraging Natural Language Processing to Extract Features of Colorectal Polyps From Pathology Reports for Epidemiologic Study. *JCO Clin Cancer Inform*. 2023;7:e2200131. doi:10.1200/CCI.22.00131

7. Fevrier HB, Liu L, Herrinton LJ, Li D. A Transparent and Adaptable Method to Extract Colonoscopy and Pathology Data Using Natural Language Processing. *J Med Syst*. 2020;44(9):151. doi:10.1007/s10916-020-01604-8

8. Gupta S, Earles A, Bustamante R, et al. Adenoma Detection Rate and Clinical Characteristics Influence Advanced Neoplasia Risk After Colorectal Polypectomy. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. 2023;21(7):1924-1936.e9. doi:10.1016/j.cgh.2022.10.003

9. Harkema H, Chapman WW, Saul M, Dellon ES, Schoen RE, Mehrotra A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. *J Am Med Inform Assoc*. 2011;18(Supplement_1):i150-i156. doi:10.1136/amiajnl-2011-000431

10. Nayor J, Borges LF, Goryachev S, Gainer VS, Saltzman JR. Natural Language Processing Accurately Calculates Adenoma and Sessile Serrated Polyp Detection Rates. *Dig Dis Sci*. 2018;63(7):1794-1800. doi:10.1007/s10620-018-5078-4

11. Imler TD, Morea J, Kahi C, Imperiale TF. Natural Language Processing Accurately Categorizes Findings From Colonoscopy and Pathology Reports. *Clin Gastroenterol Hepatol*. 2013;11(6):689-694. doi:10.1016/j.cgh.2012.11.035

12. Raju GS, Lum PJ, Slack RS, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. *Gastrointest Endosc*. 2015;82(3):512-519. doi:10.1016/j.gie.2015.01.049

13. Bae JH, Han HW, Yang SY, et al. Natural Language Processing for Assessing Quality Indicators in Free-Text Colonoscopy and Pathology Reports: Development and Usability Study. *JMIR Med Inform*. 2022;10(4):e35257. doi:10.2196/35257

14. Syed S, Angel A, Syeda H, et al. The h-ANN Model: Comprehensive Colonoscopy Concept Compilation using Combined Contextual Embeddings: In: *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications; 2022:189-200. doi:10.5220/0010903300003123

15. Carrell DS, Schoen RE, Leffler DA, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc*. 2017;24(5):986-991. doi:10.1093/jamia/ocx039

16. Acharya V, Kumaresan V, England J, Mehta S, Sussman D, Deshpande A. Use of Large Language Models to Identify Surveillance Colonoscopy Intervals—A Feasibility Study. *Gastroenterology*. Published online October 2024:S0016508524055380. doi:10.1053/j.gastro.2024.09.032

17. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214-223. doi:10.1016/j.jclinepi.2015.09.016

18. Gerganov G. llama.cpp. Published online 2024. https://github.com/ggerganov/llama.cpp

19. Gemma Team, Thomas Mesnard, Cassidy Hardin, et al. Gemma. doi:10.34740/KAGGLE/M/3301

20. Wu Y, Sun Z, Yuan H, Ji K, Yang Y, Gu Q. Self-Play Preference Optimization for Language Model Alignment. Published online 2024. doi:10.48550/ARXIV.2405.00675

21. AI@Meta. Llama 3 Model Card. Published online 2024. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

22. Hou JK, Kramer JR, Richardson P, Mei M, El-Serag HB. The Incidence and Prevalence of Inflammatory Bowel Disease Among U.S. Veterans: A National Cohort Study. *Inflamm Bowel Dis*. 2013;19(5):1059-1064. doi:10.1097/MIB.0b013e31828028ca

23. Liu L, Bustamante R, Earles A, Demb J, Messer K, Gupta S. A strategy for validation of variables derived from large-scale electronic health record data. *J Biomed Inform*. 2021;121:103879. doi:10.1016/j.jbi.2021.103879

24. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. Published online 2023. doi:10.48550/ARXIV.2303.08774

25. Chiang WL, Zheng L, Sheng Y, et al. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. Published online 2024. doi:10.48550/ARXIV.2403.04132

26. Ye T, Dong L, Xia Y, et al. Differential Transformer. Published online October 7, 2024. Accessed October 8, 2024. http://arxiv.org/abs/2410.05258

27. Abts D, Kimmell G, Ling A, et al. A software-defined tensor streaming multiprocessor for large-scale machine learning. In: *Proceedings of the 49th Annual International Symposium on Computer Architecture*. ACM; 2022:567-580. doi:10.1145/3470496.3527405

28. Prabhakar R, Sivaramakrishnan R, Gandhi D, et al. SambaNova SN40L: Scaling the AI Memory Wall with Dataflow and Composition of Experts. Published online 2024. doi:10.48550/ARXIV.2405.07518

29. Lie S. Cerebras Architecture Deep Dive: First Look Inside the Hardware/Software Co-Design for Deep Learning. *IEEE Micro*. 2023;43(3):18-30. doi:10.1109/MM.2023.3256384

30. Johnson B, Curtius K. Digital twins are integral to personalizing medicine and improving public health. *Nat Rev Gastroenterol Hepatol*. Published online 2024. doi:https://doi.org/10.1038/s41575-024-00992-3

31. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. Published online 2023. doi:10.48550/ARXIV.2310.06825

32. Zhu B, Frick E, Wu T, Zhu H, Jiao J. Starling-7b: Improving LLM helpfulness & harmlessness with rlaif. Published online November, z2023. https://huggingface.co/berkeley-nest/Starling-LM-7B-alpha

33. teknium. OpenHermes-2.5-Mistral-7B. https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B

34. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of Experts. Published online 2024. doi:10.48550/ARXIV.2401.04088

35. Hendrycks D, Burns C, Basart S, et al. Measuring Massive Multitask Language Understanding. Published online 2020. doi:10.48550/ARXIV.2009.03300

36. AI@Meta. Llama 3.2 Model Card. Published online September 25, 2024. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md

37. Frantar E, Ashkboos S, Hoefler T, Alistarh D. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. Published online 2022. doi:10.48550/ARXIV.2210.17323

Acknowledgements

This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration, and was supported by MVP000 as well as Merit Review Award I01 BX005958 from the United States (U.S.) Department of Veterans Affairs Biomedical Laboratory Research and Development Service. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government. This work was supported by AGA Research Foundation (AGA Research Scholar Award AGA2022-13-05), NIH grants (R01 CA270235, P30 CA023100), and National Library of Medicine Training Grant (NIH grant T15LM011271). The study was supported in part by the NIDDK-funded San Diego

Digestive Diseases Research Center (P30 DK120515). Detailed MVP Core team acknowledgments are included in the Supplementary Information.

Figure

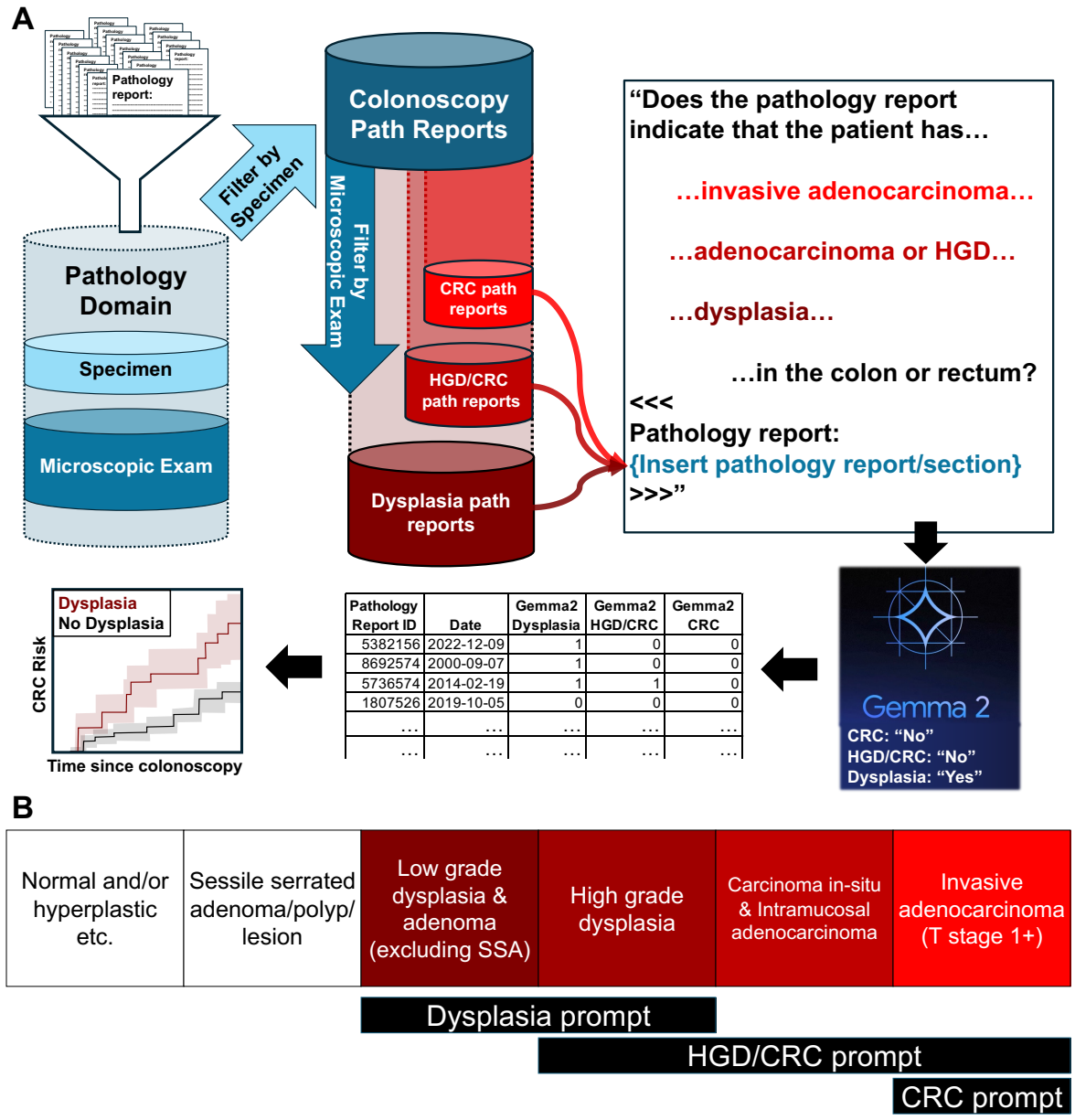

Figure 1. Workflow for extracting diagnoses from electronic health data. A: The Pathology Domain sections were used for simple term filtering designed to reduce the total number of reports fed to the LLM, while still capturing all pathology reports with possible diagnoses. Then the Microscopic Exam section of the report (or the full pathology report) is integrated into a prompt fed to the LLMs. A partial prompt is shown in the figure for illustrative purposes, and the full prompts are available in Materials and Methods and Supplementary Information. The LLM

answers (“Yes” or “No”) are converted to structured data, which can be used for downstream applications, such as estimating a stratified risk of CRC based on colonoscopy findings. **B:** Diagnoses to be obtained by 3 prompts utilized in this study. By applying each of the three prompts, the most advanced diagnosis can be stratified into each of the following four buckets: no dysplasia or adenocarcinoma, low-grade dysplasia, high-grade dysplasia OR intramucosal adenocarcinoma OR carcinoma in-situ, and invasive adenocarcinoma. HGD/CRC = High-grade dysplasia or adenocarcinoma; CRC = Colorectal cancer (invasive colorectal adenocarcinoma); Path = Pathology.

Tables

Task	Source	PPV (LB - UB)	NPV (LB - UB)	Recall (Sensitivity)	Specificity	F1	MCC
CRC	MVP	0.962 (0.92-0.99)	0.993 (0.96-1.00)	0.980	0.987	0.971	0.961
CRC	CDW	0.928 (0.86-0.97)	0.961 (0.90-0.99)	0.896	0.974	0.912	0.879
HGD/CRC	MVP	0.961 (0.92-0.99)	0.993 (0.96-1.00)	0.968	0.992	0.964	0.957
HGD/CRC	CDW	0.960 (0.90-0.99)	1.000 (0.96-1.00)	1.000	0.989	0.980	0.975
Dysplasia	MVP	0.987 (0.95-1.00)	0.987 (0.95-1.00)	0.956	0.996	0.971	0.963
Dysplasia	CDW	0.980 (0.93-1.00)	1.000 (0.96-1.00)	1.000	0.991	0.990	0.985

Table 1: Validation results for IBD patients in MVP and CDW using Gemma-2. Performance in the validation set for IBD cohort. 95% confidence intervals for PPV and NPV are calculated using a binomial distribution. Green shading progression with lower value = 0.9 and upper value = 1. CRC = Colorectal cancer (invasive colorectal adenocarcinoma); HGD/CRC = High-grade dysplasia or adenocarcinoma; IBD = Inflammatory bowel disease; PPV = Positive predictive value; NPV = Negative predictive value; LB = Lower bound; UB = Upper bound; MCC = Matthew’s correlation coefficient.

Task	Source	PPV (LB - UB)	NPV (LB - UB)	Recall (Sensitivity)	Specificity	F1	MCC
CRC	MVP	0.968 (0.93-0.99)	0.952 (0.90-0.98)	0.957	0.964	0.962	0.920
CRC	CDW	0.941 (0.88-0.98)	0.959 (0.90-0.99)	0.962	0.937	0.952	0.900
HGD/CRC	MVP	0.981 (0.95-1.00)	0.993 (0.96-1.00)	0.984	0.992	0.982	0.975
HGD/CRC	CDW	0.990 (0.95-1.00)	0.990 (0.94-1.00)	0.985	0.993	0.988	0.979
Dysplasia	MVP	0.987 (0.95-1.00)	0.993 (0.96-1.00)	0.999	0.899	0.993	0.938
Dysplasia	CDW	0.990 (0.95-1.00)	0.990 (0.94-1.00)	0.998	0.938	0.994	0.958

Table 2: Validation results for non-IBD colitis patients in MVP and CDW using Gemma-2. Performance in the validation set for non-IBD cohort. 95% confidence intervals for PPV and NPV are calculated using a binomial distribution. Green shading done in Excel with lower value = 0.9 and upper value = 1. CRC = Colorectal cancer (invasive colorectal adenocarcinoma); HGD/CRC

= High grade dysplasia or adenocarcinoma; IBD = Inflammatory bowel disease; PPV = Positive predictive value; NPV = Negative predictive value; LB = Lower bound; UB = Upper bound; MCC = Matthew's correlation coefficient.

Task	Input type	PPV (LB - UB)	NPV (LB - UB)	Recall (Sensitivity)	Specificity	F1	MCC
CRC	Microscopic exam	0.962 (0.92-0.99)	0.993 (0.96-1.00)	0.980	0.987	0.971	0.961
CRC	Full pathology report	0.903 (0.84-0.95)	1.000 (0.96-1.00)	1.000	0.968	0.949	0.935
HGD/CRC	Microscopic exam	0.961 (0.92-0.99)	0.993 (0.96-1.00)	0.968	0.992	0.964	0.957
HGD/CRC	Full pathology report	0.917 (0.85-0.96)	0.991 (0.95-1.00)	0.958	0.983	0.937	0.924
Dysplasia	Microscopic exam	0.987 (0.95-1.00)	0.987 (0.95-1.00)	0.956	0.996	0.971	0.963
Dysplasia	Full pathology report	0.992 (0.95-1.00)	0.992 (0.95-1.00)	0.973	0.997	0.982	0.977

Table 3: Validation results using full pathology report in IBD population in MVP. Comparison of Microscopic Exam section (these rows are repeated from Table 1) and full pathology report as input to LLM. Full pathology reports evaluated by LLMs in all cases where the Pathology Domain entry had a matching full note. There were many instances where the full note was not available, for reasons we do not know. As such, the number validated for the full pathology report is less than the number validated for "Microscopic exam". All analyses still had > 100 model positive and model negative cases validated. See Supplementary Table S6 for details on pathology report numbers. Green shading done in Excel with lower value = 0.9 and upper value = 1. CRC = Colorectal cancer (invasive colorectal adenocarcinoma); HGD/CRC = High grade dysplasia or adenocarcinoma; PPV = Positive predictive value; NPV = Negative predictive value; LB = Lower bound; UB = Upper bound; MCC = Matthew's correlation coefficient.