Article

Not peer-reviewed version

Comparison of Large Language Models Versus Traditional Information Extraction Methods for Real World Evidence of Patient Symptomatology in Acute and Post-Acute Sequelae of SARS-CoV-2

Vedansh Thakkar [*,†], Greg M. Silverman [†], Abhinab Kc, Nicholas E. Ingraham, Emma Jones, Samantha King, Christopher J. Tignanelli

Posted Date: 20 September 2024

doi: 10.20944/preprints202409.1518.v1

Keywords: Information Extraction; Natural Language Processing; Large Language Models; Electronic Health Records

Article

Comparison of Large Language Models versus Traditional Information Extraction Methods for Real World Evidence of Patient Symptomatology in Acute and Post-Acute Sequelae of SARS-CoV-2

Vedansh Thakkar [1,2,*,†], Greg M. Silverman [1,2,†], Abhinab Kc [3], Nicholas E. Ingraham [4], Emma Jones [1], Samantha King [5] and Christopher J. Tignanelli [1,2,6]

[1] Department of Surgery, University of Minnesota, Minneapolis MN
[2] Natural Language Processing/Information Extraction Program, University of Minnesota, Minneapolis MN
[3] University of Minnesota Medical School, Minneapolis MN
[4] Department of Pulmonary, Allergy, Critical Care, And Sleep Medicine, University of Minnesota, Minneapolis MN
[5] Department of Surgery, University of Washington, Seattle WA
[6] Center for Learning Health Systems Sciences, University of Minnesota, Minneapolis MN
* Correspondence: vthakkar@umn.edu
† equal contribution first authors.

Abstract. Patient symptoms play a critical role in disease progression and diagnosis, yet they are most often captured in unstructured clinical notes. This study explores use of large language models (LLMs) for extracting patient symptoms from clinical notes, comparing their performance against rule-based information extraction (IE) systems, like BioMedICUS. By fine-tuning an LLM on diverse corpora from multiple healthcare institutions, we aimed to improve symptom extraction accuracy and efficiency with symptoms related to acute and post-acute sequelae of SARS-CoV-2. We also conducted prevalence analysis to highlight significant differences in symptom prevalence across corpora and performed fairness analysis to assess the model's equity across race and gender. Our findings indicate that while LLMs can match the effectiveness of rule-based IE methods, they face significant challenges related to demographic bias and generalizability due to variability in training corpora. This evaluation revealed overfitting and insufficient generalization, especially when models were trained predominantly on limited datasets with single annotator bias. This study also revealed that LLMs offer substantial advantages in efficiency, adaptability, and scalability for biomedical IE, marking a transformative shift in clinical data extraction processes. These results provide real-world evidence of the necessity for diverse, high-quality training datasets and robust annotation processes to enhance LLM's performance and reliability in clinical applications.

Keywords: information extraction; natural language processing; large language models; electronic health records

1. Introduction

Patient symptoms are critical to understanding disease progression and diagnosis [1,2]. Indeed, symptoms offer both diagnostic and prognostic capacity [2]. Structured electronic health record (EHR) data rarely capture patient symptoms, leaving them mostly within unstructured sections of clinical notes such as the history of present illness (HPI) [2]. Thus, there is a need to develop natural language processing (NLP) solutions that can reliably extract this information from clinical notes to facilitate their integration into downstream applications. This study provides a comprehensive exploration of using large language models (LLMs) for generative question-answering (QA) tasks in NLP for symptom extraction from unstructured clinical notes. It evaluates their performance against rule-based information extraction (IE) methods, particularly those reliant on the unified medical

 doi:10.20944/preprints202409.1518.v

language system (UMLS). By demonstrating the comparable effectiveness of LLMs to rule-based IE methods and emphasizing their inherent advantages, such as handling concept synonymy without extensive manual intervention, this study envisions a future where labor-intensive lexica creation by subject matter experts (SMEs) becomes obsolete.

1.1. Information Extraction

Rule-based Concept Extraction. The UMLS metathesaurus integrates biomedical concepts from diverse vocabularies into concept unique identifiers (CUIs) [3]. With over 4 million concepts grouped into 15 semantic groups, 133 semantic types, and 54 semantic relationships, the UMLS facilitates various applications, including IE [3–5]. PASCLex is a comprehensive lexicon of UMLS concepts and their synonyms related to Post-Acute Sequelae of SARS-CoV-2 (PASC, also known as "Long COVID") symptoms that was created by analyzing a large corpus of clinical notes [6]. PASCLex aims to improve the identification and analysis of symptoms beyond those specific to PASC. Clinical NLP systems like The BioMedical Information Collection and Understanding System (BioMedICUS), developed by the University of Minnesota's Natural Language Processing/Information Extraction Program, leverages the UMLS for large-scale extensive text analysis, enabling annotation of concepts and their negations, sentence boundaries, and note section headers [7–9].

LLMs. LLMs autonomously glean relevant features from data, bypassing the laborious and expertise-dependent manual feature engineering process [10]. Moreover, they adeptly navigate intricate linguistic structures like long-distance dependencies, anaphora, and ambiguity, thereby augmenting accuracy in IE tasks [11]. By fine-tuning, LLMs swiftly acclimate to diverse domains and IE tasks, requiring minimal additional training data, thus rendering them versatile for myriad applications [12]. Given that LLMs continue to improve and become more accessible, they are poised to play a significant role in the future of IE.

1.2. Symptomatology

Background. Clinical practice faces challenges in patient diagnosis, particularly in primary care, where, for example, fatigue is a common symptom among many diseases and in isolation may not provide significant predictive power. Predictive models that use a patient's presenting symptoms to provide clinical decision support can help reduce diagnostic error. However, in order for these systems to be accurate, there is a need to extract reliable information related to patient symptomatology, which most often exists in unstructured data. Studies have shown that NLP methods using patient symptoms mapped to UMLS concepts can increase efficacy for predictive modeling in various contexts, including better colorectal cancer prediction [13] enhanced influenza prediction [14], and detection of surgical complications [15].

Recognizing the worldwide burden of PASC and the need for rapid diagnosis, in 2020, our team developed an NLP system as part of an artificial intelligence (AI) automation pipeline that could autonomously extract acute COVID-19 and PASC symptoms from longitudinal clinical notes at scale across a network of 12 Midwest U.S. hospitals and 60 primary care clinics [2]. Both structured clinical data and NLP-based features engineered from patient symptoms extracted at scale from clinical notes were used to populate a registry of COVID-19 patients [2,16]. We utilized our AI automation pipeline for monitoring PASC at the patient level for over 83,850 patients diagnosed with acute COVID-19, showing that these data could effectively surveil post-COVID patients to accurately identify patients suspected of having PASC [17]. This previous methodology utilized both rule-based and UMLS-based approaches for autonomous symptom extraction within our AI pipeline [2,16]. Over the past 4 years, LLMs have been developed and are rapidly becoming the status quo in general purpose NLP. However, their role is less certain within biomedical applications, and previous work has shown that bidirectional encoder representations from transformers (BERT) [18] based approach may outperform LLMs for certain specific biomedical tasks [19].

 doi:10.20944/preprints202409.1518.v

1.3. Objective

The objective of this study was to fine-tune an LLM for extraction of patient symptoms from unstructured clinical notes and demonstrate the comparable effectiveness of the fine-tuned LLM to rule-based IE methods. This study aimed to compare the LLM against BioMedICUS for symptom detection from clinical notes. We also aimed to explore the possibility of integrating the fine-tuned LLM into the BioMedICUS pipeline. Notably, BioMedICUS stands out for its capability to handle notes in rich text format (RTF), a feature not commonly found in other NLP systems, which made it an appealing choice for our purposes [8].

2. Materials and Methods

2.1. Data Sources

Corpora. This study utilized three different corpora of ground truth labeled clinical notes for fine-tuning and testing of our LLM and for comparison with BioMedICUS. University of Minnesota (UMN) COVID, a corpus of emergency department (ED) visit notes for patients diagnosed with acute COVID-19 as described in [2] and UMN PASC, a corpus of ED and outpatient (OP) visits for patients diagnosed with acute COVID-19 followed by lingering symptoms were provided by UMN/MHealth Fairview and were manually annotated by three subject matter experts with an inter rater reliability of 0.68 computed using Fleiss' kappa [20]. The National COVID Cohort Collaborative (N3C) COVID, a fully de-identified corpus of notes for patients diagnosed with acute COVID-19 was provided by the Mayo Clinic. These three corpora were used because they were available with ground truth annotations for symptom identification and offered an opportunity for both internal and external validation of the fine-tuned LLM. A summary of each corpus is presented in Table 1.

Table 1. Summary of Corpora.

Corpus	Source	Note Count
UMN PASC	Outpatient (OP)	387
	Emergency Department (ED)	84
UMN COVID	Emergency Department (ED)	46
N3C COVID	Not Available (NA)	148

UMN COVID. Criteria for inclusion and manual curation methods of this corpus are outlined in [2]. Demographics for this corpus are presented in Table 2.

UMN PASC. The inclusion criteria for the UMN PASC corpus mirror those of the UMN COVID corpus, with the following added criteria: patients must have (1) no baseline symptoms, (2) presented within a timeframe exceeding 30 days from their initial COVID-19 infection, and (3) at least one new or residual symptom [17]. Annotation guidelines instructed annotators to identify symptoms and their negations within the HPI section and other relevant sections. Clinical notes were randomly sampled from encounters at least 60 days after diagnosis with COVID-19. Demographics for this corpus are presented in Table 2.

N3C COVID. This corpus is a fully de-identified, manually annotated set of 148 unstructured notes across a selection of 5 inpatients and 5 outpatients with research authorizations as described by [16]. This corpus's inclusion criteria were: (1) notes documented within two weeks before and four weeks after the lab order date of the first positive COVID-19 result, and (2) notes that contained more than 1000 characters. Up to 20 notes from each patient were included in this corpus [21,22].

 doi:10.20944/preprints202409.1518.v

Table 2. Demographics of various corpora.

Corpus	Number of Patients	Median age in years, IQR (Q1, Q3)	Male %	Racial Distribution & Ethnicity %
UMN PASC	476	53.10, (38.00, 67.90)	58.00 %	5.00% Asian 11.90% Black 3.36% Hispanic 15.74% Others* 64.00% White
UMN COVID	46	54.68, (39.00, 66.64)	54.00 %	9.09% Asian 27.27% Black 4.54% Hispanic 7.10% Others* 52.00% White
N3C COVID**	148	Not Available (NA)	NA	NA

*Others include Declined races. **N3C COVID dataset is de-identified.

2.2. Prevalence Analysis

We examined how document format (PASC) might influence training a model by analyzing symptoms as keywords. BioMedICUS' annotations were converted to a normalized format using PASCLex to allow for keyword comparisons between corpora used in this study. Unlike a previous study by [6] that simply ranked terms by how often they appeared, this study used the Term Frequency-Inverse Document Frequency (TF-IDF) statistic to analyze how terms (keywords) are distributed across documents within each corpus. This approach provides a more nuanced picture than just counting total occurrences of a word in a corpus. A keyword's frequency (TF) measures how often a word appears in a document, while a term's inverse document frequency (IDF) measures how important a keyword is across all documents in the corpus. The product of TF and IDF thus helps identify keywords that are more meaningful or relevant to specific documents across the corpus. To compare keyword usage between corpora we performed an analysis of variance (ANOVA test) followed by ` [23,24]. Finally, we calculated effect size (η^2) to assess the magnitude of these differences [24]. Table A2 shows the top 10 extracted keywords ranked by TF-IDF for normalized symptoms mapped to PASCLex.

2.3. Dataset Curation for Model Fine-Tuning and Validation

We developed our model using clinical notes from 522 patients: 46 with acute COVID-19 and 476 with PASC. The median age of the patients was 53.89 years, with 55% male. We divided the PASC corpus into three sets: 65% for training, 4% for validation, and 31% for hold-out testing. All COVID-19 patient notes were used in the training set because they were annotated by a single SME without a consensus process. To test the model's generalizability, we included all patients from the N3C COVID corpus in our hold-out test set. This resulted in 351 patients in the training set, 20 in the validation set, and 299 in the hold-out test set. The demographic characteristics of each of these distinct corpora are delineated in Table 3.

 doi:10.20944/preprints202409.1518.v

Table 3. Breakdown of distinct corpora for LLM fine-tuning, validation, and hold-out testing.

Corpora	Number of Patients	Median age in years, IQR (Q1, Q3)	Male %	Racial Distribution & Ethnicity %
Training	351	53.89, (38.00, 67.00)	55.00 %	7.00% Asian 19.50% Black 3.90% Hispanic 11.60% Others* 58.00% White
Validation	20	53.50, (38.00, 67.00)	56.00 %	6.00% Asian 18.00% Black 4.00% Hispanic 12.00% Others* 60.00% White
Hold-out Test**	299	53.00, (38.00, 67.00)	56.00 %	6.5% Asian 20% Black 4.00% Hispanic 11.5% Others* 58% White

*Others include Declined races. **The demographics of the test set include information about UMN PASC patients only, since the N3C COVID data are de-identified.

2.4. Model Validation

Statistical Measures. Metrics including recall, precision, and F1-score were calculated to determine the models's performance in applying extracted symptoms for labeling of clinical notes.

Fairness Analysis. To assess the LLM's and BioMedICUS' equity across race and gender, their performance was evaluated on the UMN PASC hold-out test set for which data on race or ethnicity and gender were available.

Error Analysis. To determine sources of false negatives (FN) for BioMedICUS and the LLM used in this study, we examined a few symptoms in which the false negative rate (FNR) was: (1) high for the LLM, but not BioMedICUS (**fever** for N3C corpus); (2) high for BioMedICUS, but not for the LLM (**diarrhea** for UMN PASC corpus); and (3) both high for BioMedICUS and LLM (**chest pain** for N3C corpus).

3. Results

3.1. Prevalence Analysis

Corpora Keywords Ranked by TF-IDF. Cohorts in each of the corpora used in this study had top 10 extracted keywords ranked by TF-IDF for normalized symptoms mapped to PASCLex as summarized in Table A1.

Hypothesis Testing/Keyword Analysis. We observed a statistically significant difference in the logarithm of TF-IDF values based on symptoms ($F(16, 158) = 6.48$, $p < 0.0001$). Furthermore, there was a statistically significant difference in the logarithm of TF-IDF values based on corpus ($F(2, 158) = 79.17$, $p < 0.0001$). Results of post hoc testing indicated significant differences between UMN COVID and UMN PASC ($p < 0.0001$) as well as UMN PASC and N3C COVID ($p < 0.0001$). However, there was no difference between UMN COVID and N3C COVID. Furthermore there were large effects for both corpus, $\eta^2 = 0.46$ and symptoms, $\eta^2 = 0.4$.

3.2. Performance Evaluation

To validate the LLM's performance and compare the results with BioMedICUS, we used the hold-out test set to evaluate extraction performance. Table 4 symptom-wise weighted macro-average

 doi:10.20944/preprints202409.1518.v

scores of performance metrics (as described in Section 2.4) calculated over UMN PASC and N3C COVID symptoms. Figure 1 and Figure 2 shows a graphical comparison of extraction performance of BioMedICUS and LLM on the hold-out test set of manually annotated reference over positive mentions for UMN PASC and N3C COVID respectively.

Table 4. Extraction performance of BioMedICUS and LLM on the hold-out test set of manually annotated reference over positive (+) and negative (-) symptom mentions for UMN PASC and N3C COVID. The values represent recall, precision and F1-score in that order (noted as r, p, f1). The last row shows the average scores across all symptoms.

Symptom	Model	UMN PASC (+) r*, p*, f1*	UMN PASC (-) r, p, f1	N3C COVID (+) r, p, f1	N3C COVID (-) r, p, f1
bone or joint pain	BioMedICUS	0.57, 0.51, 0.32	0.64, 0.58, 0.46	0.81, 0.54, 0.46	0.56, 0.50, 0.47
	LLM	0.66, 0.56, 0.56	0.53, 0.54, 0.53	0.68,0.53,0.50	0.45, 0.48, 0.47
mental status changes	BioMedICUS	0.82, 0.64, 0.68	0.86, 0.62, 0.66	0.90, 0.83, 0.86	0.65, 0.59, 0.61
	LLM	0.54, 0.57, 0.55	0.48, 0.48, 0.48	0.53, 0.53,0.53	0.48, 0.48, 0.48
chest pain or palpitations	BioMedICUS	0.91, 0.74, 0.79	0.95, 0.91, 0.92	0.70, 0.70, 0.70	0.83, 0.71, 0.76
	LLM	0.72, 0.74, 0.73	0.62, 0.77, 0.64	0.50, 0.50, 0.50	0.46, 0.47, 0.47
cough	BioMedICUS	0.87, 0.79, 0.82	0.81, 0.73, 0.74	0.92, 0.90, 0.91	0.65, 0.68, 0.66
	LLM	0.63, 0.67, 0.65	0.66, 0.69, 0.67	0.53, 0.53, 0.53	0.44, 0.47, 0.45
diarrhea	BioMedICUS	0.76 ,0.73, 0.74	0.89, 0.86, 0.87	0.89, 0.91, 0.90	0.87, 0.83, 0.85
	LLM	0.85, 0.68, 0.73	0.63, 0.64, 0.64	0.58, 0.55, 0.55	0.59, 0.55, 0.56
shortness of breath	BioMedICUS	0.85, 0.71, 0.74	0.86, 0.82, 0.83	0.85, 0.78, 0.79	0.72, 0.69, 0.71
	LLM	0.75, 0.70, 0.72	0.61, 0.65, 0.62	0.48, 0.48, 0.48	0.50, 0.50, 0.50
fatigue	BioMedICUS	0.86, 0.75, 0.78	0.81, 0.63, 0.67	0.84, 0.80, 0.82	0.98, 0.62, 0.69
	LLM	0.75, 0.70, 0.72	0.67, 0.59, 0.61	0.64, 0.58, 0.59	0.47, 0.49, 0.48
fever	BioMedICUS	0.82, 0.61, 0.61	0.85, 0.86, 0.86	0.84, 0.85, 0.85	0.88, 0.91, 0.89
	LLM	0.68, 0.57, 0.58	0.61, 0.62, 0.61	0.62, 0.58, 0.59	0.50, 0.50, 0.49
headaches	BioMedICUS	0.84, 0.71, 0.75	0.91, 0.81, 0.85	0.77, 0.82, 0.79	0.84, 0.80, 0.82
	LLM	0.72, 0.72, 0.72	0.66, 0.66, 0.66	0.65, 0.55, 0.56	0.45, 0.47, 0.46
changes in sleep	BioMedICUS	0.94, 0.63, 0.68	0.48, 0.49, 0.48	0.99, 0.66, 0.74	0.75, 0.99, 0.83
	LLM	0.74, 0.86, 0.79	0.50, 0.50, 0.50	0.50, 0.50, 0.50	0.50, 0.49, 0.49
nausea/vomiting	BioMedICUS	0.93, 0.71, 0.76	0.91, 0.89, 0.90	0.79, 0.82, 0.81	0.78, 0.77, 0.77
	LLM	0.72, 0.62, 0.65	0.67, 0.68, 0.68	0.66, 0.62, 0.63	0.56, 0.58, 0.57
nasal congestion and obstruction	BioMedICUS	0.80, 0.65, 0.69	0.86, 0.88, 0.87	0.49, 0.49, 0.49	0.99, 0.75, 0.83
	LLM	0.62, 0.62, 0.66	0.63, 0.68, 0.65	0.43, 0.49, 0.45	0.47, 0.49, 0.48
sore throat	BioMedICUS	0.92, 0.92, 0.92	0.90, 0.84, 0.87	0.99, 0.85, 0.91	0.75, 0.99, 0.82
	LLM	0.60, 0.55, 0.57	0.60, 0.63, 0.61	0.64, 0.54, 0.54	0.47, 0.48, 0.47
loss of smell and taste	BioMedICUS	0.79, 0.73, 0.76	0.65, 0.65, 0.65	0.74, 0.81, 0.77	0.65, 0.65, 0.65
	LLM	0.66, 0.97, 0.71	0.66, 0.74, 0.70	0.48, 0.48, 0.48	0.49, 0.48, 0.49
	BioMedICUS	**0.81, 0.69, 0.70**	**0.79, 0.75, 0.74**	**0.82, 0.76, 0.77**	**0.77, 0.74, 0.74**
	LLM	0.68, 0.68, 0.68	0.61, 0.63, 0.62	0.56, 0.53, 0.54	0.48, 0.50, 0.49

* r is recall, p is precision, and f1 is f1-score.

As seen in Table 4, Overall, BioMedICUS outperforms LLM across most symptoms and datasets. However, LLM shows superior f1-scores (on UMN PASC + and -) in specific cases, such as **changes in sleep** and **bone or joint pain**. Despite this, there are instances where both models perform similarly. For example, they achieve nearly identical F1-scores (on UMN PASC +) for **diarrhea, shortness of breath**, **headaches, nasal congestion.** These examples indicate that while BioMedICUS has a broader consistent advantage, LLMs can be more effective for certain symptoms and performs comparably to BioMedICUS in specific scenarios, demonstrating the robustness of both models in extracting clinical symptoms across different datasets.

 doi:10.20944/preprints202409.1518.v

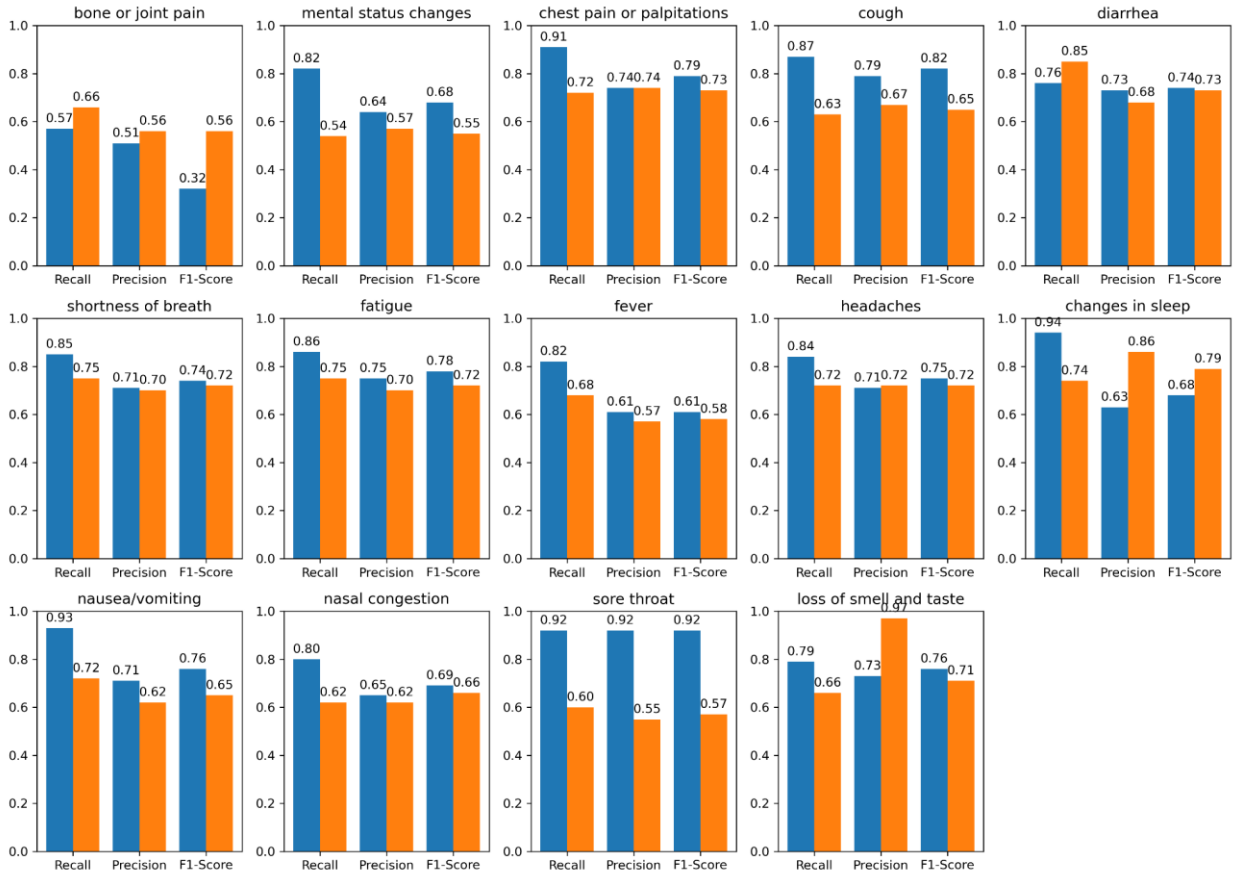

Figure 1. Extraction performance of BioMedICUS and LLM on the hold-out test set of manually annotated reference over positive mentions for UMN PASC.

 doi:10.20944/preprints202409.1518.v

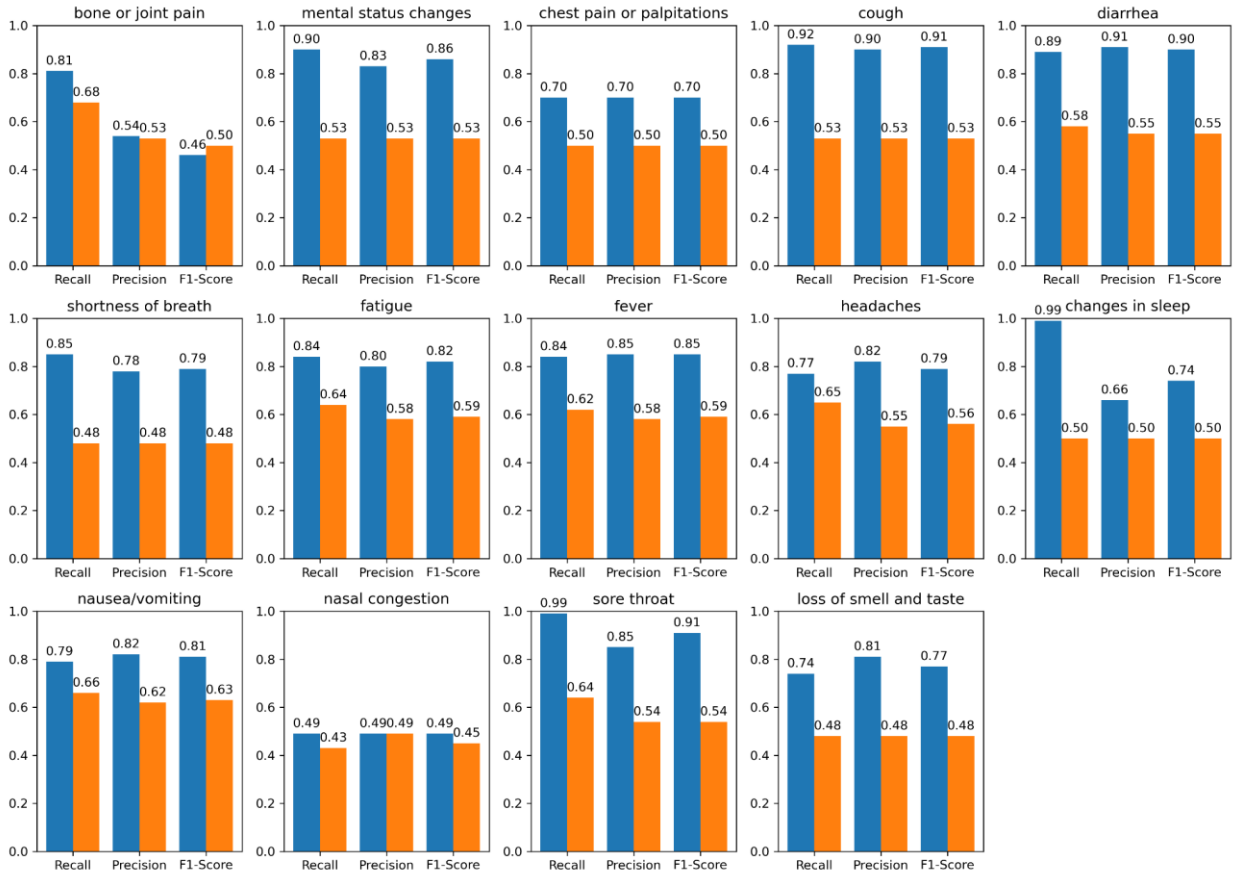

Figure 2. Extraction performance of BioMedICUS and LLM on the hold-out test set of manually annotated reference over positive mentions for N3C COVID.

Fairness analysis. For gender, both models perform slightly better for males, with recall, precision, and F1-scores higher than those for females in both positive (symptom presence) and negative (symptom negation) symptom mentions. In terms of race, both models show the highest performance for the 'White' and 'Other' categories in positive mentions. However, for the LLM, negative symptom mentions demonstrate reduced performance across all racial groups, particularly for those of Asian descent. Table 5 shows the results of LLM's and BioMedICUS' equity in the UMN PASC corpus.

Table 5. Metrics for evaluation of the LLMs' and BioMedICUS' equity for race and gender in positive (+) and negative (-) symptom mentions for UMN PASC.

Model	Category	r*, p*, f1* (+)	r, p , f1 (-)
LLM	Male	**0.69, 0.71, 0.70**	**0.62, 0.65, 0.63**
	Female	0.67, 0.66, 0.66	0.61, 0.62, 0.63
	Asian	0.64, 0.65, 0.64	0.58, 0.62, 0.60
	Black	0.67, 0.66, 0.66	0.60, 0.64, 0.61
	Other**	0.70, 0.70, 0.70	0.62, 0.65, 0.63
	White	**0.70, 0.72, 0.71**	**0.63, 0.66, 0.64**
BioMedICUS	Male	**0.82, 0.71, 0.76**	**0.80, 0.76, 0.77**
	Female	0.80, 0.67, 0.73	0.78, 0.74, 0.76
	Asian	0.75, 0.65, 0.69	0.75, 0.70, 0.72
	Black	0.80, 0.68, 0.73	0.79, 0.74, 0.76
	Other**	0.84, 0.70, 0.76	0.81, 0.76, 0.78
	White	**0.85, 0.73, 0.78**	**0.83, 0.80, 0.81**

*r is recall, p is precision, and f1 is F1-score. **"Other" includes declined races.

 doi:10.20944/preprints202409.1518.v

Error Analysis. A manual audit of a small sampling of cases from each corpus with false negatives for positive mentions of "**fever**" (in the N3C COVID corpus) and "**diarrhea**" (in the UMN PASC corpus) are shown in Table 6. BioMedICUS had a false negative rate of 0.28, 0.44 and 0.55 respectively for "**fever**," "**diarrhea**" and "**chest pain**" while the LLM had a FNR of 0.54, 0.22 and 0.88 for "**fever**," "**diarrhea**" and "**chest pain**", respectively.

Table 6. Representative examples of false negatives for positive mentions of "**fever**" in N3C COVID corpus, "**diarrhea**" in UMN PASC corpus and "**chest pain**" in N3C corpus as returned by BioMedICUS and LLM along with explanations.

Corpus/ Symptom	Ground Truth	BioMedICUS	LLM	Issues
N3C COVID/Fever	1	0	0	both UMLS and LLM did not pick up 38 °C' following heading of 'Temp' for positive mention; other similar
	1	1	0	contextual : "had been sick with fever... over 12 days prior to his admission.", etc.
	1	0	1	UMLS did not pick up "increase in temperature" for positive mention; "highest temp 37. 5," etc.
UMN PASC/Diarrhea	1	0	0	classified as negation: "Mild diarrhea today, kaopectate resolved"
	1	1	0	none
	1	0	1	improper negation: "Constipation - senna not doing well, still either rabbit turds or diarrhea"
N3C COVID/Chest Pain	1	0	0	missed lexical synonym: "Persistent pain or pressure in the chest," etc.
	1	1	0	missed case of both positive and negative mention due to "possible" symptom
	1	0	1	none

4. Discussion

This study aimed to fine-tune an LLM for extraction of patient symptoms from unstructured clinical notes and demonstrate its comparable effectiveness to rule-based IE methods. We found that the LLM-based approach not only matches but in some cases exceeds the performance of rule-based methods. This study had 3 key findings: (1) BioMedICUS showed higher recall for certain symptoms, while the LLM excelled in precision and f1-scores for others, while both had similar performance for some symptoms; (2) Manual audits revealed higher FN for BioMedICUS and higher false positives (FP) for the LLM, often due to misinterpreting context or section headers; (3) The LLM had generalizability and bias issues due to overfitting from training on a non-diverse dataset and single annotator bias, highlighting the need for diverse, high-quality datasets and robust annotations.

4.1. Prevalence Analysis

As shown in Table A1, the ranking of top symptoms using TF-IDF varies across corpora, except between UMN COVID and N3C COVID. For example, the symptom "**fever**" is ranked fifth in the UMN PASC corpus, but it ranks first and second in the UMN COVID and N3C COVID corpora, respectively. Additionally, the number of times "**fever**" appears in these corpora differs even more. Symptom mentions of "**diarrhea**" do not appear among top rankings of UMN PASC by TF-IDF. Lastly, "**chest pain**" does not appear among top ranked symptoms in any of these corpora. Results of the ANOVA test confirm that both structure and content of these corpora significantly influence the TF-IDF scores. This is likely because the corpora are organized differently or contain different information. Moreover, the large effect sizes ($\eta^2 = 0.4$ for corpora and $\eta^2 = 0.42$ for symptoms) show that both the corpus format and the presence or absence of symptoms have a substantial influence on the TF-IDF score distribution across the corpora.

 doi:10.20944/preprints202409.1518.v

4.2. Error Analysis

As shown in Table 6, for the N3C COVID corpus, both BioMedICUS and the LLM missed various terms for "**fever**" not present in either the UMLS or the LLM's lexicon. Additionally, the LLM overlooked contextual sentences ending with "**had been sick**". However, while the UMLS lexicon missed terms like "**increase in temperature**", the LLM detected them, demonstrating its sensitivity to such phrases due to its pre-training on general English. For "**diarrhea**", both BioMedICUS and the LLM incorrectly negated sentences containing "**diarrhea**" that ended with "**resolved**" missing the temporality of these mentions. However, the LLM correctly contextualized a negating phrase unrelated to having "**diarrhea**", whereas BioMedICUS did not. For "**chest pain**", both classifiers missed terms not found in the UMLS or the LLM's lexicon, and the LLM failed to classify terms with both positive and negative mentions of "**chest pain**" accurately.

For both the LLM and BioMedICUS various symptoms had been mislabeled as FP due to neither using section header detection to filter out irrelevant mentions from notes. On the other hand, certain templated sections within clinical notes were correctly identified by LLM but not by BioMedICUS. Consequently, we noted a significant difference between RTF structured and unstructured versions of the UMN PASC corpus using BioMedICUS' rule-based selection header detection: In RTF notes, symptoms under "HPI" were identified 785 times compared to 424 times in unstructured notes, suggesting that RTF's structure aids in distinguishing relevant section headers, with potential for reducing FP.

4.3. Performance Evaluation

Comparing the generalizability of the fine-tuned LLM and BioMedICUS for extracting symptoms from clinical notes reveals key differences. BioMedICUS, which uses techniques like "normalized bag of words" and UMLS, shows high generalizability due to its universal rules not limited by specific corpora, effectively identifying symptoms across diverse datasets. In contrast, the LLM's performance is closely tied to its training data, mostly from UMN PASC patients, with only 12% from UMN COVID patients. Consequently, the LLM performed similarly to BioMedICUS on the UMN PASC hold-out test set but underperformed on the acute N3C COVID test set, highlighting issues of overfitting and inadequate acute COVID-19 patient representation in the training data. Additionally, the LLM's UMN COVID training labels, annotated by a single annotator without inter-annotator agreement, likely reduced its performance. These findings suggest that while LLMs offer flexibility in capturing nuanced symptoms, their generalizability can be significantly improved with well-annotated datasets, regularization, and robust prompt engineering.

We found that for positive mentions of UMN PASC symptoms, choosing between LLM and BioMedICUS depends on priorities: BioMedICUS prioritizes maximizing identification with higher recall, while the LLM prioritizes minimizing FP with better precision. For negative mentions, while BioMedICUS appears more robust across a broad range of symptoms, LLM's strengths in specific areas suggest its value for targeted symptom identification. For N3C COVID symptoms, BioMedICUS consistently outperformed the LLM across most positive and negative symptom mentions, indicating its ability to generalize knowledge to new data.

Our results align with experiments performed by Patra, *et al*. who compared the performance of rule-based systems with LLMs for extracting information from clinical psychiatry notes. They found that the rule-based systems outperformed the LLMs across most of their cases [25]. Our findings are further supported by Chen, *et al.*, who compared LLMs like GPT-3.5 and GPT-4 [26] against fine-tuned BERT and BART [27] models for biomedical natural language processing applications. They found that GPT-4's performance was competitive or better than the fine-tuned BERT and BART models in 6 out of 12 benchmarks especially for reasoning and generative tasks [28].

4.4. Limitations

The limitations of our study include: (1) Generalizability of the LLM: being trained on limited datasets, as presented in Table 4, the LLM lacks generalizability, furthermore, effect size of corpora

 doi:10.20944/preprints202409.1518.v

differences may limit model generalizability. (2) Lack of reliable section header detection: the inability to detect and leverage section headers hinders the performance of our models. (3) Length of notes: some notes might be too large to be processed by our fine-tuned LLM. In such cases, the notes have to be processed in chunks, which slows down the process. (4) Lexica used in this study may be incomplete thus leading to increase in FN across some symptoms.

4.5. Future Work

Future work includes: (1) Expanding LLM generalizability: expanding the diversity of the training dataset, applying robust regularization techniques, involving multiple annotators for high-quality labels, using domain adaptation techniques, and employing hybrid approaches to enhance model output. (2) Training models on desired section headers from the notes: use of RTF tags to identify section headers like HPI, has potential to enhance model performance and thus resolve the issue of some notes being too large for the LLM model to process. (3) Training models on structured notes: RTF structured notes offer formatting features with tags, including text styling, tables, and hyperlinks, aiding in symptom identification and severity assessment. (4) Comparison with other fine-tuned LLMs (5) Exploring methods that combine LLMs and rule-based systems. (6) Detecting the severity of symptoms: future research aims to integrate severity detection for patient symptoms allowing clinicians to track changes in symptom severity over time, aiding in treatment adjustments with potential to improve patient outcomes.

5. Conclusions

This study highlights the significant issues of model bias and generalizability in our fine-tuned LLM due to demographic variability and corpus differences. Our evaluation revealed that the LLM, trained predominantly on UMN PASC patients with limited acute COVID-19 representation, showed overfitting and insufficient generalization, potentially exacerbated by single annotator bias. These findings provide real-world evidence of the challenges LLMs may face, underscoring the necessity for diverse, high-quality training datasets and robust annotation processes to mitigate bias and enhance generalizability. Our findings also indicated that the rule-based approach and the LLM approach each have strengths and limitations in performing IE tasks applied to clinical NLP.

Acknowledgments. This work was partially supported by the NIH's National Center for Complementary & Integrative Health under grant number U54AT012307. The authors would like to acknowledge support from the Center for Learning Health System Sciences, a partnership between the Medical School and School of Public Health at the University of Minnesota. The authors are grateful to Angela M. Bailey, MD, MS, and Molly Diethelm, PMP, for their valuable feedback, which significantly improved this paper.

Appendix A

Prevalence Analysis

Table A1. Top 10 normalized symptoms ranked by TF-IDF with prevalence within each corpus noted*.

UMN PASC	UMN COVID	N3C COVID
Pain (0.131)	Fever (0.119)	Pain (0.090)
Shortness of Breath (0.079)	Shortness of Breath (0.118)	Fever (0.059)
Cough (0.077)	Cough (0.100)	Shortness of Breath (0.084)
Myalgia (0.053)	Myalgia (0.053)	Nausea or Vomiting (0.064)
Fever (0.049)	Nausea or Vomiting (0.058)	Cough (0.053)
Rash (0.029)	Rash (0.009)	Diarrhea (0.040)
Anxiety (0.025)	Fatigue (0.041)	Abdominal Pain (0.039)
Nausea or Vomiting (0.054)	Abdominal Pain (0.014)	Respiratory Depression (0.037)
Fatigue (.060)	Diarrhea (0.031)	Fatigue (0.034)

 doi:10.20944/preprints202409.1518.v

Headaches (0.052)	Headaches (0.023)	Mental Status Change (0.031)

*The number in parentheses is the prevalence normalized with respect to the number of symptoms.

Appendix B

BioMedICUS/Microservice Text Analysis Platform (MTAP) Concept Extraction

The BioMedICUS/MTAP pipeline incorporates: (1) Sentence boundary detection using a pre-trained bi-LSTM model [7] (2) Rule-based matching for section header detection [29] (3) Concept labeling using various matching methods, including normalized bag-of-words matching against UMLS terms [30], and (4) Negation detection using NegEx [31].

Appendix C

Lexica

This study utilized four lexica: (1) A lexicon of acute COVID-19 symptoms as discussed in Silvermal, *et al.* based on guidelines provided by the Centers for Disease Control and Prevention (CDC) [2,32]; (2) A lexicon of PASC symptoms based on CDC (2022) guidelines and [33]; (3) A lexicon of acute COVID-19 symptoms developed by the N3C consortium [21]; and (4) PASCLex developed by [6]. These lexica were normalized using PASCLex to allow for mappings between each corpus (please see https://shorturl.at/iqBC3 for normalized mappings between lexica used in this study).

Appendix D

LLM Model Development

Model Used. In order to extract patient symptoms from clinical notes, we fine-tuned the LLaMA2-13b-chat model [34], for a generative QA task on our manually annotated corpora. To create such a QA dataset, an example QA prompt was used as shown in Table A2. The model outputs a list of positive and negative mentions of symptoms found within each clinical note. The LLaMA2 model was selected at the time of this study, because it was trained on two trillion tokens of data [34].

Table A2. Example QA prompt for symptom extraction. "note_text" is the text from the patient's clinical note. "SymptomA/B" are the ground truth positive mentions of the patient's symptoms. "SymptomP/Q" are the ground truth negative mentions.

Question	Answer
"What are the positive and negative symptoms of the patient given the following clinical text:\n Clinical Text: note_text\n"	"Positive symptoms are [symptomA, symptomB]. Negative symptoms are [symptomP, symptomQ]."

Fine-tuning Strategy. We fine-tuned the LLM model on two NVIDIA A100 GPUs each with a memory of 80GB. We loaded a base LLaMA2 model using the PyTorch library [35] and fine-tuned it on our training set. To reduce VRAM usage, we used 4-bit precision parameter-efficient fine-tuning (PEFT) techniques including Quantized Low-Rank Adaptation of LLMs (QLoRA) [36]. We chose QLoRA because it is an efficient fine-tuning approach that reduces memory usage enough to fine-tune a 65 billion parameter model on a single 48GB GPU while preserving full 16-bit fine-tuning task performance. Our model was trained directly using a 4-bit precision format (NF4) for 50 epochs. A batch size of four and a gradient clipping with value 0.3 was used. The initial learning rate was set to 2e-4 with a polynomial decay schedule with warm up. A weight decay of 1e-3 was used. We used AdamW [37] optimizer with its default settings. The selection of these hyperparameters was based on the model's performance on the validation set during fine-tuning.

References

1. Pakhomov S, Finley GP, McEwan R, *et al.* Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*. 2016;32 23:3635–44.
2. Silverman GM, Sahoo HS, Ingraham NE, *et al.* NLP Methods for Extraction of Symptoms from Unstructured Data for Use in Prognostic COVID-19 Analytic Models. *Journal of Artificial Intelligence Research*. 2021;72:429–74.
3. National Center for Biotechnology Information. *UMLS® Reference Manual*. National Library of Medicine (US) 2009. https://www.ncbi.nlm.nih.gov/books/NBK9676/ (accessed 5 October 2021)
4. Bodenreider O. The UMLS Semantic Network. The UMLS Semantic Network. 2020. https://semanticnetwork.nlm.nih.gov (accessed 3 February 2020)
5. He Z, Perl Y, Elhanan G, *et al.* Auditing the assignments of top-level semantic types in the UMLS semantic network to UMLS concepts. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2017;1262–9.
6. Wang L, Foer D, MacPhaul E, *et al.* PASCLex: A comprehensive post-acute sequelae of COVID-19 (PASC) symptom lexicon derived from electronic health record clinical notes. *J Biomed Inform*. 2022;125:103951.
7. Knoll BC, Lindemann EA, Albert AL, *et al.* Recurrent Deep Network Models for Clinical NLP Tasks: Use Case with Sentence Boundary Disambiguation. *Stud Health Technol Inform*. 2019;264:198–202.
8. Knoll BC, McEwan R, Finzel R, *et al.* MTAP - A Distributed Framework for NLP Pipelines. *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. Rochester, MN, USA: IEEE 2022:537–8. https://doi.org/10.1109/ICHI54592.2022.00102
9. Knoll BC, Gunderson M, Rajamani G, *et al.* Advanced Care Planning Content Encoding with Natural Language Processing. *Stud Health Technol Inform*. 2024;310:609–13.
10. Xu D, Chen W, Peng W, *et al.* Large Language Models for Generative Information Extraction: A Survey. Published Online First: 2023. doi: 10.48550/ARXIV.2312.17617
11. Yang Z, Dai Z, Yang Y, *et al.* XLNet: Generalized Autoregressive Pretraining for Language Understanding. Published Online First: 2019. doi: 10.48550/ARXIV.1906.08237
12. Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. Published Online First: 2020. doi: 10.48550/ARXIV.2004.05150
13. Hoogendoorn M, Szolovits P, Moons L, *et al.* Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. *Artificial intelligence in medicine*. 2016;69:53–61.
14. Stephens KA, Au MA, Yetisgen M, *et al.* Leveraging UMLS-driven NLP to enhance identification of influenza predictors derived from electronic medical record data. Bioinformatics 2020. https://doi.org/10.1101/2020.04.24.058982
15. Skube SJ, Hu Z, Simon GJ, *et al.* Accelerating Surgical Site Infection Abstraction With a Semi-automated Machine-learning Approach. *Ann Surg*. Published Online First: 14 October 2020. doi: 10.1097/SLA.0000000000004354
16. Sahoo HS, Silverman GM, Ingraham NE, *et al.* A fast, resource efficient, and reliable rule-based system for COVID-19 symptom identification. *JAMIA Open*. 2021;4:ooab070.
17. Silverman GM, Rajamani G, Ingraham NE, *et al.* A Symptom-Based Natural Language Processing Surveillance Pipeline for Post-COVID-19 Patients. In: Bichel-Findlay J, Otero P, Scott P, *et al.*, eds. *Studies in Health Technology and Informatics*. IOS Press 2024. https://doi.org/10.3233/SHTI231087
18. Devlin J, Chang M-W, Lee K, *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*. 2019.
19. Jahan I, Laskar MTR, Peng C, *et al.* A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*. 2024;171:108189.
20. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 1971;76:378–82.
21. He Y, Yu H, Ong E, *et al.* CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Sci Data*. 2020;7:181.
22. N3C. OHNLP/N3C-NLP-Documentation Wiki. 2022. https://github.com/OHNLP/N3C-NLP-Documentation/wiki (accessed 30 April 2024)
23. Nagamine T, Gillette B, Kahoun J, *et al.* Data-driven identification of heart failure disease states and progression pathways using electronic health records. *Sci Rep*. 2022;12:17871.
24. Pojanapunya P, Watson Todd R. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*. 2018;14:133–67.
25. Patra BG, Lepow LA, Kumar PKRJ, *et al.* Extracting Social Support and Social Isolation Information from Clinical Psychiatry Notes: Comparing a Rule-based NLP System and a Large Language Model. 2024. https://doi.org/10.48550/ARXIV.2403.17199
26. OpenAI, Achiam J, Adler S, *et al.* GPT-4 Technical Report. 2023. https://doi.org/10.48550/ARXIV.2303.08774
27. Lewis M, Liu Y, Goyal N, *et al.* BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2019. https://doi.org/10.48550/ARXIV.1910.13461

28. Chen Q, Du J, Hu Y, *et al.* Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. 2023. https://doi.org/10.48550/ARXIV.2305.16326
29. University of Minnesota NLP/IE Program. BioMedICUS Section Header Detector. GitHub. 2024. https://github.com/nlpie/biomedicus/blob/main/java/src/main/java/edu/umn/biomedicus/sections/RuleBasedSectionHeaderDetector.java (accessed 23 April 2024)
30. University of Minnesota NLP/IE Program. BioMedICUS UMLS Concept Detection Algorithm. GitHub. 2024. https://github.com/nlpie/biomedicus/tree/main/java/src/main/java/edu/umn/biomedicus/concepts (accessed 23 April 2024)
31. Chapman WW, Bridewell W, Hanbury P, *et al.* A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*. 2001;34:301–10.
32. CDC. Symptoms of Coronavirus. Centers for Disease Control and Prevention. 2020. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html (accessed 12 December 2020)
33. Davis HE, Assaf GS, McCorkell L, *et al.* Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine*. 2021;38:101019.
34. Touvron H, Martin L, Stone K, *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. Published Online First: 2023. doi: 10.48550/ARXIV.2307.09288
35. Paszke A, Gross S, Massa F, *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Published Online First: 2019. doi: 10.48550/ARXIV.1912.01703
36. Dettmers T, Pagnoni A, Holtzman A, *et al.* QLoRA: Efficient Finetuning of Quantized LLMs. Published Online First: 2023. doi: 10.48550/ARXIV.2305.14314
37. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. Published Online First: 2017. doi: 10.48550/ARXIV.1711.05101