medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

LLM-based data extraction for a large cancer registry, the Ontario Hereditary Cancer Research Network

Andres Felipe Melani De La Hoz, MSc[1,2,‡]; Jochen Weile, PhD[1,‡]; Pratham Hemlani, BSc[1,‡]; Elif Tuzlali, BSc[1]; Sarah Ridd, BASc[1,3]; Brandon Chan, BASc[1]; Lauren K. Hughes, MS[1]; Kathy Chun, PhD[11,12]; Harriet Feilotter, PhD[1,7,8]; Daria Grafodatskaya, PhD[17,18]; Jordan Lerner-Ellis, PhD[12,20]; Laila Schenkel, PhD[9,10]; Amanda Smith, PhD[14,15]; Andrea Vaags, PhD[12,13]; Hong Wang, PhD[19]; Raymond H. Kim, MD, PhD[1,3]; Lincoln Stein, MD, PhD[1,16]; Benjamin Haibe-Kains, PhD[2,3,4,5]; Melanie Courtot, PhD[1,2,6]

[‡]**These authors contributed equally to this work.**

Affiliations

[1] Ontario Institute for Cancer Research, Toronto, Ontario, Canada
[2] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[3] Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
[4] Structural Genomics Consortium, Toronto, Ontario, Canada
[5] Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
[6] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[7] Genome Diagnostics, Laboratory Medicine Program, University Health Network, Toronto, Ontario, Canada
[8] Department of Pathology and Molecular Medicine, Queen's University, Kingston, Ontario, Canada
[9] Department of Pathology and Laboratory Medicine, Western University, London, Ontario, Canada
[10] Molecular Diagnostics Division, London Health Sciences Centre, London, Ontario, Canada
[11] Hospital for Sick Children, Toronto, Ontario, Canada
[12] Department of Laboratory Medicine & Pathobiology, University of Toronto, Toronto, Ontario, Canada
[13] Department of Laboratory Medicine and Genetics, Trillium Health Partners, Mississauga, Ontario, Canada
[14] Children's Hospital of Eastern Ontario (CHEO), Ottawa, Ontario, Canada
[15] Department of Genetics, University of Ottawa, Ottawa, Ontario, Canada
[16] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[17] Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada
[18] Hamilton Regional Laboratory Medicine Program, Hamilton Health Sciences, Hamilton, Ontario, Canada
[19] Molecular Genetics and Cytogenetics Laboratories, Genetics Program, North York General, Toronto, Ontario, Canada
[20] Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Sinai Health, Toronto, Ontario, Canada

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Importance: Manual data extraction from genomic lab reports for on-line registries and databases is time-consuming for human resources such as clinical research coordinators. Automated tools, especially LLMs, can address these issues. Efficient and accurate data processing is crucial for building a reliable database.

Objective: To streamline the data extraction and curation process for genetic testing lab reports using an LLM-based approach.

Design: Nine sample molecular lab reports were selected for manual data extraction by two expert curators. The process was timed, and the results served as gold-standard for validating automated extraction. Eighteen fields from the OHRCN's data model were selected as extraction targets.

Setting: The study was conducted within OHCRN, which unifies research, genomic, and clinical patient data from clinics and laboratories across Ontario, Canada.

Participants: Nine laboratories agreed to share sample molecular lab reports and two clinical research coordinators affiliated with OHCRN participated as data curators.

Exposure: LLM-based Extraction of Information (LEI), an automated data extraction pipeline, was developed using regular expressions, Trie search, and LLMs to extract data from molecular lab reports and structure it for inclusion into OHCRN's database.

Main Outcomes and Measures: LEI was evaluated by measuring the F1-score on the extraction task of 18 entity types. These measures were compared against 15 extraction tools in the biomedical domain. Extraction time was also measured and compared against manual extraction times.

Results: LEI demonstrated quality on par with and surpassing other existing LLM-based extraction methods. Reference tools showed F1-scores around 70%, while LEI achieved an average score of 87.4%. LEI reduced extraction time by approximately 2-fold, with an average

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

time of 7.59 minutes per report including results review by curators, compared to 14.88 minutes per report for manual extraction.

Conclusions and Relevance: LEI facilitates standardized, accurate, and efficient healthcare data extraction from unstructured texts, significantly improving the current OHCRN workflow. By automating the extraction process, LEI allows expert curators to focus on validating results rather than performing manual data entry. LEI's simple interface enables researchers to easily guide extraction tasks and supports adaptability across diverse biomedical scenarios. Future improvements in accuracy may be achieved through fine-tuning techniques and ongoing advancements in LLM technologies.

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Large-scale clinical databases are essential for healthcare operations, research, and evidence-based decision-making[1,2]. Examples of these databases include COSMIC[3], with a pool of data from over 29,000 scientific publications; the National Cancer Database[4], with over 34 million records of patients in the United States; or TDBRAIN[5], a data collection from more than 1,200 psychiatric patients. The exponential growth and format heterogeneity of clinical data demand innovative approaches to efficiently manage high-volume datasets.

The Ontario Hereditary Cancer Research Network (OHCRN)[6] is an example of such a clinical database. OHCRN is a province-wide program for individuals with hereditary cancer syndromes in Ontario. Hereditary cancer accounts for approximately 10% of all cancer cases[7] and presents a unique opportunity for data sharing and research to improve early detection, biological understanding, and treatment. By centralizing clinical and germline genomic data and making them accessible to researchers, OHCRN aims to facilitate cancer research and connect participants with researchers and clinical trials. A significant challenge lies in the manual curation of germline genetic testing reports, which are provided as text documents in a variety of formats conforming to individual laboratory practices. This manual extraction of the relevant data points and their transformation into structured data is time-consuming, expensive and prone to human errors (e.g., typographical errors). As a representative example of multiple large databases used in biomedical research and clinical operations, OHCRN would benefit from a computational tool for extracting data from unstructured sources, which could also be applied to similar initiatives.

Traditional automated extraction tools have design and setup limitations[8]. For example, clinical data extraction for the MSK-CHORD[9] project employed multiple Natural Language Processing transformer models that required individual training on different curated data sets, plus additional

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

rule-based models to process more structured data. Recently, Large Language Models (LLMs) have made great strides in improving both the power and ease of use for data extraction applications. In the biomedical domain, BioBERT[10], MT-clinical BERT[11], and Med-PaLM[12] demonstrate capabilities in entity extraction and normalization. Modern foundation models, such as GPT[13], LLAMA[14] and Claude[15] offer unstructured data extraction via in-context learning. For instance, GL-NER[16] employs a few-shot approach and proposes a Universal Prompt template to improve the model's Named Entity Recognition (NER) tasks. By contrast, GPT-NER[17] re-frames the NER task from sequence labelling to text generation, making it easier for LLMs to recognize entities while still following the general paradigm of in-context learning.

Furthermore, genetic testing reports present unique challenges for data extraction such as intricate tables, genetic variants nomenclature and highly specialized terminology, and the presence of established or novel variants. Format also varies across laboratories, requiring deep contextual understanding to correctly interpret findings.

To address these limitations, we propose LEI (LLM-based Extraction of Information), an integrated framework that leverages computer vision and LLMs to extract healthcare data from non-standardized, unstructured text sources, and provide machine-readable output suitable for database curation. The primary purpose of LEI is to streamline the data extraction and curation process of molecular lab reports, empowering data curators to process reports at a much accelerated pace and comparable accuracy.

Methods

As part of the OHCRN initiative, germline genetic testing results for at-risk individuals undergoing testing in one of nine provincially funded laboratories will be collected from 20 genetics clinics

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

across Ontario. Data will be centralized within a central repository (Supplementary Figure 1). The lab reports must be processed manually by OHCRN's data curators, who locate and extract the information to be entered into the OHCRN data repository in accordance with its data model (Supplementary Figures 2A, 2B).

The LEI data extraction framework

LEI consists of two modules: a computer vision module for converting physical documents into digital text, and a data extraction module, which employs multiple strategies, such as regular expression searches and LLM-based entity recognition to extract genetic testing data. LEI's output is represented as JSON-formatted key-value pairs (Figure 1). To address the unique characteristics of germline genetic test reports, LLM prompts were designed with the support of an expert curator and LEI employs multiple standard nomenclatures for the definition of regular expressions and the trie structure.

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

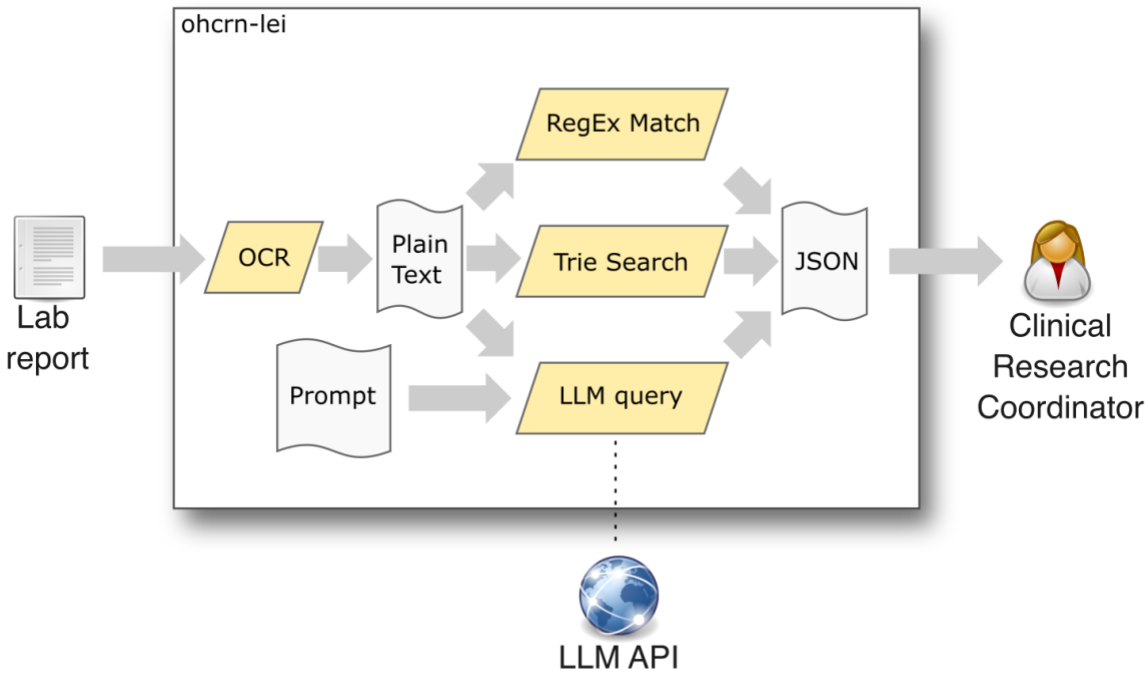


Figure 1. Overview of LEI. *Lab reports are digitized using the OCR tool, and the text is then passed to the LLM, RegEx and Trie Search. The LLM also takes as input the prompts defining the entities to extract and a system message that guides the model'*

The computer vision module

The computer vision module accepts scanned or photographed documents in PDF or JPG format and converts them to digital text. When designing this module, multiple optical character recognition (OCR) libraries were considered, namely Tesseract[18], EasyOCR[19] and Amazon Textract[20]. A comparative analysis by Vedhaviyassh and colleagues found that EasyOCR outperforms Tesseract OCR accuracy by 5% due to its deep learning capabilities for object recognition[21]. Textract was discarded for not being open-sourced and the EasyOCR library was chosen instead. The LEI computer vision module was implemented in Python.

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The data extraction module

The data extraction module takes the digital text output from the computer vision model and extracts relevant biomedical data, such as clinical report types, dates, and descriptors of identified genomic variants. Extraction was performed using a combination of pattern matching with regular expressions, Trie searches, and zero-shot LLM prompting.

Pre-defined regular expressions for a number of biological entity identifiers were obtained from the following sources: patterns for variant identifiers in HGVS format[22] from the hgvs-regexp repository[23] for the DNA level and mavehgvs[24] for the protein level nomenclature. Accession patterns for variant accessions and cross references from OMIM[25], ClinVar[26], dbSNP[27], COSMIC[3], ClinGen[28], and UniProt[29] were obtained from their respective authoritative sources. Finally, chromosome identifier patterns were created manually.

To recognize official gene symbols and aliases as defined by the HUGO human gene nomenclature committee (HGNC)[30], a Trie Search algorithm[31] was implemented. The HGNC gene definitions file was downloaded and all symbols and aliases were compiled into a prefix tree structure. This structure was then used as a finite state machine to step through text documents, identifying all matching substrings.

For the LLM component, OpenAI GPT-4o[32] was used, as it has previously been demonstrated to be highly accurate and flexible, enhancing its ability to comprehend and generate cross-modal content[33]. Access to GPT-4o was handled via the Chat Completions API[34], which provides two different communication roles: a system message role to set the expected behavior from the model and a user message role to provide requests to which the assistant will respond[35]. The latter was used to input text from the molecular reports, from where the data was to be extracted.

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

To ensure consistency in the LLM's response when faced with the same input text, a softmax temperature parameter of zero was used. This value favors deterministic and highly probable outputs[36].

A prompt-engineering paradigm described by Zamfirescu-Pereira et al.[37] was used, which describes three relevant strategies:

1. desired interactions: examples for fields that had specific formats or needed additional context to be found by the LLM, such as identifiers;
2. code-like or template-based prompts: prompts using Markdown syntax and
3. repeated command prompts: prompts in which repeated commands were inserted for the cases when the resulting data from the extraction pipeline was missing information or when the LLM ignored commands initially present at just one point in the input prompt.

A Python application for LEI was provided to run the regular expression matching, connect to the OpenAI API, process the prompt and input text, retrieve the result and merge the final JSON output. Prompts for common extraction tasks were packaged with the application, while new custom prompts were provided as plain text files.

Evaluation method

Nine PDF test files containing scanned de-identified molecular lab reports were used, each from a different clinical laboratory in Ontario, covering a variety of gene panels. The reports varied in length from 2 to 8 pages. Each had a different structure, depending on the issuing institution. To differentiate between the performance of the OCR component and the data extraction component, manual plain-text conversions of the documents that could be used to bypass the OCR were also

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

created. To create a gold standard extraction benchmark, manual extractions were performed multiple times on all nine documents by two expert curators and the results reconciled.

Focus was placed on the evaluation of three different extraction tasks: “Report”, “Molecular Test”, and “Variant”, each of which extracted a different set of entities (Supplementary Table 1). A Python script was used to normalize and compare the JSON output against the gold standard and calculate the number of true positive, false positive and false negative entries.

The expected time savings afforded by LEI was quantified by measuring the time needed to perform the Report, Molecular Test and Variant extraction tasks across the nine molecular reports by LEI, as well as by a human curator performing the task manually.

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Results

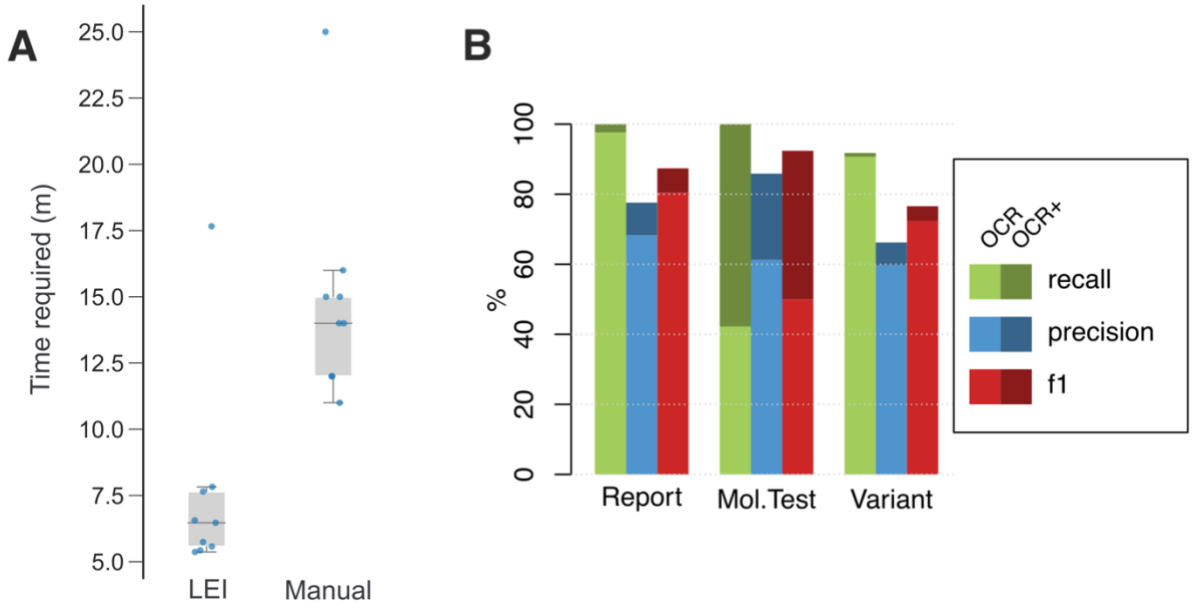


Figure 2. Extraction results. (A) *Time taken to extract the information from all reports, using LEI vs manual extraction. LEI times include the time required to review results by expert curator* ***(B)*** *Precision, recall and F1-score for each extraction task with and without OCR errors. Light colors show the first scenario, which included OCR errors produced by the quality of scanned documents. Dark colors show the second scenario, in which the OCR errors were manually corrected to ensure input text quality*

LEI was capable of performing the extraction ~2 times faster than the manual process. The mean extraction time by LEI, including the time needed to review results, was 7.59 minutes, while the mean manual extraction time by the human curator was 14.88 minutes (Figure 2A).

Extraction accuracy

To evaluate the reliability of the results, the automatic extractions were compared against the manually curated gold standard, determining true positives, false positives, and false negatives for each entity type. True positives represent data that is correctly extracted and accurately mapped to the appropriate OHCRN data model field. Conversely, false positives are extracted values that are incorrectly mapped or are not present in the original text, akin to "hallucinations." False negatives are missed entities that should have been extracted, effectively representing

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

"missing values." We then calculated precision, recall, and F1-scores. Precision is the ratio of true positives to all extracted positives. Recall measures the ratio of true positives to all actual entities meant for extraction. Finally, the F1-score provides a harmonic mean of precision and recall.

Errors can originate from the OCR process or the information extraction process; consequently performance metrics were measured in two different scenarios. The first scenario considers both the OCR and the extraction processes together. In the second scenario, OCR issues were manually corrected to ensure accurate text input for extraction. In this scenario, the Report extraction task successfully achieved the desired score range, with an F1 of 87.4%. The Molecular Test task achieved an F1-score of 92.5, with 100% recall. Notably, the use of OCR had the biggest impact here. Lastly, the Variant task showed an F1-score of 76.6%, explained by a high recall of 90.7% and an acceptable precision of 66.2% (Figure 2B).

Diving deeper into the performance metrics, the “Ordering Clinic” for the “Report” task presented the largest hurdle. The LLM used the testing lab name instead since this value was not included in most reports, generating false positives. For “Molecular Test”, false positives stemmed from the Trie search identifying words identical to real gene names. Besides, OCR issues caused many gene symbols to be missed, particularly in a single report with a continuous list of over 800 gene symbols. The “Sample type” recall suffered as most values are implicit in the reports and require further genetics expertise to deduce them. Finally, the “Variant” task precision was affected by false positives in the “variation code” and “coding HGVS” entity types. The former was caused by persistent hallucinations of the LLM, while the latter was caused by regular expressions incorrectly extracting variants mentioned in the report that are not the targeted data, i.e., the variant identified in the patient (Supplementary Figure 3).

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

LEI is an adaptable framework that does not require set-up, pre-training or curated data, which can also be extended to new tasks described in natural language. These features render LEI accessible to clinical researchers with minimal computational experience. With an overall precision exceeding 80% and over 2-fold time savings, LEI enables genetic testing data extraction from unstructured texts in a standardized, accurate and fast manner. LEI replaces the manual extraction subprocess (Supplementary Figure 2C) and assists curators in the validation step. This enables curators to review and update if needed to ensure the quality and correctness of the data.

We demonstrate how LEI improves the efficiency of the data extraction process for lab reports that are part of the OHCRN database. LEI accurately extracts genetic testing report data by combining the LLM capabilities to process unstructured texts with the ability of Trie search and regular expressions to identify data unique to genetic testing. Thanks to the remote API calls, LEI's runtime is generally independent of local hardware and can thus be run on a low-end laptop.

Recent similar approaches on biomedical information extraction show precision, recall, and F1-scores between 70% and 90%[11,16,38–40] (Supplementary Table 2). Even if an automated extraction system does not reach perfect accuracy, having a recall of e.g., 70% means the curator will have to manually process the remaining 30% of the total data, which is easier than processing all the information[41]. With LEI, the burden of manual extraction was reduced for the OHCRN clinical coordinator, reducing the time needed for this recurrent process and eliminating the chance of human data input errors when sending information to OHCRN's repository.

The precision-recall analysis of LEI's output showed its quality to be on par with, and even surpass, other existing LLM-based extraction methods. In all tasks, LEI exceeds the established

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

F1-score threshold of 70%. LEI's extraction accuracy also surpasses manual extraction by humans. Mathes et al.[42] assessed the frequency of data extraction errors in manual systematic reviews and found error rates could reach up to 50%. This fact underscores the importance of building a gold standard benchmark through consensus between multiple manual extraction passes performed by multiple curators that could be used to validate LEI's results.

While most errors were caused by incorrect extractions by the LLM, a considerable number stemmed from OCR problems, where the quality of scans hindered character recognition. OCR accuracy varied significantly across documents, influenced by scan quality, typefaces, and formatting. The variability caused by these errors propagated through our pipeline, especially when leveraging LLMs to extract relevant information. OCR also commonly struggled with recognizing certain variant information, such as protein and coding changes, resulting in incomplete data extraction.

LEI provides a simple interface with pre-configured extraction tasks as well as support for defining new tasks using natural language with a few formatting rules. This enables researchers or clinicians with few or no programming experience to easily direct LEI's results, and opens up the possibility of using the solution in multiple contexts with minimal effort. The sole use of LLMs for data extraction would be ideal for simplifying the pipeline even further, as developing regular expressions can be intimidating or unapproachable without specific expertise[43], which can be the case for the average researcher or clinician. Our pipeline could then be improved by trading the regexes for LLM-prompted extraction.

While our current data extraction pipeline demonstrates significant advantages in time efficiency and accuracy for extracting relevant information from unstructured healthcare documents, certain

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

limitations need to be acknowledged. These limitations are areas for possible future improvements.

LEI was developed using a limited set of reports from nine molecular testing laboratories across Ontario. The manual extraction performed by the expert curators ensures an effective gold standard benchmark for validation. However, the limited number of sample cases available for testing means that our evaluation may not account for the full variability of report formats and structures across institutions. As the OHCRN initiative develops, more reports will become available for a larger test dataset, thus enabling more extensive evaluation. In combination with feedback from clinical data curators, these new reports can be used to continuously improve LEI.

A major concern regarding all current LLMs is hallucinations, i.e., the presence of false assertions in the generated responses of the model[44]. In earlier iterations of LEI, hallucinations were observed, especially for extracted variant descriptors, where the LLM produced results that complied with the descriptor format, but with values that were absent in the reports. To avoid hallucinations, regular expressions were used for fields with standardized formats. Additionally, LEI's implementation of the prompting strategies mentioned by Zamfirescu-Pereira et al.[37] proved to be effective for the data extraction task.

LEI relies on GPT-4o as the primary LLM. However, open-source and domain-specific models offer promising alternatives. These models provide flexibility to customize and adapt LEI for specific healthcare contexts. A significant advantage of open-source models is their ability to be deployed locally, eliminating the need to transmit sensitive patient data to third-party servers and thereby reducing privacy concerns. Local deployment can also increase LEI's adoption by other institutions with similar interests to OHCRN, allowing them to confidently implement our pipeline whilst retaining full control over the data and maintaining patient anonymity.

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Open-source fine-tuning tools such as Unsloth[45] can be integrated into LEI, offering a customizable and efficient framework for fine-tuning LLMs to better suit data extraction tasks. However, a recent study by Dorfner et al.[46] found that biomedical-specific LLMs often underperform compared to generalist models on tasks involving unseen medical data. Yet, the study highlights the use of retrieval-augmented generation (RAG) to combine the strengths of general purpose models with external domain-specific knowledge bases. Incorporating RAG can enable an LLM to extract relevant information more accurately without sacrificing general performance capabilities. Combining fine-tuning and RAG techniques could enhance LEI's performance, particularly for tasks requiring specialized knowledge, such as the Variant task.

Conclusions

In summary, we have shown that LEI provides useful assistance to clinical database human curators in terms of time savings and extraction accuracy. While it will not replace the curators, it will assist them by automatically extracting data for their review, reducing the time required for the process and reducing possible human errors. Its user-friendly, yet flexible, implementation allows for simple expansion into additional data extraction tasks as needed. While a number of limitations still exist, future iterations may soon be able to address them thanks to the rapid pace of progress in the machine learning field.

Data availability

The LEI application, as well as the code used for evaluation, is available at https://github.com/courtotlab/ohcrn_lei. It can be used and shared freely under the terms of the GNU Public License (GPL) 3.0.

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Acknowledgements

We gratefully acknowledge the contributions of the following laboratories for providing sample laboratory reports that were essential to this study: Children's Hospital of Eastern Ontario, Hamilton Health Sciences, Kingston Health Sciences Centre, London Health Sciences, Sinai Health System, North York General Hospital, Hospital for Sick Children, Trillium Health Partners, University Health Network. Their efforts and collaboration were critical to the success of this work.

This study is conducted with the support of the Ontario Institute for Cancer Research through funding provided by the Government of Ontario.

Conflicts of interest

The authors declare that they have no conflict of interest.

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Dagliati A, Malovini A, Tibollo V, Bellazzi R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. *Brief Bioinform*. 2021;22(2):812-822. doi:10.1093/bib/bbaa418

2. Cook JA, Collins GS. The rise of big clinical databases. *Br J Surg*. 2015;102(2):e93-e101. doi:10.1002/bjs.9723

3. Sondka Z, Dhir NB, Carvalho-Silva D, et al. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res*. 2023;52(D1):D1210-D1217. doi:10.1093/nar/gkad986

4. Boffa DJ, Rosen JE, Mallin K, et al. Using the National Cancer Database for Outcomes Research: A Review. *JAMA Oncol*. 2017;3(12):1722. doi:10.1001/jamaoncol.2016.6905

5. Van Dijk H, Van Wingen G, Denys D, Olbrich S, Van Ruth R, Arns M. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Sci Data*. 2022;9(1):333. doi:10.1038/s41597-022-01409-z

6. Farncombe KM, Hughes LK, Tuzlali E, others. Development of the Ontario Hereditary Cancer Research Network, a unified registry as a resource for individuals with inherited cancer syndromes: an observational registry creation protocol. *BMJ Open*. 2024;14(11):e087023. doi:10.1136/bmjopen-2024-087023

7. Garutti M, Foffano L, Mazzeo R, et al. Hereditary Cancer Syndromes: A Comprehensive Review with a Visual Tool. *Genes*. 2023;14(5):1025. doi:10.3390/genes14051025

8. Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, Ghosh P. A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. *Appl Sci*. 2021;11(18):8319. doi:10.3390/app11188319

9. Jee J, Fong C, Pichotta K, et al. Automated real-world data integration improves cancer outcome prediction. *Nature*. 2024;636(8043):728-736. doi:10.1038/s41586-024-08167-5

10. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Wren J, ed. *Bioinformatics*. 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682

11. Mulyar A, Uzuner O, McInnes B. MT-clinical BERT: scaling clinical information extraction with multitask learning. *J Am Med Inform Assoc*. 2021;28(10):2108-2115. doi:10.1093/jamia/ocab126

12. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

13. Radford A, Narasimhan K, Salimans T, Sutskever I, others. Improving language understanding by generative pre-training. Published online 2018.

14. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 Herd of Models. Published online 2024. doi:10.48550/ARXIV.2407.21783

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

15. Enis M, Hopkins M. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. Published online 2024. doi:10.48550/ARXIV.2404.13813

16. Zhu X, Dai F, Gu X, Li B, Zhang M, Wang W. GL-NER: Generation-Aware Large Language Models for Few-Shot Named Entity Recognition. In: Wand M, Malinovská K, Schmidhuber J, Tetko IV, eds. *Artificial Neural Networks and Machine Learning – ICANN 2024*. Vol 15022. Lecture Notes in Computer Science. Springer Nature Switzerland; 2024:433-448. doi:10.1007/978-3-031-72350-6_30

17. Wang S, Sun X, Li X, et al. GPT-NER: Named Entity Recognition via Large Language Models. Published online 2023. doi:10.48550/ARXIV.2304.10428

18. Smith R. An Overview of the Tesseract OCR Engine. In: *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*. IEEE; 2007:629-633. doi:10.1109/ICDAR.2007.4376991

19. EasyOCR. Published online 2024. https://github.com/JaidedAI/EasyOCR

20. Deborah RN, Rajiv SA, Vinora A, Soundarya M, Arif GM, Arif SM. Efficient Information Retrieval: AWS Textract in Action. In: *Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2024*. Vol 1. Springer Nature; 2024:95.

21. Vedhaviyassh DR, Sudhan R, Saranya G, Safa M, Arun D. Comparative Analysis of EasyOCR and TesseractOCR for Automatic License Plate Recognition using Deep Learning Algorithm. In: *2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE; 2022:1-6. doi:10.1109/iceca55336.2022.10009215

22. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat*. 2016;37(6):564-569. doi:10.1002/humu.22981

23. Rogers T. hgvs-regexp. Published online 2018. https://github.com/7ravis/hgvs-regexp

24. Rubin AF. mavehgvs (Version 0.4.0) [Computer software]. Published online 2021. doi:10.5281/zenodo.5148054

25. McKusick-Nathans Institute of Genetic Medicine M Johns Hopkins University (Baltimore. Online Mendelian Inheritance in Man, OMIM ®. Published online 2024.

26. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(1):D980-D985. doi:10.1093/nar/gkt1113

27. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311. doi:10.1093/nar/29.1.308

28. Rehm HL, Berg JS, Brooks LD, et al. ClinGen — The Clinical Genome Resource. *N Engl J Med*. 2015;372(23):2235-2242. doi:10.1056/nejmsr1406261

29. Bateman A, Martin MJ, Orchard S, et al. UniProt: the Universal Protein Knowledgebase in

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

2023. *Nucleic Acids Res*. 2022;51(D1):D523-D531. doi:10.1093/nar/gkac1052

30. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO Gene Nomenclature Committee (HGNC). *Hum Genet*. 2001;109(6):678-680. doi:10.1007/s00439-001-0615-0

31. De La Briandais R. File searching using variable length keys. In: *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference on XX - IRE-AIEE-ACM '59 (Western)*. ACM Press; 1959:295-298. doi:10.1145/1457838.1457895

32. OpenAI, Hurst A, Lerer A, et al. GPT-4o System Card. Published online 2024. doi:10.48550/ARXIV.2410.21276

33. Zhu N, Zhang N, Shao Q, Cheng K, Wu H. OpenAI's GPT-4o in surgical oncology: Revolutionary advances in generative artificial intelligence. *Eur J Cancer*. 2024;206:114132. doi:10.1016/j.ejca.2024.114132

34. OpenAI. Chat completions API. Published online 2024. https://api.openai.com/v1/chat/completions

35. OpenAI. Chat completions Docs. Published online 2024. https://platform.openai.com/docs/guides/chat-completions

36. Evstafev E. The Paradox of Stochasticity: Limited Creativity and Computational Decoupling in Temperature-Varied LLM Outputs of Structured Fictional Data. Published online 2025. doi:10.48550/ARXIV.2502.08515

37. Zamfirescu-Pereira JD, Wong RY, Hartmann B, Yang Q. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM; 2023:1-21. doi:10.1145/3544548.3581388

38. Sung M, Jeong M, Choi Y, Kim D, Lee J, Kang J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Borgwardt K, ed. *Bioinformatics*. 2022;38(20):4837-4839. doi:10.1093/bioinformatics/btac598

39. Jantscher M, Gunzer F, Kern R, Hassler E, Tschauner S, Reishofer G. Information extraction from German radiological reports for general clinical text and language understanding. *Sci Rep*. 2023;13(1):2353. doi:10.1038/s41598-023-29323-3

40. Xu T, Gu Y, Xue M, Gu R, Li B, Gu X. Knowledge graph construction for heart failure using large language models with prompt engineering. *Front Comput Neurosci*. 2024;18:1389475. doi:10.3389/fncom.2024.1389475

41. Hobbs JR. Information extraction from biomedical text. *J Biomed Inform*. 2002;35(4):260-264. doi:10.1016/S1532-0464(03)00015-7

42. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC Med Res Methodol*. 2017;17(1). doi:10.1186/s12874-017-0431-4

medRxiv preprint doi: https://doi.org/10.1101/2025.08.20.25334127; this version posted August 26, 2025. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

43. Friedl JEF. *Mastering Regular Expressions*. 3rd ed. O'Reilly Media; 2006.

44. Martino A, Iannelli M, Truong C. Knowledge Injection to Counter Large Language Model (LLM) Hallucination. In: *Lecture Notes in Computer Science*. Springer Nature Switzerland; 2023:182-185. doi:10.1007/978-3-031-43458-7_34

45. Han D, Han M. Unsloth. Published online 2023. https://unsloth.ai/

46. Dorfner FJ, Dada A, Busch F, et al. Biomedical Large Languages Models Seem not to be Superior to Generalist Models on Unseen Medical Data. Published online 2024. doi:10.48550/ARXIV.2408.13833