

Fine-tuning Large Language Models for Structured Extraction of Infectious Disease-related Information from Clinical Notes in Japanese Primary Care: Development and Validation Study

Hiroshi Yoshihara, Haruka Maeda, Yuriko Hagiwara, Daichi Sato, Kei Kitajima, Akihiro Iwata, Nicolas Van de Velde, Yuta Nakamura, Yosuke Yamagishi, Ataru Igarashi

Submitted to: Journal of Medical Internet Research
on: September 28, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 19

 Multimedia Appendixes..... 20

 Multimedia Appendix 1..... 20

CONSORT (or other) checklists..... 21

 CONSORT (or other) checklist 0..... 21

Fine-tuning Large Language Models for Structured Extraction of Infectious Disease-related Information from Clinical Notes in Japanese Primary Care: Development and Validation Study

Hiroshi Yoshihara¹ MSc; Haruka Maeda² MD, PhD, MPH; Yuriko Hagiwara¹; Daichi Sato³ PhD; Kei Kitajima³ MEng; Akihiro Iwata³ MIA; Nicolas Van de Velde⁴ PhD; Yuta Nakamura⁵ MD, PhD; Yosuke Yamagishi⁶ MD, MSc; Ataru Igarashi¹ PhD

¹Department of Health Policy and Public Health Graduate School of Pharmaceutical Sciences The University of Tokyo Tokyo JP

²Department of Respiratory Infections Institute of Tropical Medicine Nagasaki University Nagasaki JP

³M3 Inc. Tokyo JP

⁴Moderna, Inc. Cambridge US

⁵Department of Computational Diagnostic Radiology and Preventive Medicine Graduate School of Medicine The University of Tokyo Tokyo JP

⁶Division of Radiology and Biomedical Engineering Graduate School of Medicine The University of Tokyo Tokyo JP

Corresponding Author:

Hiroshi Yoshihara MSc

Department of Health Policy and Public Health

Graduate School of Pharmaceutical Sciences

The University of Tokyo

7-3-1 Hongo, Bunkyo-ku

Tokyo

JP

Abstract

Background: The Coronavirus Disease 2019 (COVID-19) pandemic highlighted the importance of timely infectious disease surveillance.

Objective: We aimed to develop a natural language processing algorithm to extract structured information on symptoms and vaccination history from free-text clinical notes in Japanese primary care. This approach enables low-latency monitoring using real-world electronic health record data.

Methods: A total of 773 clinical notes, originating from 526 unique patients, were provided by M3 Inc. through the Japan Medical Data Survey and used for analysis. Three clinicians annotated information related to infectious disease symptoms and vaccination history. The data were divided into 622 training cases and 151 evaluation cases, ensuring no patient overlap between the two sets. We developed a rule-based extraction algorithm using clinician-designed regular expressions and conducted few-shot learning and fine-tuning with both open-source and commercial large language models (LLMs). We compared the extraction accuracy (F1 score) of these approaches.

Results: Generally, the format of clinical notes tends to be similar across clinicians, allowing rule-based algorithms to achieve reasonable extraction accuracy. LLMs using few-shot prompts could extract unstructured information, such as vaccination history, with high accuracy, achieving high performance across all evaluation metrics (Anthropic Claude 3.5 Sonnet, macro-average F1 score 0.902). We also demonstrated that a fine-tuned small open-source LLM can achieve remarkably higher accuracy (Google Gemma 2 2B, macro-average F1 score 0.932) than commercial models.

Conclusions: We showed that a fine-tuned LLM can accurately extract and structure infectious disease-related information from free-text clinical notes, which implies the feasibility of digital surveillance using electronic medical record information.

(JMIR Preprints 28/09/2025:84974)

DOI: <https://doi.org/10.2196/preprints.84974>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/preprint/84974>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/84974>



Original Manuscript

Fine-tuning Large Language Models for Structured Extraction of Infectious Disease-related Information from Clinical Notes in Japanese Primary Care: Development and Validation Study

Hiroshi Yoshihara 1, Haruka Maeda 2, Yuriko Hagiwara 1, Daichi Sato 3, Kei Kitajima 3, Akihiro Iwata 3, Nicolas Van de Velde 4, Yuta Nakamura 5, Yosuke Yamagishi 6, Ataru Igarashi 1

1. Department of Health Policy and Public Health, Graduate School of Pharmaceutical Sciences, The University of Tokyo, Tokyo, Japan
2. Department of Respiratory Infections, Institute of Tropical Medicine, Nagasaki University, Nagasaki, Japan
3. M3, Inc., Tokyo, Japan
4. Moderna, Inc., Cambridge, MA, USA
5. Department of Computational Diagnostic Radiology and Preventive Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
6. Division of Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Address for correspondence:

Hiroshi Yoshihara MS

Department of Health Policy and Public Health, Graduate School of Pharmaceutical Sciences, The University of Tokyo

Address: 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Telephone: +81 3-5841-4828

Email: hiroshechka-yoshihara@g.ecc.u-tokyo.ac.jp

Abstract

Background: The Coronavirus Disease 2019 (COVID-19) pandemic highlighted the importance of timely infectious disease surveillance.

Objective: We aimed to develop a natural language processing algorithm to extract structured information on symptoms and vaccination history from free-text clinical notes in Japanese primary care. This approach enables low-latency monitoring using real-world electronic health record data.

Methods: A total of 773 clinical notes, originating from 526 unique patients, were provided by M3 Inc. through the Japan Medical Data Survey and used for analysis. Three clinicians annotated information related to infectious disease symptoms and vaccination history. The data were divided into 622 training cases and 151 evaluation cases, ensuring no patient overlap between the two sets. We developed a rule-based extraction algorithm using clinician-designed regular expressions and conducted few-shot learning and fine-tuning with both open-source and commercial large language models (LLMs). We compared the extraction accuracy (F1 score) of these approaches.

Results: Generally, the format of clinical notes tends to be similar across clinicians, allowing rule-based algorithms to achieve reasonable extraction accuracy. LLMs using few-shot prompts could extract unstructured information, such as vaccination history, with high accuracy, achieving high performance across all evaluation metrics (Anthropic Claude 3.5 Sonnet, macro-average F1 score 0.902). We also demonstrated that a fine-tuned small open-source LLM can achieve remarkably higher accuracy (Google Gemma 2 2B, macro-average F1 score 0.932) than commercial models.

Conclusions: We showed that a fine-tuned LLM can accurately extract and structure infectious disease-related information from free-text clinical notes, which implies the feasibility of digital surveillance using electronic medical record information.

Keywords

Natural language processing; Large language model; Electronic health records; EHR; Infectious diseases; Public health

Introduction

The Coronavirus Disease 2019 (COVID-19) pandemic highlighted the importance of infectious disease surveillance, and timely surveillance has become more crucial than ever in the post-COVID-19 era. After the pandemic's peak, multiple respiratory virus infections such as influenza, COVID-19, and respiratory syncytial virus (RSV) infection have been co-circulating, sometimes resulting in simultaneous outbreaks. For example, during the winter of 2022–2023, the United States experienced a “triple-demic” of SARS-CoV-2, influenza, and RSV infections that strained healthcare systems and even overwhelmed hospital capacity in some regions [1]. These pathogens also frequently present overlapping clinical symptoms, making distinguishing one infection from another difficult. Such shared symptomatology complicates timely diagnosis and highlights the need for improved testing and active surveillance to identify cases accurately [2]. Co-infections pose an additional challenge; patients infected with both influenza and COVID-19, for instance, have been shown to face a markedly higher risk of severe outcomes like intensive care unit admission [3]. Collectively, influenza, RSV infection, and COVID-19 continue to present a substantial global health challenge in the post-pandemic world, underscoring the importance of robust infectious disease surveillance systems.

In Japan, harnessing electronic health records (EHRs) data has the potential to offer particular advantages for infectious disease surveillance compared to traditional claims data. The country's conventional surveillance systems, such as the national sentinel provider network for influenza, typically aggregate case reports on a weekly basis and often incur reporting lags of about one week [4]. This delay can slow down outbreak detection. To achieve lower latency, researchers have turned to alternative data sources such as healthcare insurance claims and pharmacy records. Notably, a real-time prescription surveillance system introduced in Japan demonstrated that daily tracking of antiviral drug prescriptions could closely estimate influenza case numbers in near real time [4]. However, while claims and prescription data provide faster updates, they carry only limited clinical details, primarily billing codes for diagnoses or medications, and may not capture the nuanced symptom information available in a clinical note. Claims-based metrics also risk bias from changing coding practices and may lack clinical fidelity [5]. By contrast, EHR clinical notes contain rich descriptions of patient symptoms, findings, and vaccination history recorded at the point of care. These narratives can be accessed soon after the medical encounter, giving a timelier and more detailed picture of emerging infections. Studies have found that surveillance based on EHR-derived clinical data can provide more objective and reliable estimates of disease incidence than surveillance based on claims data [5].

Natural language processing (NLP) and deep learning are promising tools to address some of these surveillance challenges. NLP is a branch of artificial intelligence that focuses on enabling computers to understand and generate human language [6]. In recent years, deep learning techniques have advanced NLP capabilities, allowing automated analysis of large volumes of text. This is especially significant in healthcare, where enormous amounts of data are recorded as unstructured text in EHRs [6]. Applications of NLP in healthcare are broad: systems have been developed to classify medical notes by content, recognize and extract clinical entities (such as symptoms, diagnoses, or medications) from text, summarize patient records, and even translate medical jargon into lay language [7]. Recent large language models (LLMs) are used to parse free-text EHR narratives and extract structured clinical concepts with unprecedented accuracy. Many studies concluded that commercial and open-source LLMs could outperform traditional NLP systems in extracting various information from medical notes [8,9].

Internationally, numerous studies have demonstrated the utility of deep learning-based NLP models for identifying COVID-19 cases or extracting related clinical information from EHRs [10,11]. In Japan, epidemiologists have mainly used structured data (e.g., outpatient pharmacy prescription records [12]) to monitor infectious disease trends. To our knowledge, no research has yet attempted to directly extract infectious disease information from unstructured Japanese clinical narratives. Here, we aim to address this gap by developing a NLP algorithm that leverages state-of-the-art LLM to extract infectious disease-related symptoms and vaccination details from free-text clinical notes written in Japanese. Such an algorithm serves as a valuable foundation for infectious disease surveillance, supporting more timely public health interventions and enhancing responsive infectious disease surveillance capacity.

Methods

Data Source and Study Population

A pseudonymized subset of clinical notes in the Japan Medical Data Survey (JAMDAS), an outpatient healthcare database of nationwide primary care clinics curated by M3 Inc [13], was used for analysis. This dataset included 2,020 clinical notes from 1,036 patients, aged 0 to 84 years, who visited five outpatient clinics for patients with fever or suspected infectious disease between October 2020 and September 2023. Patients were randomly sampled, with no filtering based on disease names or prescriptions. Therefore, the dataset includes patients with a variety of medical conditions.

After removing missing values and duplicate records, three clinicians within the research team performed annotations on the following items: body temperature, fatigue, muscle and joint pain, headache and dizziness, respiratory symptoms, visible changes in the throat, gastrointestinal symptoms, sensory abnormalities, other symptoms, vaccination information, and onset date. Detailed annotation items, criteria, and extraction formats are described in Table 1. The raw onset description, parsed onset date, body temperature, and description of vaccination history were annotated as text data types, while all other items were classified as binary variables. Due to resource constraints, we analyzed 773 clinical notes from 526 patients for which annotations by at least two clinicians agreed.

Ethical Considerations

The Nihonbashi-Sakura Clinic Ethics Committee approved this study, and informed consent was waived.

Algorithm Development

The dataset was randomly divided into 622 training cases and 151 evaluation cases, ensuring no patient overlap between the two sets. We developed and compared the following approaches: a rule-based algorithm crafted by clinicians; few-shot learning using commercial and open-source LLMs; and fine-tuning of open-source LLMs. The rule-based algorithm was developed by clinicians who carefully reviewed the training dataset and translated patterns for each extraction item into regular expressions.

After several preliminary performance experiments, this study selected Anthropic Claude 3.5 Sonnet (version 20241022) [14] as the representative commercial LLM and the Google Gemma 2 family (27B and 2B models) [15] as the representative open-source LLMs. We tested several prompt patterns and settled on one that outputs results in a JSON format, where each extraction item serves

as a key and the corresponding extracted value as its value (e.g., {“fever”: 1, “temperature”: 38.3, “headache”: 0}). The prompt used is shown in Table 2.

Few-shot learning is a technique in which a model is given a small number of example inputs and outputs to learn from before performing a task [16]. This allows the model to better understand the task and generate more accurate results without updating its parameters. We initially planned to compare zero-shot inference with few-shot learning inference. However, we found that zero-shot outputs often did not adhere strictly to the JSON format, making reliable evaluation difficult. Therefore, we used the few-shot learning results as the baseline for LLM performance. We randomly sampled 18 examples, ensuring that each extraction item had at least one positive instance included.

To further improve performance, we applied supervised fine-tuning to the open-source LLMs. The training dataset was split into 500 samples for training and 122 samples for validation, with the validation set used to monitor overfitting during training. We performed fine-tuning using the instruction-tuned Google Gemma 2 models (gemma-2-2b-it and gemma-2-27b-it). Full fine-tuning was conducted on the 2B model, while due to hardware limitations, we applied Quantized Low-Rank Adaptation (QLoRA) for fine-tuning the 27B model. LoRA (Low-Rank Adaptation) is a technique that reduces the number of trainable parameters by injecting low-rank matrices into a pre-trained model’s weights, enabling efficient fine-tuning without updating the entire model [17]. QLoRA combines LoRA with quantization, compressing model weights to reduce memory usage while maintaining accuracy, allowing large models to be fine-tuned on limited hardware resources [18]. The specific training parameters are detailed in Supplementary Table 1 in Multimedia Appendix 1. Fine-tuning was performed using the transformers [19] and trl [20] libraries on a single NVIDIA RTX 6000 Ada GPU.

Evaluation

We grouped extraction items by category and calculated the F1 score for each group. The macro-average F1 score across all groups was used as the model’s performance metric. The F1 score measures accuracy by balancing precision and recall, with 1 meaning perfect performance. For items extracted as text, the F1 score was computed based on exact matches. For reference, sensitivity and specificity were also calculated.

Results

The data flowchart and an overview of the experiments are shown in Figure 1. Background information and distribution of extraction items for the training and evaluation datasets are presented in Table 1.

As shown in Supplementary Table 2 in Multimedia Appendix 1, although there were some systematic differences in clinical note templates across clinics and clinician preferences within clinics, the notes were generally written in the SOAP (Subjective, Objective, Assessment, Plan) format, showing overall similarity. The performance evaluation results for each model are presented in Table 3. Some extraction items exhibited relatively low variability in their descriptions, indicating they were similar across many clinicians; consequently, the rule-based algorithm achieved high performance on these items. For items with high variability in descriptions, such as vaccination information and onset dates, the rule-based algorithm showed low accuracy; however, few-shot learning with LLMs achieved high accuracy. In particular, the commercial model Anthropic Claude 3.5 Sonnet achieved a macro-average F1 score of 0.902, with F1 scores above 0.8 for most extraction items.

In the fine-tuning experiments, both the Google Gemma 2 2B and 27B models showed substantial performance improvements. The 2B model achieved a macro-average F1 score of 0.840, while the 27B model fine-tuned with QLoRA reached 0.932.

A qualitative error analysis of misclassifications by the LLM revealed that many errors occurred when multiple extraction targets were listed consecutively, separated by commas or periods, followed by a collective negation (e.g., “*Diarrhea, nausea negative.*”, “*Abdominal pain and vomiting. Wheezing negative.*”). These cases can be interpreted in multiple ways and are examples where even human experts would find it difficult to reach a consensus.

Discussion

Main Findings

We demonstrated that LLMs can extract infectious disease-related information and vaccination data from free-text clinical notes written in Japanese in EHRs with very high accuracy. Notably, for items with high variability in descriptions, LLMs showed remarkable improvement over rule-based algorithms using regular expressions. In particular, the commercial model Anthropic Claude 3.5 Sonnet (with few-shot learning) achieved high performance with an F1 score of 0.902 without additional training, while fine-tuning the open-source Google Gemma 2 27B model using QLoRA on only 500 samples enabled us to achieve even higher performance with an F1 score of 0.932.

Given concerns about security risks when sending clinical note texts containing sensitive personal information to external commercial models, achieving comparable or better performance with a smaller open-source model demonstrates the feasibility of safely processing data within a closed environment. This advantage enhances the real-world applicability of the algorithm. Going forward, we intend to investigate the relationship between the algorithm’s extraction outputs and actual infectious disease trends observed in real-world settings.

Limitations

This study has several limitations. First, the analysis dataset was derived solely from five clinics. The algorithm’s accuracy may decrease in other clinics with markedly different clinical note styles. Second, we were unable to establish an external evaluation dataset. Ideally, external validation using data from facilities not involved in training would be preferred; however, due to resource constraints, this was not feasible in the present study. These two limitations primarily relate to the external validity of the model’s performance. Finally, the performance evaluation of this study may be insufficient. For many patients, only a single clinical note exists, making it unclear whether the algorithm can accurately capture temporal changes in symptoms over time. In summary, future research should include a more comprehensive performance evaluation using an external validation dataset collected from clinics different from those used in this study. Additionally, studies should focus on developing algorithms to predict infectious disease trends based on the extracted symptom information.

Conclusions

By fine-tuning the open-source LLM Google Gemma 2, we developed a model specialized in

extracting and structuring infectious disease-related information from Japanese free-text clinical notes that outperforms commercial models.

Tables

Table 1. Dataset characteristics and extraction item definitions

	Training dataset	Test dataset	p-value	Definition	Format	Example
N	622	151	-			
Demographic information						
Age, mean(std)	16.2 (15.7)	18.3 (16.8)	0.156			
Male, n(%)	350 (56.3%)	70 (46.4%)	0.036			
Symptom onset						
Parsed date mentioned, n(%)	496 (79.7%)	119 (78.8%)	0.886	Date of initial symptom onset determined from the visit date and clinical note description.	Text	2022-10-15
Raw description mentioned, n(%)	503 (80.9%)	121 (80.1%)	0.928	Raw description of the initial symptom onset date in the clinical note.	Text	昨日朝 (En: yesterday morning)
Body temperature						
Body temperature mentioned, n(%)	420 (67.5%)	99 (65.6%)	0.716	Highest body temperature recorded in the clinical note. Annotations were standardized to the format "3x.x".	Text	39.4
Slight fever, n(%)	78 (12.5%)	18 (11.9%)	0.945	Whether slight fever is mentioned in the clinical note or patient's highest body temperature between 37.0°C and 37.4°C.	Binary	1
Fever, n(%)	369 (59.3%)	89 (58.9%)	1.000	Whether fever is mentioned in the clinical note or patient's highest body temperature 37.0°C and higher.	Binary	0
High fever, n(%)	190 (30.5%)	47 (31.1%)	0.968	Whether high fever is mentioned in the clinical note or patient's highest body temperature 38.0°C and higher.	Binary	1
Chills, n(%)	12 (1.9%)	3 (2.0%)	1.000	Presence of the symptom.	Binary	0
Fatigue and muscle/joint pain						
Fatigue, n(%)	89 (14.3%)	21 (13.9%)	1.000	Presence of the symptom.	Binary	1
Muscle/joint pain, n(%)	18 (2.9%)	5 (3.3%)	0.997	Presence of the symptom.	Binary	0
Headache and dizziness						
Dizziness, n(%)	3 (0.5%)	1 (0.7%)	1.000	Presence of the symptom.	Binary	1
Headache, n(%)	139 (22.3%)	33 (21.9%)	0.983	Presence of the symptom.	Binary	0
Respiratory symptoms						
Runny nose, n(%)	222 (35.7%)	54 (35.8%)	1.000	Presence of the symptom.	Binary	1
Nasal congestion, n(%)	17 (2.7%)	4 (2.6%)	1.000	Presence of the symptom.	Binary	0
Sore throat, n(%)	239 (38.4%)	57 (37.7%)	0.952	Including discomfort in the throat.	Binary	1
Cough, n(%)	275	68	0.928	Presence of the symptom.	Binary	0

	(44.2%)	(45.0%)				
Sputum, n(%)	104 (16.7%)	24 (15.9%)	0.902	Presence of the symptom.	Binary	1
Chest pain, n(%)	3 (0.5%)	0 (0.0%)	0.900	Presence of the symptom.	Binary	0
Shortness of breath, n(%)	3 (0.5%)	0 (0.0%)	0.900	Presence of the symptom.	Binary	1
Breathing sounds, n(%)	11 (1.8%)	3 (2.0%)	1.000	Presence of abnormal breath sounds based solely on objective physical examination findings (excluding patient-reported symptoms).	Binary	0
Throat appearance						
Tonsillar hypertrophy, n(%)	1 (0.2%)	0 (0.0%)	1.000	Presence of the symptom.	Binary	1
Tonsillar exudates, n(%)	3 (0.5%)	1 (0.7%)	1.000	Presence of the symptom.	Binary	0
Pharyngeal redness, n(%)	113 (18.2%)	27 (17.9%)	1.000	Including redness of the tonsils.	Binary	1
Gastrointestinal symptoms						
Abdominal pain, n(%)	22 (3.5%)	5 (3.3%)	1.000	Presence of the symptom.	Binary	1
Diarrhea, n(%)	28 (4.5%)	7 (4.6%)	1.000	Presence of the symptom.	Binary	0
Nausea, n(%)	21 (3.4%)	5 (3.3%)	1.000	Presence of the symptom.	Binary	1
Vomiting, n(%)	21 (3.4%)	5 (3.3%)	1.000	Presence of the symptom.	Binary	0
Other symptoms						
Taste abnormality, n(%)	4 (0.6%)	0 (0.0%)	0.722	Presence of the symptom.	Binary	1
Smell abnormality, n(%)	1 (0.2%)	1 (0.7%)	0.845	Presence of the symptom.	Binary	0
Lymph node swelling, n(%)	3 (0.5%)	1 (0.7%)	1.000	Including swelling of all lymph nodes, regardless of location.	Binary	1
Rash, n(%)	11 (1.8%)	3 (2.0%)	1.000	Presence of rash or skin eruption.	Binary	0
Vaccination information						
COVID-19 vaccination, n(%)	141 (22.7%)	34 (22.5%)	1.000	Presence of COVID-19 vaccination, regardless of the number of doses.	Binary	1
Raw description mentioned, n(%)	194 (31.2%)	47 (31.1%)	1.000	Descriptions related to vaccines (including those other than COVID-19).	Text	コロナワクチン接種：2回、最終2021年10月16日、Pfizer (En: COVID-19 vaccination: 2 doses, most recent on October 16, 2021, Pfizer)

Binary variables represent the number of positive cases, while continuous and text variables represent the number of notes each item mentioned. The p-value indicates the result of a test comparing the training and test datasets (chi-square test for binary and text variables, Mann-Whitney U test for continuous variables).

Table 2. LLM prompts for data extraction

Prompt for few-shot learning

System: Carefully analyze the following instruction, drawing upon your extensive knowledge and relevant references. Keep the answer short and do not provide explanations or notes.												
#	Instructions:											
You task is to extract items from a medical record in the context, and return the results as JSON strings. If you don't find item in the context or you are not sure, skip the item.												
#	Format:											
Return	JSON	strings	ONLY	in	a	single						line.
Output	a	single	JSON	with	a	multiple						keys.
JSON	key	must	be	one	of	the	items	below,	do	NOT	change	the
JSON		value		is	the	the	extracted	such		number	as	or
Remove				units								kg.
#	Items:											
{description	binary	items,	negative	statement	=	0,	positive	statement	=	1.		
#	Examples:											
{multiple	examples}											
#	Now	please	extract	items	from	the	context	below.				
Context:	{clinical note to process}											
Answer:												
Prompt for fine-tuning												
#	Instructions:											
You task is to extract items from a medical record in the context, and return the results as JSON strings. If you don't find item in the context or you are not sure, skip the item.												
#	Format:											
Return	JSON	strings	ONLY	in	a	single						line.
Output	a	single	JSON	with	a	multiple						keys.
JSON	key	must	be	one	of	the	items	below,	do	NOT	change	the
JSON		NOT		is	change	the	extracted	such		number	as	or
Remove		value		units								kg.
#	Items:											
{description	binary	items,	negative	statement	=	0,	positive	statement	=	1.		
#	Examples:											
{single	example}											
#	Now	please	extract	items	from	the	context	below.				
Context:	{clinical note to process}											
Answer:												

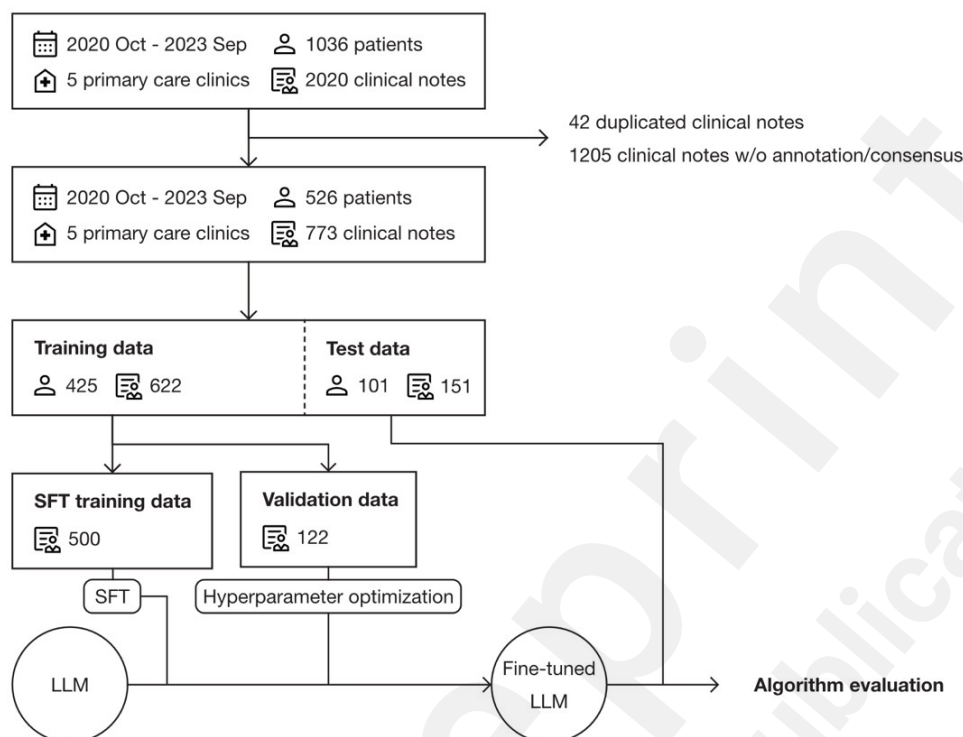
Table 3. Macro-average F1 score for each extraction category by algorithm

Category	Rule-based algorithm	Gemma 2 2B + FSL	Gemma 2 27B + FSL	Claude Sonnet 3.5 + FSL	Gemma 2 2B + SFT	Gemma 2 27B + QLoRA SFT
Body temperature	0.878	0.680	0.885	0.920	0.880	0.966
Fatigue and muscle/joint pain	0.964	0.756	0.906	0.982	0.868	0.964
Headache and dizziness	0.899	0.805	0.973	0.973	0.973	0.973
Respiratory symptoms	0.912	0.695	0.949	0.960	0.890	0.962
Throat appearance	0.741	0.680	0.866	0.857	0.918	0.951
Gastrointestinal symptoms	0.767	0.829	0.933	0.917	0.818	0.978
Other symptoms	0.533	0.000	0.750	0.800	0.500	0.750
Vaccination information	0.443	0.324	0.747	0.810	0.818	0.884
Onset date	0.175	0.832	0.928	0.900	0.892	0.962
Overall (macro-average)	0.701	0.622	0.882	0.902	0.840	0.932

FSL, few-shot learning; SFT, supervised fine-tuning.

Figures

Figure 1. Data flowchart and study overview



SFT, supervised fine-tuning; LLM, large language model.

Acknowledgments

We're grateful to M3, Inc. for their dedicated efforts in maintaining and developing JAMDAS database, which was used for this analysis. Grammatical corrections to the manuscript were performed with the aid of Google's Gemini 2.5 Flash/Pro.

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article. HY, Ataru Igarashi, DS, KK, and Akihiro Iwata receive research grant from Moderna Inc, but not relevant to this study.

Authors' Contributions

All authors contributed to the conceptualization of the study and the final review of the manuscript. HY was responsible for the methodology, formal analysis, visualization, and drafting of the original manuscript. In collaboration with HY, YN and YY also performed key work on the methodology and formal analysis. Data curation was carried out by DS, KK, and Akihiro Iwata. Validation of the work was performed by HY, HM, YH, NV, and Ataru Igarashi. Finally, Ataru Igarashi provided overall supervision for the project. All authors have approved the final manuscript and are accountable for all aspects of the work.

Conflicts of Interest

None declared.

Data Availability

The data used in this study were obtained under a restricted-use agreement with M3, Inc. and are therefore not freely available to the public. Any inquiries regarding the data can be directed to the corresponding author.

Abbreviations

COVID-19: Coronavirus Disease 2019

EHR: electronic health record

FSL: few-shot learning

JAMDAS: Japan Medical Data Survey

LLM: large language model

LoRA: Low-Rank Adaptation

NLP: natural language processing

QLoRA: Quantized Low-Rank Adaptation

RSV: respiratory syncytial virus

SFT: supervised fine-tuning

SOAP: Subjective, Objective, Assessment, Plan

References

1. Boehm AB, Wolfe MK, White BJ, Hughes B, Duong D, Bidwell A. More than a tripledemic: Influenza A virus, respiratory syncytial virus, SARS-CoV-2, and human metapneumovirus in wastewater during winter 2022-2023. *Environ Sci Technol Lett* 2023 Aug 8;10(8):622–627. PMID:37577361
2. Hanage WP, Schaffner W. Burden of acute respiratory infections caused by influenza virus, respiratory syncytial virus, and SARS-CoV-2 with consideration of older adults: A narrative review. *Infect Dis Ther Springer Science and Business Media LLC*; 2025 Jan;14(Suppl 1):5–37. PMID:39739200
3. Cong B, Deng S, Wang X, Li Y. The role of respiratory co-infection with influenza or respiratory syncytial virus in the clinical severity of COVID-19 patients: A systematic review and meta-analysis. *J Glob Health* 2022 Sept 17;12:05040. PMID:36112521
4. Sugawara T, Ohkusa Y, Ibuka Y, Kawanohara H, Taniguchi K, Okabe N. Real-time prescription surveillance and its application to monitoring seasonal influenza activity in Japan. *J Med Internet Res JMIR Publications Inc.*; 2012 Jan 16;14(1):e14. PMID:22249906
5. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, Kadri SS, Angus DC, Danner RL, Fiore AE, Jernigan JA, Martin GS, Septimus E, Warren DK, Karcz A, Chan C, Menchaca JT, Wang R, Gruber S, Klompas M, CDC Prevention Epicenter Program. Incidence

- and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *JAMA* 2017 Oct 3;318(13):1241–1249. PMID:28903154
6. Jerfy A, Selden O, Balkrishnan R. The growing impact of Natural Language Processing in healthcare and public health. *Inquiry* 2024 Jan;61:469580241290095. PMID:39396164
 7. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, Turner K. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Comput Biol Med Elsevier BV*; 2023 Mar;155(106649):106649. PMID:36805219
 8. Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, Papadopoulos N, Odavic D, Risteski P, Haeussler A, Dzemali O. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. *BMJ Health Care Inform* 2025 Jan 19;32(1). PMID:39832824
 9. Huang J, Yang DM, Rong R, Nezafati K, Treager C, Chi Z, Wang S, Cheng X, Guo Y, Klesse LJ, Xiao G, Peterson ED, Zhan X, Xie Y. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ Digit Med Springer Science and Business Media LLC*; 2024 May 1;7(1):106. PMID:38693429
 10. Homburg M, Meijer E, Berends M, Kupers T, Olde Hartman T, Muris J, de Schepper E, Velek P, Kuiper J, Berger M, Peters L. A natural language processing model for COVID-19 detection based on Dutch general practice electronic health records by using bidirectional encoder representations from transformers: Development and validation study. *J Med Internet Res* 2023 Oct 4;25:e49944. PMID:37792444
 11. Wagner T, Shweta F, Murugadoss K, Awasthi S, Venkatakrishnan AJ, Bade S, Puranik A, Kang M, Pickering BW, O'Horo JC, Bauer PR, Razonable RR, Vergidis P, Temesgen Z, Rizza S, Mahmood M, Wilson WR, Challener D, Anand P, Liebers M, Doctor Z, Silvert E, Solomon H, Anand A, Barve R, Gores G, Williams AW, Morice WG 2nd, Halamka J, Badley A, Soundararajan V. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. *Elife eLife Sciences Publications, Ltd*; 2020 July 7;9. PMID:32633720
 12. Miyawaki A, Kitajima K, Iwata A, Sato D, Tsugawa Y. Antibiotic prescription for outpatients with COVID-19 in primary care settings in japan. *JAMA Netw Open* 2023 July 3;6(7):e2325212. PMID:37490294
 13. Corporate Information. M3, Inc. Available from: <https://corporate.m3.com/en/corporate/> [accessed June 25, 2025]
 14. Claude 3.5 Sonnet. Available from: <https://www.anthropic.com/news/claude-3-5-sonnet> [accessed June 25, 2025]
 15. Gemma Team, Riviere M, Pathak S, Sessa PG, Hardin C, Bhupatiraju S, Hussenot L, Mesnard T, Shahriari B, Ramé A, Ferret J, Liu P, Tafti P, Friesen A, Casbon M, Ramos S, Kumar R, Lan CL, Jerome S, Tsitsulin A, Vieillard N, Stanczyk P, Girgin S, Momchev N, Hoffman M, Thakoor S, Grill J-B, Neyshabur B, Bachem O, Walton A, Severyn A, Parrish A, Ahmad A, Hutchison A, Abdagic A, Carl A, Shen A, Brock A, Coenen A, Laforge A, Paterson A, Bastian B, Piot B, Wu B, Brandon Royal, Chen C, Kumar C, Perry C, Welty C, Choquette-Choo CA, Sinopalnikov D, Weinberger D, Vijaykumar D, Rogozińska D, Herbison D, Bandy E, Wang E, Noland E, Moreira E, Senter E, Eltyshv E, Visin F, Rasskin G, Wei G, Cameron G, Martins G, Hashemi H,

- Klimczak-Plucińska H, Batra H, Dhand H, Nardini I, Mein J, Zhou J, Svensson J, Stanway J, Chan J, Zhou JP, Carrasqueira J, Iljazi J, Becker J, Fernandez J, van Amersfoort J, Gordon J, Lipschultz J, Newlan J, Ji J-Y, Mohamed K, Badola K, Black K, Millican K, McDonell K, Nguyen K, Sodhia K, Greene K, Sjoesund LL, Usui L, Sifre L, Heuermann L, Lago L, McNealus L, Soares LB, Kilpatrick L, Dixon L, Martins L, Reid M, Singh M, Iverson M, Görner M, Velloso M, Wirth M, Davidow M, Miller M, Rahtz M, Watson M, Risdal M, Kazemi M, Moynihan M, Zhang M, Kahng M, Park M, Rahman M, Khatwani M, Dao N, Bardoliwalla N, Devanathan N, Dumai N, Chauhan N, Wahltinez O, Botarda P, Barnes P, Barham P, Michel P, Jin P, Georgiev P, Culliton P, Kuppala P, Comanescu R, Merhej R, Jana R, Rokni RA, Agarwal R, Mullins R, Saadat S, Carthy SM, Cogan S, Perrin S, Arnold SMR, Krause S, Dai S, Garg S, Sheth S, Ronstrom S, Chan S, Jordan T, Yu T, Eccles T, Hennigan T, Kocisky T, Doshi T, Jain V, Yadav V, Meshram V, Dharmadhikari V, Barkley W, Wei W, Ye W, Han W, Kwon W, Xu X, Shen Z, Gong Z, Wei Z, Cotruta V, Kirk P, Rao A, Giang M, Peran L, Warkentin T, Collins E, Barral J, Ghahramani Z, Hadsell R, Sculley D, Banks J, Dragan A, Petrov S, Vinyals O, Dean J, Hassabis D, Kavukcuoglu K, Farabet C, Buchatskaya E, Borgeaud S, Fiedel N, Joulin A, Kenealy K, Dadashi R, Andreev A. Gemma 2: Improving open language models at a practical size. arXiv [csCL]. 2024. Available from: <http://arxiv.org/abs/2408.00118>
16. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. arXiv [csCL]. 2020. Available from: <http://arxiv.org/abs/2005.14165>
 17. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-Rank Adaptation of large language models. arXiv [csCL]. 2021. Available from: <http://arxiv.org/abs/2106.09685>
 18. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv [csLG]. 2023. Available from: <http://arxiv.org/abs/2305.14314>
 19. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S, Drame M, Lhoest Q, Rush AM. Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations Online: Association for Computational Linguistics; 2020. p. 38–45.
 20. von Werra L, Belkada Y, Tunstall L, Beeching E, Thrush T, Lambert N, Huang S, Rasul K, Gallouédec Q. TRL: Transformer Reinforcement Learning. GitHub repository. GitHub; 2020. Available from: <https://github.com/huggingface/trl>

Supplementary Files

Multimedia Appendixes

Parameters used for model training and examples of clinical notes.

URL: <http://asset.jmir.pub/assets/6eca16612ca831f75b84aab0c1da91a8.docx>

CONSORT (or other) checklists

TRIPOD-LLM checklist.

URL: <http://asset.jmir.pub/assets/112141cf3c956e7dc4227873c7b4f71e.pdf>