

RECOGNIZING HEALTH CONCEPTS IN TWITTER DATA USING LARGE LANGUAGE MODEL'S

by

Soniya Sagar Chavan

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science



College of Technology

Hammond, Indiana

May 2025

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Keyuan Jiang, Chair

College of Technology

Dr. Ashok Vardhan Raja

College of Technology

Dr. George Stefanek

College of Technology

Approved by:

Dr. Ge Jin

*To my incredible family,
my parents, Sagar Rajaram Chavan and Sarita Sagar Chavan, and my brother, Sattyam Sagar
Chavan — whose strength, love, and belief in me made everything possible.*

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Keyuan Jiang, whose unwavering support and invaluable guidance have shaped my academic journey since my very first semester. His mentorship, encouragement, and insightful feedback have played a crucial role in my learning and growth. Throughout the course of this thesis, he has not only helped me refine my research but also pushed me to improve at every step, assisting me in revising drafts and strengthening my understanding. I am truly grateful for his patience, expertise, and dedication, which have been instrumental in bringing this work to fruition.

I also extend my sincere appreciation to Dr. Ashok Vardhan Raja for his continuous support and guidance during my master's studies. His willingness to assist me whenever I sought advice has been incredibly valuable. Additionally, I am grateful to Dr. George Stefanek for being a part of my thesis journey and providing valuable insights that contributed to my research.

I would also like to acknowledge Purdue University Northwest and everyone who has supported me throughout my master's studies. The faculty, staff, and peers have made this journey enriching and have helped me grow both academically and personally.

Last but certainly not least, I am deeply thankful to my family for their unwavering support from day one to this very moment. Their encouragement, belief in me, and sacrifices have been my greatest source of strength, and I could not have accomplished this milestone without them.

Thank you all.

TABLE OF CONTENTS

LIST OF TABLES	8
LIST OF FIGURES	9
ABSTRACT	10
DEFINITIONS	11
1. INTRODUCTION	15
1.1 Background	15
1.2 Introduction	16
1.3 Why LLMs and Prompts?	18
2. STATEMENT OF PROBLEM	20
3. RESEARCH QUESTION	22
3.1 Research Aim	22
3.2 Research Questions	22
4. SIGNIFICANCE OF THE STUDY	23
4.1 Practical and Public Health Significance	23
4.2 Academic Significance	23
4.3 Technological Relevance	24
5. SCOPE OF THE STUDY	25
6. ASSUMPTIONS	26
7. LIMITATIONS	27
8. DELIMITATIONS	28
9. LITERATURE REVIEW	29
9.1 Health Concept Extraction	29
9.2 Methods of Extraction	30
9.3 Characteristics of Social Media Data	31
9.4 Health Discussions on Social Media (Twitter/X and Other Platforms)	32
9.5 Challenges in Processing Social Media Data for Health Concept Extraction	34
9.6 Solutions to Overcome Social Media Data Challenges	35
9.7 Health Concept Extraction with Large Language Models (LLMs)	37
9.8 Normalization of Health Concepts Using Large Language Models (LLMs)	39

9.9	Named Entity Recognition (NER) in Health Concept Extraction Using LLMs.....	40
9.10	Non-Deterministic Nature of Large Language Models (LLMs)	42
9.11	Analysis and Synthesis	43
10.	METHOD	47
10.1	Overview of the Research Workflow	47
10.2	Data Collection and Preprocessing.....	50
10.3	Ground Truth Annotation	52
10.4	Baseline Model Using Schema-Guided Extraction	53
10.5	Prompt Engineering Strategy	54
10.5.1	Prompt Development and Deployment	55
10.5.2	Worked Example: Semantic Match Prompting.....	55
10.6	Multi-Run Execution and Consensus Selection	57
10.7	Performance Evaluation	59
10.8	Symptom Normalization	61
10.9	Symptom Categorization	62
10.10	Data Visualization and Analysis	63
10.11	Implementation Notes and Runtime Considerations	64
10.12	Summary	66
11.	RESULTS	68
11.1	Model Performance Overview	68
11.2	Symptom Normalization Insights.....	73
11.3	Categorization Distribution	74
11.4	Multidimensional Visualization – 3D Data cube	79
12.	DISCUSSIONS.....	81
12.1	Addressing Research Questions	81
12.2	Reassessing Limitations in Context.....	82
12.3	Implications for Public Health Informatics and LLM Applications	82
12.4	Opportunities for Future Research	83
13.	CONCLUSION.....	85
	APPENDIX A. PROMPT TEMPLATES.....	87
	APPENDIX B. FULL SYMPTOM NORMALIZATION MAPPINGS	89

APPENDIX C. 3D CUBE SLICING.....	94
REFERENCES	102

LIST OF TABLES

Table 1. Literature Review Summary	46
Table 2. Use of Prompt-Based Instructions Across Tasks and Models	54
Table 3. Semantic Match Results for a Single Tweet Across Models	56
Table 4. Example of Multi-Run Output and Manual Consensus	58
Table 5. Confusion Matrix Components (TP, FP, FN)	68
Table 6. Evaluation Metrics	69
Table 7. Compact Comparison of Normalized Symptom Expressions Across Models	74
Table 8. Symptom Category Comparison	76

LIST OF FIGURES

Figure 1. Baseline and Experimental Extraction Pipeline	49
Figure 2. After Extraction Processing: Semantic Matching, Normalization, Categorization, and Visualization	50
Figure 3. Stacked bar chart showing Confusion Matrix components produced by each model during symptom extraction.	69
Figure 4. Accuracy Comparison Across Models	70
Figure 5. Precision Comparison Across Models.....	71
Figure 6. Recall Comparison Across Models	71
Figure 7. F1 Score Comparison Across Models	72
Figure 8. Combined Performance Comparison of All Models	72
Figure 9. Bar chart showing the total frequency of each symptom category after normalization. Respiratory and systemic symptoms were the most prevalent, consistent with the COVID-19. .	77
Figure 10. Heatmap of symptom category frequency by day (Day 1 to Day 14). Each cell represents the count of normalized symptoms within a category for that specific day.	78
Figure 11. 3D Symptom Cube	80

ABSTRACT

Social media users share valuable public health information, particularly discussions related to symptoms and health conditions. The Social media platforms reflect timely experiences of individuals during health events such as the COVID-19 pandemic. However, discovering meaningful insights from such informal, user-generated data presents challenges due to the presence of slang, abbreviations, emojis, various phrasing, misspellings, incomplete sentences and contextual ambiguity. Traditional Natural Language Processing (NLP) methods struggle with these complexities, limiting the quality and reliability of extracted information.

This thesis investigates a Large Language Model (LLM)-based framework for extracting, normalizing, and categorizing COVID-19 symptoms from 635 manually annotated tweets. Human annotation was used as the ground truth for evaluation. The study assessed the performance of three LLM setups—GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash—with GPT-4 Turbo, and Gemini 2.0 Flash executed three times with the same prompt and data. A consensus-based decision-making mechanism was used to drive the consistent output from multiple runs. All extracted outputs were then processed uniformly using Gemini 2.0 Flash for semantic match model, normalization model, and categorization model, ensuring consistency across models.

Performance Evaluation indicated that GPT-4-0613 via LangChain outperformed the others, achieving an accuracy of 0.78, precision of 0.84, recall of 0.91, and F1 score of 0.87. GPT-4 Turbo and Gemini 2.0 Flash followed with notably lower performance. Results were visualized using interactive multi-dimensional data cubes.

The results demonstrate that integrating LLMs with structured downstream processes—such as schema-based normalization and consensus-based output—can improve the quality and consistency of health information extraction from social media. This research provides a reproducible and scalable methodology for health-related text mining in noisy, real-world data. The proposed framework may be extended in future work for timely public health monitoring, multilingual symptom recognition, and broader application to other disease contexts.

DEFINITIONS

Concept Normalization

The process of ensuring that different ways of describing the same concept are mapped to a common term in a controlled vocabulary. For example, "burning up," "feeling hot," and "high temp" all refer to "fever" and are normalized accordingly in this study.

Consensus Output

The final output derived from multiple responses generated by the same LLM using an ensemble mechanism. It represents the most consistent result, selected through a voting mechanism. In this study, consensus outputs were used to reduce variability and improve consistency in symptom extraction from GPT-4 Turbo and Gemini 2.0 Flash.

COVID-19 Symptoms

Health problems linked to COVID-19, such as fever, cough, fatigue, or shortness of breath. This study focuses on tracking these symptoms from informal Twitter posts.

Data Cube

A data cube is a multi-dimensional data structure used for organizing and analyzing information across multiple axes, such as symptom, day, and category. In this study, data cubes were used to visualize extracted symptoms across time.

Data Preprocessing

The process of cleaning tweets before they are processed by downstream LLMs. This includes removing special characters, emojis, links, and non-health-related text to isolate symptom-related content.

Ensemble Mechanism

A method used to improve output consistency by running the same LLM multiple times with the same prompt and data tweet and selecting the most consistent output. This mitigates the non-deterministic behavior of the LLM outputs.

F1 Score

The F1 Score is a performance metric that balances Precision and Recall to assess the model's overall accuracy. It is calculated as the harmonic mean of Precision and Recall.

Generative Pre-trained Transformer 4 (GPT-4)

GPT-4 is a multimodal large language model developed by OpenAI. It is capable of understanding and generating human-like text across diverse tasks. In this study, two versions of GPT-4 were utilized—GPT-4- 0613 and GPT-4 Turbo—to extract symptom expressions from informal social media data.

Gemini 2.0 Flash

A state-of-the-art large language model developed by Google. In this study, Gemini 2.0 Flash was used consistently across all LLM outputs for semantic similarity matching, symptom normalization, and categorization to ensure uniformity in downstream processing.

Human Annotation

The manual process of labeling tweets to identify COVID-19 symptoms and its timely day information, serving as the gold-standard benchmark (ground truth) against which LLM performance was evaluated.

LangChain

LangChain is a framework designed for developing applications that utilize large language models. In this study, LangChain was used to orchestrate prompt executions and manage structured input/output interactions with GPT-4-0613.

Large Language Model (LLM)

An advanced AI model trained on massive text corpora to understand and rewrite natural language. In this research, LLMs such as GPT-4 and Gemini 2.0 were used to extract, normalize, and structure health symptom mentions from tweets. They were also utilized for semantic matching and categorization of symptom mentions.

Natural Language Processing (NLP)

A subfield of artificial intelligence that enables machines to read, interpret, and generate human language. It forms the foundation of this study's symptom extraction and analysis pipeline.

Precision

A metric that measures the proportion of correctly predicted symptoms out of all symptoms predicted by the model. It reflects the model's ability to avoid false positives.

Prompt Engineering

The technique of crafting targeted instructions or questions to guide LLMs in producing desired and structured responses. In this study, it was essential to develop prompts that are able to extract symptoms, performing semantic matching, normalization, and categorization in a consistent JSON format.

Public Health Informatics

The application of information science, including artificial intelligence, to support the surveillance, analysis, and management of public health data. This study contributes to the field by proposing an LLM-driven method to identify symptom trends from social media.

Recall

A metric that measures the proportion of correctly predicted symptoms out of all actual symptoms present in the data. It reflects the model's ability to avoid false negatives.

Symptom Categorization

The process of assigning normalized symptoms to broader medical categories (e.g., grouping "sore throat" and "cough" under "Respiratory Symptoms"). Categorization in this study was performed using Gemini 2.0 Flash after normalization.

Temporal Information Extraction

The process of identifying time references in tweets (e.g., "yesterday", "for three days") and converting them into structured formats (e.g., Day 2) to track symptom change over time.

Twitter Data

Twitter data refers to publicly available content collected from the social media platform Twitter (now X), including tweet text, user metadata, timestamps, and contextual elements such as hashtags and mentions. In this study, the tweet text served as the foundation for extracting and analyzing COVID-19 symptoms across multiple days.

1. INTRODUCTION

1.1 Background

Social media has transformed how people communicate about health, offering a timely lens into personal experiences, symptom progression, and public sentiment during health crises. Among the many platforms, Twitter (now X) has become a particularly valuable resource due to its large number of active users and rapid information sharing. During the COVID-19 pandemic, millions of individuals used Twitter to report symptoms, express concerns, and document their health status, often before seeking medical help. These publicly available data streams provide rich opportunities for supplementing traditional health surveillance systems with early-warning signals and population-level trends (Lamsal, 2021; Lee et al., 2015).

Despite its potential, mining consistent health-related information from Twitter data remains challenging. Unlike structured medical records or peer-reviewed clinical notes, Twitter data is informal, noisy, and highly variable. People describe the same symptom using drastically different expressions — for instance, “burning up,” “temp’s going crazy,” and “my body’s on fire” may all signify fever. This linguistic diversity poses a fundamental barrier to symptom recognition using conventional Natural Language Processing (NLP) techniques (Correia et al., 2020). Furthermore, social media posts often include sarcasm, exaggeration, or metaphorical language. A tweet like “Guess I’m dying from this headache lol” might not indicate a genuine medical concern, further complicating interpretation (Sarker & Gonzalez, 2016).

Temporal information extraction is another significant hurdle. Tweets frequently reference symptoms using various expressions such as “yesterday” or “been feeling sick for a few days.” These time markers need to be interpreted and normalized into structured formats (e.g., Day 1, Day 2) for effective tracking, which traditional models fail to do reliably (Wang et al., 2018). Moreover, existing biomedical terminologies like UMLS and SNOMED CT were developed for structured clinical data and often do not generalize well to informal, user-generated text (Lu et al., 2024).

Recent advancements in Large Language Models (LLMs) offer a promising alternative. LLMs such as GPT-4 and Gemini 2.0 have demonstrated the ability to interpret complex, informal language and produce structured outputs aligned with user-defined schemas (Busch et al., 2025; Guo et al., 2024). These models can potentially bridge the gap between noisy social media posts and clinically meaningful insights by extracting and normalizing symptom mentions with high semantic accuracy. However, their performance remains underexplored in the context of health concept recognition from social media — particularly when dealing with temporal ambiguity, slang, and variability across repeated model runs.

This study is motivated by the need to address these limitations through a unified LLM-based framework for health symptom recognition. By systematically evaluating multiple LLM configurations (e.g., GPT-4-0613 via LangChain, GPT-4 Turbo, Gemini 2.0 Flash) and applying an ensemble-based output stabilization strategy, the research aims to enhance the consistency of symptom extraction from informal text. The proposed approach contributes to public health informatics by laying the groundwork for scalable, AI-driven systems that support timely monitoring of health signals from social media platforms.

1.2 Introduction

The widespread adoption of Large Language Models (LLMs) has transformed the landscape of natural language understanding, enabling more flexible and efficient interpretation of informal, domain-specific, and context-rich text (Bommasani et al., 2021). Unlike traditional Natural Language Processing (NLP) pipelines that rely on rigid rule-based or shallow learning algorithms, LLMs demonstrate strong capabilities in semantic understanding, generation, and pattern recognition from diverse linguistic inputs (Lu et al., 2024; Zhang et al., 2025). These advancements have opened new opportunities in the field of public health informatics, where user-generated content on social media platforms like Twitter can serve as a timely source of population-level health insights.

This study investigates the use of LLMs to extract, normalize, and categorize COVID-19-related symptoms from Twitter data. Specifically, it evaluates the performance of multiple LLM variants—including GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash—in

recognizing health concepts expressed through informal language. The LLMs are tasked with interpreting tweets that vary widely in structure, tone, and symptom descriptions, ranging from explicit statements like “I’ve had a fever since Monday” to more nuanced expressions such as “burning up again, ughhh.” In addition to linguistic variation, tweets often reference temporal information using various phrases like “for the past few days” or “yesterday,” requiring the models to infer and align symptoms to a consistent time frame. Accurately interpreting this implicit temporal context is essential for mapping symptom progression and enabling structured public health analysis.

To address the non-determinism in outputs typically observed in generative models, this research introduces an ensemble mechanism that involves running each model three times with the same tweets and prompt and selecting the most consistent output based on semantic agreement. This approach enhances output consistency, a critical factor for downstream tasks such as normalization and symptom categorization. After extraction, all model outputs are uniformly processed using Gemini 2.0 Flash to perform semantic similarity matching against ground truth annotations. This step allows the system to recognize when different expressions—such as “burning up” or “feeling really hot”—mean the same thing, like “fever.” Instead of checking for exact word matches, it compares the meaning of phrases to ensure accurate evaluation. Once the symptoms are matched, they are then standardized into defined symptom categories to enable structured analysis.

In addition to symptom extraction, this framework emphasizes temporal context understanding—transforming relative phrases like “last night” or “day three” into cardinal format that support trend visualization across days. Prompt engineering plays a vital role in enabling this functionality throughout the framework. Carefully crafted prompts guide the LLMs to generate consistent and structured outputs in machine-readable formats. For instance, specifying a format like { "day": 1, "symptom_list": ["sore throat", "fever"] } not only facilitates temporal alignment but also supports downstream tasks such as normalization, categorization, and performance evaluation.

Ultimately, this research seeks to demonstrate how the integration of advanced LLMs, consensus-based refinement, and structured prompt design can improve the consistency and interpretability of symptom information extracted from informal Twitter posts. By doing so, it builds a bridge between social media discourse and structured public health data pipelines, contributing to the

ongoing advancement of AI-driven health surveillance systems (Jiang et al., 2023; Ntinopoulos et al., 2025; Guo et al., 2024).

1.3 Why LLMs and Prompts?

Extracting structured health information from social media presents significant challenges due to the informal, context-dependent, and highly variable language used by individuals online. Unlike clinical notes or standardized health records, social media posts frequently contain slang, abbreviations, emojis, and metaphorical expressions. For instance, symptom descriptions such as “can’t stop coughing my lungs out” or “my body’s on fire” may both indicate respiratory distress or fever but lack clinical terminology. Prior research has shown that this linguistic variability severely limits the effectiveness of traditional NLP systems and even domain-specific biomedical models (Liu et al., 2022; Sarker & Gonzalez, 2016).

Large Language Models (LLMs) have emerged as a promising alternative due to their ability to understand context, generalize across tasks, and adapt to varied input styles. However, even advanced LLMs like GPT-4 and Gemini are subject to non-determinism—that is, the same prompt can yield different outputs on different runs. This behavior introduces instability in extraction pipelines, particularly in sensitive applications such as health informatics, where consistency and reproducibility are crucial (Guo et al., 2024; Ntinopoulos et al., 2025).

To mitigate such variability, recent studies have explored ensemble-style inference, where a model is executed multiple times and outputs are aggregated using majority voting or semantic similarity (Jiang et al., 2023). This strategy improves reliability by capturing a consensus across multiple generations rather than relying on a single response, which may be affected by randomness or incomplete interpretation. Some research also advocates the use of model ensembles—leveraging the strengths of multiple LLMs—to enhance coverage and reduce bias in domain-specific extraction tasks (Kumar et al., 2022).

Another critical factor in improving LLM performance is prompt engineering. Prompt design plays a central role in shaping the behavior of generative models and guiding them toward producing structured and interpretable outputs. Recent work demonstrates that prompts specifying structured

output formats, such as JSON templates or slot-filling instructions, significantly improve accuracy and ease downstream processing (Zhang et al., 2025; Shen et al., 2023). This is particularly relevant in the context of health symptom extraction, where consistent formatting allows for easier mapping to standard categories and timelines.

Together, these findings underscore the need for a multi-model, multi-prompt strategy that balances interpretability, flexibility, and output consistency. By integrating insights from these studies, the present work adopts LLM approach combined with targeted prompt engineering to improve the robustness of health information extraction from social media. Each model was evaluated independently to better understand its performance characteristics prior to ensemble decision-making and downstream normalization. Specific implementation details and evaluation procedures are discussed in the following methodology section.

2. STATEMENT OF PROBLEM

Social media platforms, particularly Twitter (now X), have emerged as informal yet timely sources of public health information during global health crises. Individuals frequently use these platforms to share personal health experiences, including symptom progression and emotional responses to illness. The COVID-19 pandemic magnified this trend, generating massive volumes of self-reported data that, if properly extracted and interpreted, could enhance traditional disease surveillance systems (Lamsal, 2021; Sinnenberg et al., 2017). Despite this potential, mining consistent and structured health-related information from social media remains a significant challenge.

Unlike clinical records or structured surveys, tweets often include colloquial expressions, abbreviations, sarcasm, emojis, and fragmented language. For example, phrases like “my lungs are on fire” or “feeling dead today” may ambiguously imply symptoms such as shortness of breath or fatigue. Additionally, the brevity and lack of medical context make it difficult to distinguish genuine symptom reports from non-serious posts (Chancellor et al., 2016). This linguistic variability complicates the task of identifying precise health concepts in noisy user-generated content.

Temporal information adds another layer of complexity. Tweets rarely reference dates explicitly; instead, they use relative terms like “yesterday,” “for the past few days,” or “day three.” Extracting structured symptom progression from such language requires interpretation and temporal normalization—tasks that traditional NLP systems struggle to handle reliably (Lin et al., 2013). While efforts have been made to extract time expressions from clinical notes, those methods are often not transferable to informal platforms due to contextual ambiguity and missing anchors.

Tools such as MetaMap and cTAKES—designed to identify medical terms using resources like the Unified Medical Language System (UMLS), a controlled biomedical vocabulary—have been applied to extract symptoms from text. However, these tools rely heavily on formal terminology and exact matches, making them poorly suited for processing informal or metaphorical expressions

common in tweets (Denecke, 2014). Their performance degrades when applied to non-standard language that lacks clinical structure or context.

Recent advances in large language models (LLMs), including GPT-4 and Gemini, have significantly improved the ability to interpret informal text. These models are capable of understanding context, paraphrasing, and semantic equivalence, offering a promising alternative for health information extraction. However, their non-deterministic nature—producing different outputs for the same input—limits their utility in clinical or epidemiological pipelines where consistency is essential (Guo et al., 2024; Xie et al., 2025). Furthermore, most existing work using LLMs focuses on one-off extractions and lacks systematic evaluation pipelines that include semantic matching, normalization, and categorization aligned with human-annotated data.

Although individual studies have demonstrated the potential of LLMs in health-related social media mining (Jiang et al., 2023; Guo et al., 2024), no comprehensive framework currently exists that combines multiple LLM configurations with prompt-based schema enforcement, ensemble consensus selection, and downstream symptom structuring. The lack of reproducible, scalable solutions for temporally aligned symptom extraction hinders the broader application of LLMs in public health informatics.

This thesis addresses the problem of designing a consistent, prompt-driven, and reproducible framework for extracting, normalizing, and categorizing COVID-19 symptom information from informal Twitter posts using large language models. The framework is designed to handle temporal alignment, manage model output variability, and support structured evaluation using human-annotated ground truth.

By focusing on consistency across multiple LLMs, employing ensemble-style inference, and incorporating downstream semantic processing, this research contributes a scalable methodology for health-related symptom mining from noisy, real-world data sources. It bridges the gap between unstructured social discourse and structured health surveillance systems, laying the groundwork for broader application in public health monitoring.

3. RESEARCH QUESTION

3.1 Research Aim

The aim of this study is to design and evaluate a structured framework that leverages large language models (LLMs) for the extraction, normalization, categorization, and temporal alignment of COVID-19 symptoms from informal Twitter posts. The framework incorporates prompt engineering, ensemble-style consensus generation, and downstream structuring techniques to improve output consistency. The performance of three LLM configurations—GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash—is evaluated in extracting symptom and temporal entities, annotated ground truth dataset to assess their effectiveness in health concept recognition.

3.2 Research Questions

This research is guided by the following questions:

RQ1: To what extent can large language models accurately and consistently extract COVID-19 symptom mentions and their associated temporal references from informal Twitter posts when compared to human annotations?

RQ2: How does ensemble-based consensus generation influence the consistency and semantic alignment of LLM outputs in symptom extraction tasks?

RQ3: Can structuring performed after extraction—through normalization and categorization—enhance the interpretability of LLM outputs for public health symptom trend analysis?

4. SIGNIFICANCE OF THE STUDY

4.1 Practical and Public Health Significance

This study provides a scalable and reproducible pipeline for mining public health data from social media. In health emergencies such as the COVID-19 pandemic, conventional data sources—including clinical records and epidemiological surveys—often suffer from reporting delays, limited geographic reach, and privacy constraints. In contrast, platforms like Twitter offer timely, crowd-sourced signals that can serve as early indicators of emerging health issues. However, extracting reliable information from such platforms requires overcoming challenges related to informal language, fragmented syntax, and non-standard symptom descriptions.

By improving the consistency and structure of symptom extraction using large language models (LLMs), this research lays the groundwork for automated tools that can assist in disease monitoring, trend detection, and real-time surveillance. The proposed framework supports the alignment of extracted symptoms with day-level temporal markers, allowing for detailed symptom progression analysis. These features can enable public health dashboards, early outbreak warning systems, and dynamic mapping of symptom distributions across time.

While this thesis focuses on COVID-19 symptoms, the methodology can also be applied to other infectious diseases, chronic conditions, or even mental health monitoring. This expands the relevance of the work beyond a single disease context, making it a potentially valuable asset in broader public health surveillance efforts.

4.2 Academic Significance

This research contributes to the growing body of literature on the use of LLMs in biomedical natural language processing, particularly in the context of informal and noisy text. While most prior studies have focused on clinical documents or structured datasets, this work addresses a critical gap by systematically evaluating how multiple LLMs perform on informal, user-generated content. It advances the academic discourse by introducing a unified, end-to-end framework that spans not only extraction, but also normalization, temporal alignment, and categorization.

The study also offers a methodological contribution by incorporating prompt engineering and multi-run ensemble consensus as strategies to manage the non-deterministic nature of generative models. These techniques—although frequently acknowledged in LLM research—have rarely been formalized or evaluated in public health text mining contexts. The structured, schema-driven approach used here can inform future research in both computational linguistics and health informatics by providing a reproducible design for consistent information extraction in low-structure environments.

4.3 Technological Relevance

The study leverages advanced large language models—GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash—and evaluates their capacity to extract structured, meaningful health signals from noisy and informal text. It highlights the importance of prompt engineering, schema-guided output structuring, and ensemble-based output stabilization—areas that are increasingly critical in the development of reliable AI systems. By aligning model performance with real-world application needs, this research bridges the gap between cutting-edge NLP development and deployable solutions in health informatics.

5. SCOPE OF THE STUDY

This study focuses on evaluating the effectiveness of large language models (LLMs) in extracting, normalizing, categorizing, and temporally aligning COVID-19-related symptom mentions from informal Twitter posts. The analysis is based on a human-annotated dataset of 635 tweets, which resulted in 994 tweet-day symptom annotations used as ground truth for model evaluation.

The selected models—GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash—are assessed for their inference-time performance using structured prompt formats and ensemble-based consensus generation. The study is limited to English-language tweets specifically related to COVID-19 and does not extend to other diseases, languages, or platforms.

The scope is limited to evaluating LLMs during inference and does not involve model training, fine-tuning, or parameter modification. Upstream processes such as data collection, noise filtering, and classification of tweets as personal health narratives are outside the scope of this work, as the study uses a curated dataset of 635 tweets that were selected and manually annotated in prior research to include day-referenced COVID-19 symptom mentions.

Downstream components such as symptom normalization, temporal alignment, and symptom categorization are evaluated using Gemini 2.0 Flash to maintain consistency across outputs from all models. The focus of the evaluation is on the accuracy and structural consistency of extracted information relative to the human-annotated ground truth, with specific attention to semantic correctness and day-level symptom tracking.

6. ASSUMPTIONS

This research is guided by several methodological and technical assumptions that underpin the design of the framework, execution of the experiments, and interpretation of the results. These assumptions reflect accepted practices in the evaluation of language model performance and are necessary to ensure consistency, feasibility, and comparability across models and stages of the pipeline.

1. Human Annotation Accuracy

It is assumed that the human annotations used in the dataset are accurate and faithfully represent the presence of COVID-19 symptoms and temporal references as expressed in the tweet content. These annotations serve as the gold standard for evaluating model performance.

2. Prompt Interpretability Across Models

The structured prompts used for each LLM are assumed to be equally interpretable across different model architectures. Although model behavior may vary, the use of schema-based prompts provides a consistent format for guiding extraction tasks.

3. Sufficiency of Multi-Run Variability Sampling

Running each LLM three times per tweet is assumed to provide sufficient coverage of output variability to enable meaningful consensus selection. This number of runs balances practical resource constraints with the goal of capturing model inconsistencies.

4. Consistency of Gemini 2.0 Flash for Semantic Evaluation

Gemini 2.0 Flash is assumed to be semantically reliable and consistent for the tasks of symptom normalization, categorization, and similarity-based evaluation. It is applied uniformly across outputs from all models to ensure consistency in downstream processing.

5. Temporal Expression Mapping Without External Metadata

Temporal references in tweet text (e.g., "yesterday", "day three") are assumed to be mappable to specific numeric day values based solely on prompt-based instructions. This study does not incorporate tweet timestamps or user context for temporal disambiguation.

7. LIMITATIONS

While this study offers a structured and reproducible approach to symptom extraction using large language models (LLMs), it is important to recognize the limitations that affect the scope, generalizability, and interpretability of the findings. These limitations inform how results should be contextualized and suggest directions for future work.

1. Dataset Size and Diversity

The dataset is limited to 635 tweets, which—while sufficient for controlled evaluation—may not capture the full linguistic diversity present in larger or more heterogeneous social media corpora.

2. Language and Disease Focus

All analyzed tweets are in English and specific to COVID-19. As a result, the findings may not generalize to other diseases or to symptom expressions in non-English contexts.

3. Use of LLMs Without Fine-Tuning

The LLMs are employed in a zero-shot or few-shot capacity without task-specific fine-tuning. Although this mirrors realistic deployment scenarios, it may not reflect the maximum achievable performance of these models under domain adaptation.

4. Reliance on a Black-Box Evaluation Component

Semantic matching and normalization are performed using Gemini 2.0 Flash, which is a proprietary and opaque system. This introduces a layer of interpretability limitation, particularly in understanding how normalization decisions are made.

5. Handling of Non-Literal or Ambiguous Content

The evaluation framework does not specifically address tweets containing sarcasm, humor, exaggeration, or figurative language. Such content may mimic symptom-related phrasing but lacks clinical relevance, posing challenges for both annotation and model interpretation.

8. DELIMITATIONS

To ensure a focused and feasible research scope, this study is intentionally bounded in the following ways. These delimitations reflect strategic design choices to maintain methodological clarity and manage resource constraints, and they help define the context within which the results should be interpreted.

1. Model Selection

Only three large language models (LLMs) are evaluated: GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash. Other LLMs, such as Claude, Mistral, PaLM, or open-source alternatives, are not considered in this study.

2. Static Ground Truth Evaluation

Model performance is assessed against a fixed, human-annotated dataset. Real-time data streams, evolving annotation frameworks, or active learning strategies are outside the scope of this research.

3. Limited Consensus Strategy

The ensemble consensus mechanism is restricted to three repeated inference runs per model per tweet. Other ensemble techniques, such as uncertainty estimation, confidence weighting, or model voting across architectures, are not applied.

4. Exclusion of Population-Level Analysis

This study does not aim to explore symptom co-occurrence, clustering, or geospatial distribution patterns at the population level. Its focus remains on the extraction quality, consistency, and interpretability of model outputs.

5. Ethical and Sociotechnical Considerations

While the research assumes compliance with ethical standards in data handling and anonymization, it does not engage in a formal discussion of the ethical, privacy, or policy implications of mining health data from social media platforms.

9. LITERATURE REVIEW

9.1 Health Concept Extraction

Health concept extraction involves identifying medically relevant entities—such as symptoms, diseases, medications, and anatomical references—from unstructured text. In biomedical natural language processing (BioNLP), early efforts concentrated on structured documents like clinical notes, discharge summaries, and electronic health records (EHRs). Rule-based systems and dictionary-driven methods utilizing resources like the Unified Medical Language System (UMLS), SNOMED CT, and the Medical Subject Headings (MeSH) taxonomy provided the foundation for recognizing standardized clinical terms (Denecke, 2014).

However, these systems often struggle with informal, user-generated content. Their reliance on formal terminology limits their ability to generalize to noisy, context-rich domains such as social media. For example, a user might describe shortness of breath as “I can't catch my breath,” which lacks a direct lexical match with UMLS terms.

Correia et al. (2020) observed that health-related posts on platforms like Twitter often reflect symptoms, behaviors, and emotional states but are expressed using vague, metaphorical, or non-standard language. This linguistic variability complicates both concept detection and classification, especially for models trained exclusively on clinical text.

To address this, researchers began adapting machine learning approaches for health concept recognition in unstructured and informal contexts. Sarker et al. (2022) proposed a self-supervised method to extract COVID-19-related symptoms from Twitter using lexicons and weak supervision. Their findings emphasized the limitations of traditional models when confronted with idiomatic or creative symptom expressions.

While deep learning and contextual models have shown improvements, they still require task-specific tuning or annotated training data. This has prompted growing interest in more flexible,

context-aware systems that can handle linguistic diversity without retraining, setting the stage for the adoption of large language models in this domain.

9.2 Methods of Extraction

Early approaches to health concept extraction relied heavily on rule-based and dictionary-driven systems. These methods typically utilized static lexicons, including terminologies from the Unified Medical Language System (UMLS), SNOMED CT, and other controlled vocabularies, to match symptom keywords within clinical narratives or structured documents (Denecke, 2014). While these approaches performed well in formal medical settings, they were often brittle and inflexible when applied to informal or noisy text sources such as social media.

To address limitations in recall and linguistic coverage, machine learning techniques were introduced. Traditional classifiers such as Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) used handcrafted features like part-of-speech tags, lexical patterns, and syntactic dependencies to recognize symptom mentions. However, these models required large amounts of labeled training data and struggled with generalization in cross-domain applications (Han et al., 2023).

The development of deep learning models marked a significant shift in extraction methods. Recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models such as BERT and BioBERT improved upon previous techniques by capturing broader semantic and contextual relationships within text. Wang et al. (2018) demonstrated that these models could significantly outperform traditional methods in named entity recognition tasks within biomedical corpora. Nevertheless, when applied to informal text such as tweets, their performance often degraded without fine-tuning on domain-specific language (Lamsal, 2021).

Recent advancements in large language models (LLMs), including GPT-3, GPT-4, and Gemini 2.0 Flash, have introduced prompt-based querying as an alternative to supervised training. These models are pre-trained on massive, diverse corpora and can generalize to new tasks with minimal input specification. Guo et al. (2024) evaluated LLMs on health classification tasks and found that well-structured prompts could yield competitive results even without additional training data.

Prompt engineering has since emerged as a key strategy for improving output structure and consistency in LLM-driven extraction pipelines. Ouyang et al. (2022) showed that instruction tuning with format-specific prompts enhanced model reliability across multiple tasks. Similarly, Xu and Cheng (2023) emphasized that schema-constrained prompts can reduce hallucinations and improve alignment with evaluation benchmarks in biomedical settings.

However, LLMs remain non-deterministic, and their outputs may vary significantly across runs depending on factors like decoding strategy, temperature settings, and prompt phrasing. Bang et al. (2023) documented substantial variability in LLM performance across identical prompts, noting challenges for reproducibility in structured extraction tasks. This variability has led researchers to explore consensus-based methods such as majority voting or semantic agreement to stabilize outputs across multiple model runs (Song et al., 2024).

9.3 Characteristics of Social Media Data

Social media platforms, particularly Twitter, provide a rich stream of real-time public discourse, including health-related narratives. During public health events such as the COVID-19 pandemic, individuals frequently used Twitter to document symptoms, emotional states, and self-isolation experiences. This content offers valuable insight for syndromic surveillance and public health research, but its informal, variable nature creates significant challenges for automated processing.

One defining characteristic of social media data is its linguistic informality. Due to character limits and casual tone, users often employ abbreviations, creative spelling, emojis, and slang. Phrases like “I’m burning up,” “my head’s gonna explode,” or “coughing like crazy” can refer to fever or headache, yet are difficult for traditional NLP systems to interpret (Correia et al., 2020). These non-standard expressions reduce the precision of lexicon-based and rule-driven extraction methods.

Another limitation is the brevity and lack of context in individual posts. Tweets often appear in isolation, without a clear timeline or background information. Han et al. (2023) observed that temporal resolution is especially difficult in narrative health data, and even more so in short-form content where users may mention “day 3” without clarifying what symptoms preceded it or when they began. This complicates efforts to extract coherent health event trajectories.

Social media content also includes a variety of non-standard linguistic elements. Tweets may feature emojis, hashtags, excessive punctuation, and non-linear grammar. While these features may convey emotional tone or severity, they often obscure the underlying health content. As Lamsal (2021) noted, tweets combine multimodal and textual cues, which can both enrich and confound automated interpretation.

Another concern is semantic noise—irrelevant or misleading content that includes medical terms without health context. For example, “I have Taylor Swift fever” or “that movie gave me a headache” use symptom-related terms metaphorically. Correia et al. (2020) found that fewer than 20% of symptom-flagged tweets reflected verifiable health mentions, underscoring the importance of relevance filtering.

Despite these challenges, social media data provides a timely signal for emerging health trends. Posts are often created as symptoms occur, allowing early insight ahead of official reporting systems (Sarker et al., 2022). Furthermore, Twitter captures perspectives from individuals who may not interact with formal healthcare systems, offering a more inclusive representation of population health behavior.

Given this combination of high noise and high potential value, successful symptom extraction from social media requires models that are both linguistically flexible and capable of contextual reasoning. This has led to increasing interest in adaptive, prompt-driven NLP systems that can generalize to informal and fragmented input.

9.4 Health Discussions on Social Media (Twitter/X and Other Platforms)

Over the past decade, social media platforms such as Twitter (now X), Facebook, Reddit, and health forums have emerged as informal yet powerful channels for health-related discourse. Users share personal health experiences, report symptoms, and discuss treatment options or emotional states. During public health emergencies such as the COVID-19 pandemic, these platforms offered early insight into individual-level health behaviors and symptom progression.

One of the earliest demonstrations of Twitter’s utility in public health came from Signorini et al. (2011), who found that tweet volume and content correlated with national trends in influenza-like illness. This early work established the foundation for using social media data as a tool for syndromic surveillance. In the context of COVID-19, researchers expanded this approach to track self-reported symptoms, vaccine sentiment, and pandemic-related stress. Lamsal (2021) compiled a large-scale dataset of multilingual COVID-19 tweets, which became a foundational resource for researchers working on pandemic-related text mining.

However, extracting structured insights from such discourse remains challenging. Tweets often mix symptom mentions with humor, lifestyle commentary, or emotional reflections. For instance, a post like “Day 5: Still can’t sleep, throat’s on fire, bingeing Netflix” blends symptom information with non-health content. Without nuanced interpretation, models risk either missing important signals or misclassifying unrelated language.

Social media discourse is also characterized by emotional variability and spontaneity. Users may express frustration, fear, or sarcasm in describing their symptoms. Correia et al. (2020) noted that the emotional tone and linguistic creativity in such posts complicate direct symptom extraction, especially for models trained on formal medical text.

Importantly, social media data includes populations often underrepresented in clinical datasets—such as individuals who self-manage mild illness or lack access to formal healthcare systems. Sarker et al. (2022) reported that self-reported symptoms on Twitter often precede official case counts, highlighting the potential of social media for early outbreak detection and community-level health monitoring.

Despite concerns about data quality, representativeness, and misinformation, researchers have explored integrating social media signals into hybrid surveillance systems. Guo et al. (2024) discussed the value of combining social media data with electronic health records to improve the timeliness and diversity of public health intelligence. When paired with robust filtering and extraction techniques, social media data can serve as a complementary input to traditional health monitoring systems.

Overall, health-related discussions on social media offer a dynamic, real-time lens into population health behaviors. The challenge lies not in access to data, but in the design of NLP systems capable of interpreting symptom expressions accurately amid informal language, limited context, and emotional noise.

9.5 Challenges in Processing Social Media Data for Health Concept Extraction

Extracting health-related information from social media platforms introduces a variety of challenges due to the unstructured, informal, and often ambiguous nature of user-generated content. These challenges are distinct from those encountered in processing structured clinical texts and require tailored methods that can account for variability in language, context, and intent.

One of the most pervasive challenges is linguistic inconsistency. Users rarely use standardized clinical terms; instead, they describe symptoms using informal or metaphorical language. Expressions such as “temp’s spiking,” “burning up,” or “my bones are made of fire” may refer to fever or body aches, but they lack direct lexical matches to standardized vocabularies. Prieto et al. (2014) observed that conventional lexicon-based tools struggled to detect such expressions due to their non-standard phrasing and absence from curated medical dictionaries.

Another key difficulty is the brevity and fragmentation of context in platforms like Twitter, where messages are limited in length and often posted in isolation. Symptom mentions may be distributed across tweets or embedded within replies, hashtags, or quoted text. Dunn et al. (2012) highlighted that the absence of surrounding narrative context significantly increases the risk of misinterpretation and false positives in health information extraction.

Sarcasm and figurative language present additional complications. Phrases such as “dying from this cute little cough” or “this headache’s trying to kill me” may not reflect actual medical conditions, but they can mislead literal extraction models. Detecting sarcasm remains an open problem in NLP, particularly in health-related applications where sentiment and symptom reporting often overlap (Bang et al., 2023).

The presence of noise and irrelevant content is also common. Health-related keywords may be used metaphorically or non-clinically—for instance, “Bieber fever” or “fever dream of a concert.” Al-Garadi et al. (2016) emphasized that such false positives can dominate retrieved results if relevance filtering is not applied. Their work found that automated systems often misclassify entertainment-related posts as health mentions without careful contextual disambiguation.

Misinformation poses another substantial challenge. During public health crises, social media often includes unverified claims, conspiracy theories, or emotionally charged narratives that resemble genuine symptom reports. Charles-Smith et al. (2015) called for the development of hybrid systems that combine NLP with credibility scoring and epidemiological validation to distinguish credible health signals from noise.

Finally, ethical and privacy considerations complicate the analysis of publicly available health-related social media content. Although tweets are posted in public domains, users may not be aware that their health disclosures are subject to analysis. Chancellor et al. (2019) advocated for responsible research practices, including de-identification, consent awareness, and Institutional Review Board (IRB) oversight, especially when working with mental health data or vulnerable populations.

These challenges underscore the need for context-aware, semantically robust NLP systems designed specifically for the social media environment. Successful approaches must integrate linguistic flexibility, sarcasm detection, noise filtering, and ethical safeguards to reliably extract health concepts from online discourse.

9.6 Solutions to Overcome Social Media Data Challenges

To address the complexity of extracting health-related content from noisy, informal platforms like Twitter, researchers have developed a range of technical and methodological strategies. These include preprocessing pipelines, context-aware models, schema-guided prompting, ensemble inference, and ethical data handling protocols—all aimed at increasing the reliability and usability of social media data for public health research.

Preprocessing and normalization remain foundational steps. Techniques include the removal of URLs, emojis, and non-alphanumeric characters, as well as spelling correction and lemmatization. Tools such as TwitIE and spaCy-based pipelines have been adapted to retain medically relevant tokens while filtering noise (Sinnenberg et al., 2017). Domain-specific lexicons, including crowdsourced symptom lists, have also been employed to map informal expressions to standardized terms (Smith et al., 2020).

To mitigate semantic ambiguity, researchers have increasingly adopted contextual embedding models. Transformer-based architectures like BERT, BioBERT, and ClinicalBERT are capable of interpreting phrases in context, which is critical when distinguishing literal symptom mentions from figurative language. Wang et al. (2021) showed that fine-tuned BERT models significantly outperformed LSTM baselines on symptom classification tasks in social media datasets.

Schema-based prompting has emerged as a powerful technique in LLM-driven extraction. Instead of open-ended instructions, structured prompts define expected output formats—e.g., JSON fields for symptom and time—thus guiding model behavior. Xu and Cheng (2023) demonstrated that schema-constrained prompts improve consistency and reduce hallucination, particularly in health information extraction tasks.

To address non-determinism in LLM outputs, multi-run consensus strategies are gaining adoption. This involves executing a model multiple times on the same input and selecting the most semantically consistent result using either majority voting or similarity scoring. Song et al. (2024) found that this method improved reproducibility in LLM-based classification and extraction tasks, especially when outputs were sensitive to prompt structure.

Temporal expression resolution has also received attention. Tools like MedTime (Lin et al., 2013) and other hybrid rule-based systems have been adapted to interpret vague time references such as “day 3” or “since Tuesday” by aligning them to structured formats. For social media, anchoring expressions to metadata such as tweet timestamps or threading patterns has shown promise (Sun et al., 2013).

In managing irrelevant or misleading content, relevance classifiers trained to detect genuine health mentions have proven effective. Lyu et al. (2021) developed a tweet-level classifier that filtered out metaphorical or entertainment-related posts, improving precision without sacrificing recall. Such classifiers are now frequently used as a preprocessing or validation step in social media mining pipelines.

Finally, ethical considerations have led to stronger privacy and governance protocols in health-related social media analysis. Chancellor et al. (2019) emphasized the importance of anonymization, opt-in datasets, and IRB compliance, particularly for sensitive domains such as mental health. Recent studies are increasingly transparent about consent practices and data storage policies in response to evolving expectations around digital ethics.

Collectively, these strategies represent a maturing toolkit for transforming informal and noisy social media content into reliable, structured data suitable for health analysis. They form the basis for ongoing work that integrates large language models with rule-based validation, multi-run stability checks, and ethical design frameworks.

9.7 Health Concept Extraction with Large Language Models (LLMs)

The emergence of large language models (LLMs) such as GPT-3, GPT-4, and Gemini 2.0 Flash has significantly advanced the field of health concept extraction from unstructured text. Unlike traditional machine learning or rule-based approaches that require extensive labeled data and retraining, LLMs leverage prompt-based querying and zero-shot generalization to perform tasks with minimal supervision.

Instruction-tuned models like GPT-4 demonstrate strong capabilities in extracting symptoms, diagnoses, and temporal information when provided with well-structured prompts. Ouyang et al. (2022) showed that instruction-tuned LLMs could follow natural language instructions and produce structured outputs in JSON or other predefined formats. This capability has particular relevance for biomedical domains, where schema-constrained outputs are critical for downstream integration and analysis.

In the context of social media, Guo et al. (2024) evaluated GPT-4 on health classification tasks involving Twitter data and found that prompt engineering played a key role in determining model accuracy. Their study reported that zero-shot prompting with clearly defined entity categories yielded performance comparable to fine-tuned transformer models—despite the LLMs having no task-specific training.

Prompt formatting has been shown to reduce ambiguity and improve output reliability. Xu and Cheng (2023) found that schema-based prompting improved the consistency of symptom extraction and reduced the likelihood of hallucinated entities. These structured prompts define expected fields (e.g., day, symptom list) and allow the model to produce machine-readable output aligned with human annotation guidelines.

However, LLMs are inherently non-deterministic. Output variability across multiple runs, even with the same prompt and input, presents challenges for reproducibility and consistency. Bang et al. (2023) examined LLM performance on informal health text and documented significant fluctuation in entity recognition depending on prompt structure, decoding temperature, and the specific model variant used.

This variability is particularly problematic in public health contexts where reliable extraction pipelines are needed. Song et al. (2024) proposed consensus-based generation strategies to mitigate LLM variability. By aggregating results across multiple runs and selecting the most semantically consistent outputs, researchers were able to improve agreement with human-annotated ground truth and reduce error rates.

Despite their advantages, LLMs also face difficulties interpreting idiomatic or metaphorical symptom descriptions common in social media. Sarcasm, brevity, and emotionally charged language reduce precision in entity extraction, especially when the model lacks context. As noted by Guo et al. (2024), further refinement is needed to handle ambiguous or creative phrasing in informal user-generated content.

Still, LLMs offer a compelling solution to many of the limitations associated with previous rule-based and supervised models. Their ability to perform extraction, normalization, and classification tasks without retraining opens new avenues for health information mining in low-resource or dynamic settings.

9.8 Normalization of Health Concepts Using Large Language Models (LLMs)

Normalization is a critical step in health text processing, involving the mapping of diverse and informal symptom expressions to standardized medical terms. This process enables aggregation, comparison, and downstream analytics by aligning noisy text with terminologies such as the Unified Medical Language System (UMLS), SNOMED CT, or ICD-10. While normalization is well-supported in structured clinical texts, it remains challenging in informal domains such as social media due to linguistic variability and contextual ambiguity.

Historically, normalization relied on rule-based methods and exact string matching, which offered high precision but limited recall. These techniques struggled with informal expressions like “burning up,” “can’t breathe properly,” or “feels like fire in my chest,” which often failed to match standardized terms directly. Liu et al. (2012) noted that manual lexicon expansion was often required to improve coverage, but such approaches lacked scalability and adaptability across domains.

Advancements in deep learning, particularly the introduction of contextual embedding models such as BioBERT and ClinicalBERT, improved semantic matching between informal language and medical terms. Wang et al. (2018) demonstrated that contextual models trained on biomedical corpora could capture underlying concept similarity, although their effectiveness was often limited to formal clinical datasets.

More recently, LLMs have introduced new normalization capabilities through prompt-based semantic reasoning. Rather than relying on fixed dictionaries, LLMs can interpret symptom phrases and determine if they are semantically equivalent to standardized terms. For example, a prompt might ask: “Do the phrases ‘my body’s on fire’ and ‘fever’ refer to the same symptom?”

Respond yes or no.” This approach supports dynamic mapping without retraining or external ontologies.

Semantic similarity techniques have become central to LLM-based normalization. Song et al. (2024) showed that embeddings generated by LLMs could be compared to determine conceptual alignment, achieving high accuracy in normalization tasks even under noisy conditions. When combined with prompt constraints, this strategy reduced false positives and improved agreement with human annotations.

LLMs also support multi-step reasoning, allowing not only normalization but also categorization into broader symptom classes. For example, once “can’t stop coughing” is normalized to “cough,” it may then be categorized under “Respiratory Symptoms.” This layered interpretation supports structured health analytics and trend tracking. Denecke (2014) emphasized that decomposing health text tasks into stages—extraction, normalization, and classification—leads to better alignment with controlled vocabularies and improves downstream interpretability.

Despite these advances, LLMs are not immune to errors. Hallucination—the generation of plausible but incorrect outputs—remains a concern. Ouyang et al. (2022) highlighted the importance of using clear prompt instructions and post-hoc validation to reduce misinterpretation. Prompt design and careful quality control remain essential components of reliable normalization pipelines.

Normalization also plays a key role in cross-model comparison. Because different LLMs may produce semantically similar symptom mentions in different formats, normalization enables fair evaluation and aggregation by standardizing outputs to a common terminology.

9.9 Named Entity Recognition (NER) in Health Concept Extraction Using LLMs

Named Entity Recognition (NER) is a foundational task in natural language processing (NLP) that involves identifying and classifying entities—such as symptoms, diseases, medications, or anatomical terms—within unstructured text. In biomedical domains, NER supports a variety of downstream applications, including clinical decision support, epidemiological modeling, and

patient monitoring. While traditional NER systems have shown high accuracy in structured data environments, their effectiveness diminishes in informal settings like social media.

Early biomedical NER systems relied on dictionary lookups and rule-based approaches, using resources such as SNOMED CT and UMLS to detect medically relevant phrases (Denecke, 2014). These methods worked well for standard terminology but lacked adaptability to informal, user-generated content, where symptom mentions often appear in colloquial or metaphorical language. For example, a phrase like “my bones are on fire” may signify joint pain or fever but would not match any formal term directly.

To improve generalizability, statistical models such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) were introduced. These models used token-level features such as part-of-speech tags, orthographic patterns, and word context to detect entity boundaries (Smith et al., 2020). However, their reliance on annotated training data and limited ability to capture long-range dependencies constrained their effectiveness in noisy, short-form text.

The emergence of deep learning models, particularly BiLSTM-CRF architectures and transformer-based models like BERT, marked a significant advancement. Pre-trained models such as BioBERT and ClinicalBERT achieved high accuracy on benchmark biomedical NER tasks (Lee et al., 2020). However, even these models required domain adaptation to handle the informal and variable language found on platforms like Twitter.

Large language models (LLMs) such as GPT-4 and Gemini 2.0 Flash offer an alternative approach to NER through prompt-based learning. Rather than relying on pre-defined labels or fine-tuning, LLMs can perform NER in a zero-shot or few-shot manner. For example, a prompt like “Extract all symptoms mentioned in the tweet and return them in a list” enables the model to infer the relevant entity types without additional supervision. Guo et al. (2024) found that GPT-4, when guided by structured prompts, achieved performance comparable to fine-tuned transformer models on symptom extraction from social media.

LLMs also offer the flexibility to define custom entity categories, such as “core symptoms,” “mild symptoms,” or “multi-day symptoms.” Xu and Cheng (2023) showed that LLMs can adapt to these customized NER schemas without retraining, as long as prompts are clear and output formats are specified.

However, challenges remain. LLMs often exhibit inconsistent boundary detection, particularly when symptom mentions are compound or nested. For example, “tightness in chest with short breath” might be split into two or one entity depending on prompt phrasing or decoding strategy. Song et al. (2024) emphasized that multi-run strategies and prompt tuning are necessary to stabilize entity detection across repeated runs.

Despite their variability, LLMs represent a major advancement in NER, especially for informal domains like social media. Their ability to handle unstructured inputs, generate structured outputs, and adapt dynamically to different entity types makes them valuable tools for biomedical information extraction in real-world settings.

9.10 Non-Deterministic Nature of Large Language Models (LLMs)

One of the core technical challenges in deploying large language models (LLMs) for biomedical information extraction is their non-deterministic behavior. Unlike traditional models, which produce consistent outputs for a given input, LLMs can generate different responses each time they are run—even with the same prompt and input text. This stochastic nature, while beneficial in open-ended tasks like creative writing or dialog generation, complicates reproducibility in domains that require precision and consistency, such as health concept extraction.

The source of this variability lies in the probabilistic decoding mechanisms used during inference. Techniques such as temperature scaling, nucleus sampling (top-p), and beam search introduce randomness to improve fluency and diversity (Ouyang et al., 2022). While these methods enable LLMs to explore multiple plausible outputs, they also lead to inconsistencies in structured tasks like symptom extraction, where even small variations can affect downstream evaluations.

In health applications, where output consistency is essential, this behavior presents a significant challenge. For instance, an LLM might identify “fever” in one run but overlook it in another for the same tweet. Song et al. (2024) documented this phenomenon across multiple LLM versions, including GPT-4, noting significant variance in entity-level predictions. Such inconsistencies can propagate through extraction pipelines and compromise the validity of public health trend analyses.

To address this, researchers have proposed multi-run consensus strategies. Rather than relying on a single model output, the same prompt is executed multiple times, and the final result is selected using voting, semantic similarity scoring, or majority agreement. Xu and Cheng (2023) demonstrated that such consensus approaches reduced extraction noise and improved alignment with gold-standard annotations in health datasets.

Another strategy involves temperature tuning—lowering the temperature setting (e.g., to 0 or 0.2) to encourage more deterministic responses. While this can stabilize outputs, it often reduces the model’s flexibility, especially when interpreting informal or ambiguous symptom expressions (Bang et al., 2023). Thus, a balance must be struck between consistency and linguistic adaptability.

Prompt engineering is also key to managing output variation. Structured prompts that include explicit instructions, format constraints (e.g., JSON schemas), or example completions are more likely to yield consistent and parsable results. Ouyang et al. (2022) emphasized that instruction-tuned models respond more reliably when prompts are clear, constrained, and task-specific.

Despite these efforts, non-determinism remains an open problem in LLM deployment, particularly in high-stakes applications like clinical research, epidemiological modeling, or automated symptom monitoring. Continued research into prompt stabilization, decoding control, and ensemble-based generation is needed to ensure that LLMs can be safely and effectively used in health informatics workflows.

9.11 Analysis and Synthesis

The application of large language models (LLMs) to health concept extraction from social media represents a convergence of developments in biomedical natural language processing, prompt

engineering, and real-world data mining. The literature reviewed in this chapter reveals the evolution of methods—from rule-based systems and supervised models to transformer architectures and prompt-driven LLMs—as researchers attempt to meet the demands of extracting meaningful information from informal, noisy, and context-fragmented data sources like Twitter.

Traditional approaches to health concept extraction, including dictionary lookups and rule-based pipelines, have proven effective in structured domains such as clinical notes (Denecke, 2014). However, their limitations in handling the informal and figurative language of social media have driven the adoption of statistical and deep learning models, such as CRFs, BiLSTMs, and BERT variants (Smith et al., 2020; Wang et al., 2018). These models demonstrated improved contextual understanding but still struggled to generalize across domains without extensive fine-tuning (Lee et al., 2020).

LLMs, such as GPT-4 and Gemini 2.0 Flash, offer a flexible alternative. Through prompt engineering, they enable schema-guided, zero-shot learning, and support extraction, normalization, and categorization tasks without retraining (Ouyang et al., 2022; Xu & Cheng, 2023). Their capacity to process non-standard expressions and understand context-rich inputs positions them well for social media applications, where symptom language is often metaphorical or embedded in humor and emotional narratives (Correia et al., 2020; Guo et al., 2024).

Despite their advantages, LLMs are not without drawbacks. A recurring challenge across studies is their non-deterministic output behavior, which complicates reproducibility and performance evaluation (Bang et al., 2023; Song et al., 2024). This issue is particularly problematic in health-related domains where interpretability, auditability, and consistency are critical. Researchers have responded by exploring consensus-based output selection, temperature control, and more constrained prompt designs to stabilize model behavior (Xu & Cheng, 2023).

Semantic similarity methods and contextual embeddings have also emerged as complementary tools for improving normalization and entity mapping in LLM pipelines. These techniques support the alignment of informal symptom expressions with standardized vocabularies, enhancing the utility of extracted data for downstream analytics (Song et al., 2024; Wang et al., 2021).

Moreover, the literature emphasizes the importance of ethical safeguards in social media-based health research. Issues of user consent, data anonymization, and representational fairness must be considered when mining sensitive health disclosures (Chancellor et al., 2019; Charles-Smith et al., 2015). Recent studies advocate for transparent data governance and IRB oversight, especially in contexts involving mental health or vulnerable populations.

In summary, the reviewed body of work illustrates a clear shift toward more adaptive, semantically informed, and ethically responsible methods for extracting health-related information from informal online discourse. These studies collectively highlight the need for robust, interpretable, and context-aware NLP frameworks capable of navigating the unique challenges posed by social media health data. The next phase of research builds upon this foundation, aiming to further stabilize, scale, and integrate LLMs into public health informatics workflows.

Table 1. Literature Review Summary

Study	Data Source	Task	Method/Model	Key Findings	Limitations
Correia et al. (2020)	Twitter (COVID-19)	Symptom detection	Rule-based + keyword filtering	Highlighted metaphorical and ambiguous symptom expression challenges	Low recall for idiomatic expressions
Sarker et al. (2022)	Social media posts	COVID-19 symptom extraction	Weak supervision with lexicons	Improved over rule-based methods in informal text	Still struggled with indirect symptom phrasing
Wang et al. (2018)	Clinical notes	NER & concept normalization	BioBERT	Achieved strong performance in structured medical data	Poor transfer to noisy, informal texts
Guo et al. (2024)	Twitter	Classification & extraction	GPT-4, prompt-based	Schema prompts yield results comparable to fine-tuned models	Vulnerable to prompt ambiguity
Song et al. (2024)	Twitter (health)	Extraction & normalization	LLM + semantic similarity + consensus	Improved consistency across LLM runs	Dependent on prompt tuning
Xu & Cheng (2023)	Synthetic & real health texts	Prompt consistency study	Schema-constrained prompting	Reduced hallucinations, improved output alignment	Needs detailed prompt templates
Jiang et al. (2023)	Twitter (COVID-19)	Day-based symptom extraction & evaluation	GPT-3.5, GPT-4 (API), ChatGPT, Bard; prompt-based + semantic matching with GPT-4	GPT-3.5 & GPT-4 performed best in extracting symptoms by day without tuning; novel use of GPT-4 for semantic match evaluation	Limited to 635 annotated tweets; prompt/output inconsistency; no fine-tuning
Chancellor et al. (2019)	Reddit & Twitter (mental health)	Ethical review of social data use	Qualitative review	Emphasized need for consent, anonymization, IRB oversight	Lacked computational evaluation

10. METHOD

10.1 Overview of the Research Workflow

This study implements a modular, multi-stage framework for extracting, normalizing, and categorizing COVID-19-related symptoms from informal Twitter posts using large language models (LLMs). To improve clarity, the workflow is divided across two figures: Figure 1.a illustrates the initial extraction and consensus generation process, while Figure 1.b presents the downstream stages of evaluation, normalization, categorization, and visualization.

As depicted in Figure 1.a, the first phase begins with a pre-annotated dataset of tweets, each containing symptom mentions and their associated temporal references. These tweets are processed through two primary pipelines. The baseline pipeline employs GPT-4-0613 via LangChain to extract day-symptom pairs using a structured JSON schema. This configuration produces a single output per tweet and serves as a baseline for model evaluation. In contrast, the experimental pipeline uses GPT-4 Turbo and Gemini 2.0 Flash. Due to the non-deterministic nature of generative LLMs, each model is executed three times per tweet using the same prompt and system instructions. These multiple generations are then manually reviewed to identify the most semantically consistent response, forming the consensus output. This ensemble-style mechanism helps reduce output variability and improves consistency for downstream processing.

The second part of the framework, shown in Figure 1.b, begins with a semantic similarity evaluation using Gemini 2.0 Flash. This step compares each model’s output to human-annotated ground truth data using semantic matching rather than strict lexical equivalence. Evaluation metrics including accuracy, precision, recall, and F1-score are calculated to assess each configuration’s performance. After evaluation, all outputs undergo a normalization stage where informal expressions such as “burning up” or “feeling hot” are mapped to standardized medical terms like “fever.” This normalization is performed using Gemini 2.0 Flash, followed by a rule-based term replacement process that substitutes the original symptom phrases in the outputs with their normalized equivalents.

Next, a categorization stage assigns each normalized symptom to a predefined medical category, such as respiratory, neurological, or gastrointestinal, again using Gemini 2.0 Flash. The categorized outputs are compiled into a structured data representation in the form of a three-dimensional interactive data cube. In this cube, the X-axis corresponds to the symptom day referenced in the tweet, the Y-axis denotes the symptom category, and the Z-axis reflects the frequency of symptom mentions. This cube serves as the foundation for visualization and pattern recognition. Additional visual outputs include heatmaps, confusion matrices, and comparative performance charts that allow for temporal and categorical analysis of symptoms across models.

The overall architecture emphasizes consistency, modularity, and reproducibility. By applying the same prompt structure and semantic tools across multiple LLMs, this framework enables structured evaluation and supports scalable symptom tracking from social media data.

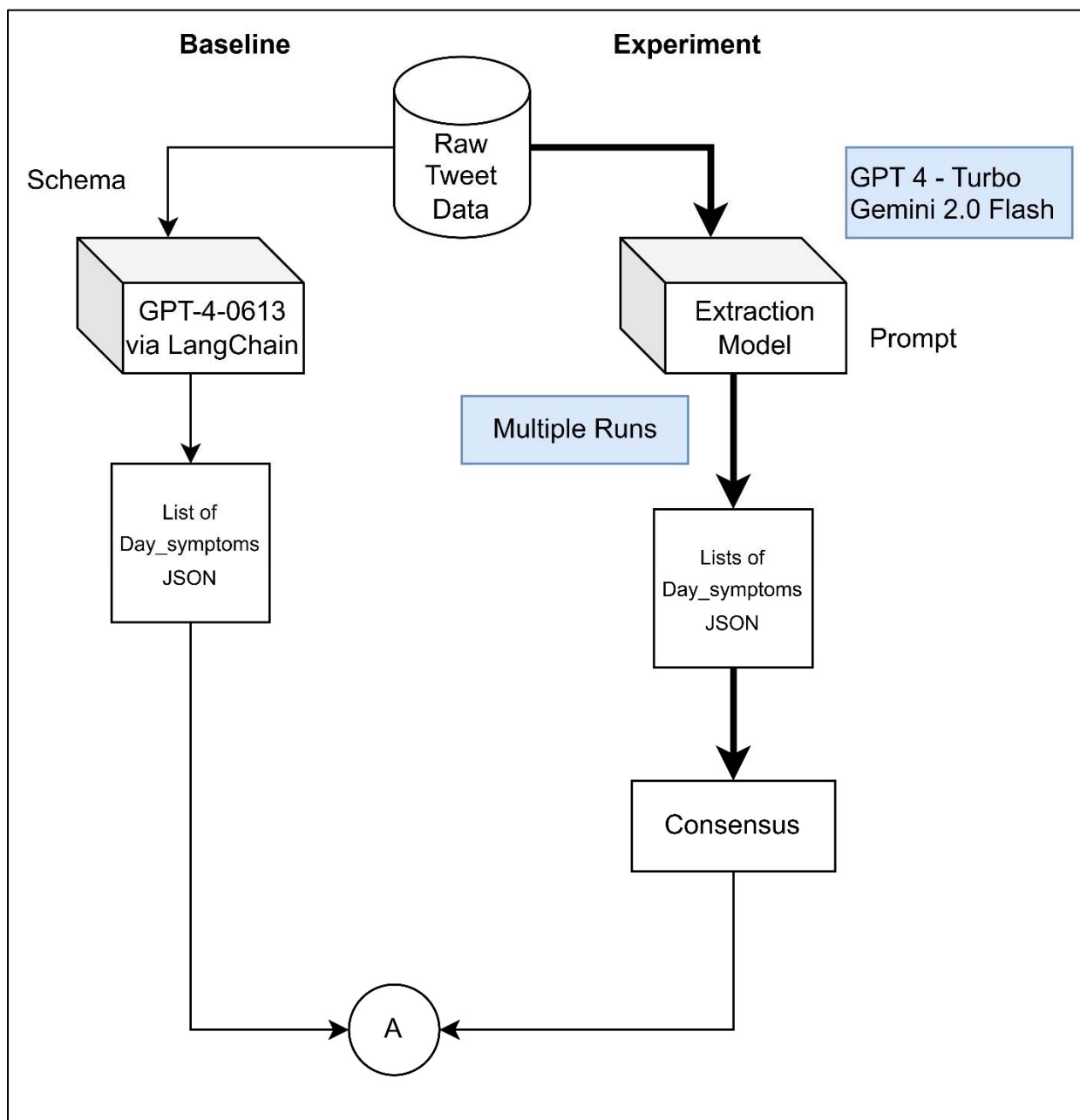


Figure 1. Baseline and Experimental Extraction Pipeline

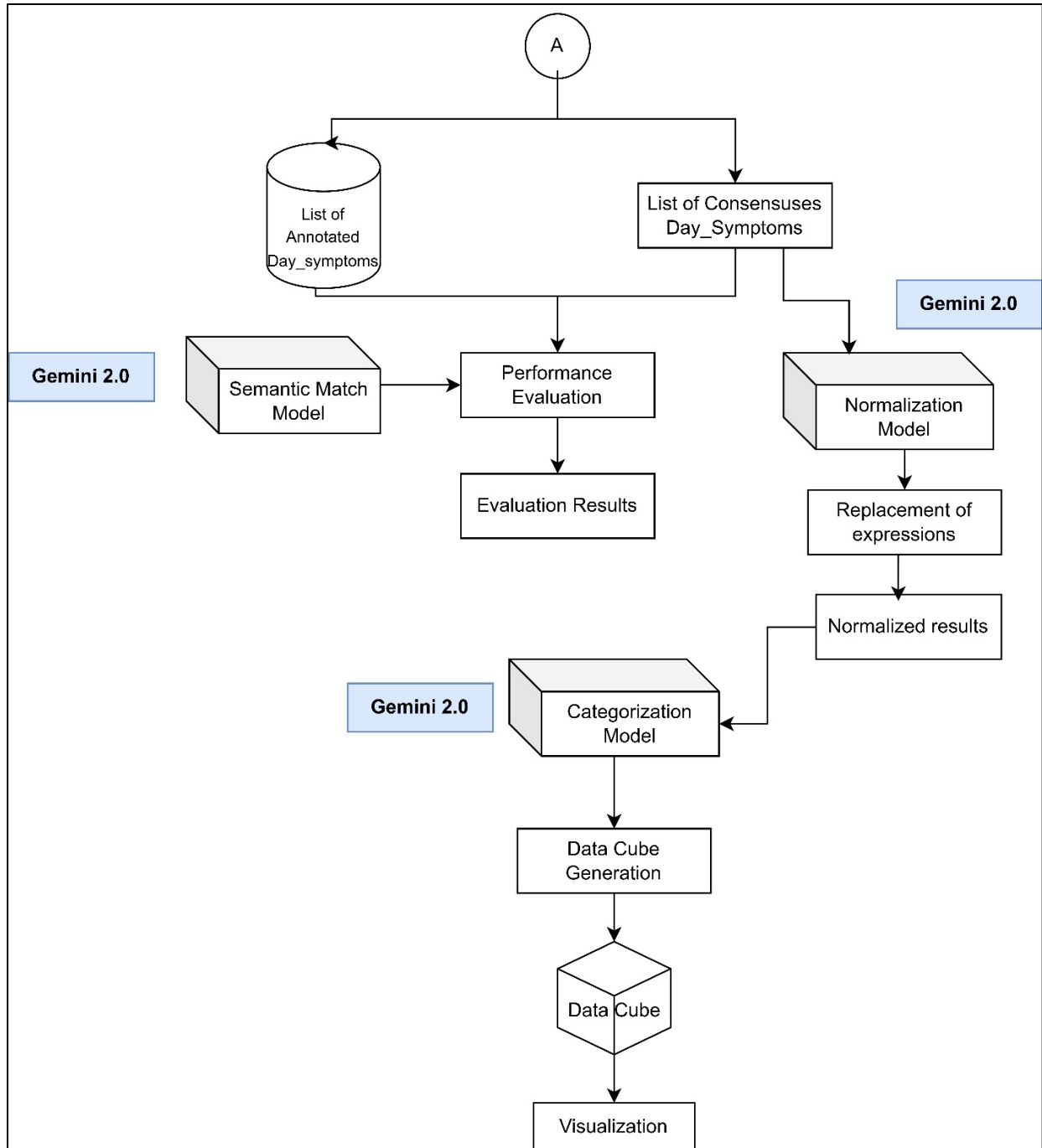


Figure 2. After Extraction Processing: Semantic Matching, Normalization, Categorization, and Visualization

10.2 Data Collection and Preprocessing

The dataset used in this study originates from prior work by Dr. Keyuan Jiang and colleagues, as documented in the publication “Detection of Day-Based Health Evidence with Pretrained Large

Language Models: A Case of COVID-19 Symptoms in Social Media Posts” (Jiang et al., 2024). In that study, a large-scale corpus of 12 million COVID-19-related tweets was collected from Twitter between March and April 2020 using a custom-built web crawler. The crawler operated in compliance with Twitter’s data usage policies and used domain-relevant keywords such as “covid19,” “COVID-19,” “coronavirus,” “Wuhan pneumonia,” and “nCoV” to retrieve tweets related to the ongoing health crisis.

Subsequent filtering steps focused on improving the relevance and usability of the dataset. Tweets were deduplicated, retweets were excluded, and only English-language posts were retained. A key aspect of the filtering process involved identifying tweets that reflected personal health experiences, as opposed to news, jokes, or commentary. This was accomplished using techniques described in earlier work on tweet classification. MetaMap Lite was then applied to ensure that only tweets containing at least one medically recognizable symptom or sign were included. To capture the temporal dimension of symptom expression, regular expressions were used to extract tweets containing day references such as “day 1,” “day1,” or “first day.” These filtering stages yielded a final dataset of 635 tweets, each manually annotated by two authors for day-specific symptom mentions. Annotations followed a guideline where the shortest span of text representing a symptom concept was extracted, with temporal references aligned accordingly.

In this thesis, the pre-annotated dataset of 635 tweets was used as the sole input for all model evaluations. Prior to LLM-based processing, a light round of data cleaning was conducted to ensure consistency and formatting alignment. This included correcting common misspellings, removing emojis and non-standard Unicode characters, and standardizing the representation of day references to align with prompt expectations (e.g., “first day” was mapped to “day 1”). The cleaned data was formatted into structured JSON objects to support schema-based prompting for the extraction models.

No additional tweet filtering, annotation, or dataset augmentation was performed. This decision was made to preserve the integrity of the ground truth annotations and ensure consistency across model comparisons. All subsequent processing steps, including extraction, consensus generation, normalization, and evaluation, were conducted using this fixed and pre-validated dataset.

10.3 Ground Truth Annotation

The dataset used in this study includes manual annotations that serve as the ground truth for evaluating model performance. These annotations were originally developed as part of the research conducted by Jiang et al. (2024), where two authors independently annotated each of the 635 study tweets for day-specific symptom mentions. The goal of the annotation process was to identify and extract the shortest span of text that clearly conveyed a symptom concept, while also linking it to the corresponding day referenced within the tweet.

Annotation guidelines followed a symptom-centric approach based on previously established health information extraction frameworks. A symptom was defined as any physical or mental condition reported by the tweet’s author that could be interpreted as an indicator of illness. To maintain semantic accuracy and consistency, annotations focused on medically interpretable expressions rather than figurative or emotional language. For instance, in a tweet containing the phrase “some pain on breathing” and a temporal reference to “day 4,” the annotated output would extract “pain” and associate it with day 4. Temporal phrases such as “today,” “yesterday,” or “on the second day” were normalized by the annotators into cardinal day values (e.g., “day 1,” “day 2”).

These human annotations serve as the benchmark against which all model-generated outputs are compared in this study. They provide both the symptom list and the day alignment necessary for evaluating the accuracy, precision, recall, and F1-score of LLM-based extraction pipelines. The dataset reflects a high degree of annotation quality, having been developed under close supervision and following consistent guidelines. No modifications were made to the annotations during the course of this study, ensuring that the evaluation remains grounded in validated expert judgments.

By relying on this gold-standard set of annotations, the study is able to quantify the effectiveness of each model configuration in extracting day-based health evidence from noisy and informal Twitter data.

10.4 Baseline Model Using Schema-Guided Extraction

he baseline model in this study was implemented using GPT-4-0613 accessed via LangChain, which supports structured output formatting through automatic prompt generation based on a user-defined schema. This setup allowed for deterministic, reproducible extractions of symptoms and their associated day references from each tweet. The model was executed once per input, with no sampling or variation, serving as a reference for evaluating the impact of more complex multi-run strategies used in the experimental pipeline.

LangChain enabled the definition of a simple JSON schema that specified the expected structure of the output. The schema contained two fields: "day", represented as a string, and "symptom_list", an array of strings. This schema was passed to LangChain's `create_extraction_chain()` function, which internally constructed the prompt and managed the formatting of the model's output. The schema (code snippet) used is shown below:

```
schema = {  
  "properties": {  
    "day": {"type": "string"},  
    "symptom_list": {  
      "type": "array",  
      "items": {"type": "string"}  
    }  
  }  
}
```

Code Snippet JSON schema used for symptom and day extraction in the baseline model (GPT-4-0613 via LangChain).

This schema ensured that each output was constrained to a consistent and machine-readable format. Unlike the experimental models described in the next section, no custom-written system instructions, natural-language prompts, or filtering criteria (such as ignoring metaphorical or emotional language) were added manually. The structure of the output was fully determined by

the schema, and LangChain’s internal mechanics handled prompt formatting and parsing without user intervention.

The result was a controlled, schema-constrained baseline that enabled consistent downstream evaluation of extraction performance, normalization, and categorization. This baseline served as a benchmark for comparison with the more dynamic, prompt-based configurations of GPT-4 Turbo and Gemini 2.0 Flash.

10.5 Prompt Engineering Strategy

To ensure consistency, interpretability, and comparability across models and downstream tasks, this study adopted a deliberate prompt engineering strategy tailored to each stage of the pipeline. While the baseline model—GPT-4-0613 via LangChain—relied on a schema-based approach, the experimental models (GPT-4 Turbo and Gemini 2.0 Flash) used natural-language prompts combined with structured output constraints to extract and analyze health-related information from tweets. The same prompt was used across multiple stages to ensure that model behavior was not influenced by variations in wording or format.

Table 2 summarizes the use of prompt-based instructions across the primary tasks and models in this study.

Table 2. Use of Prompt-Based Instructions Across Tasks and Models

	Task	GPT-4-0613 via LangChain	GPT-4 Turbo	Gemini 2.0 Flash
	Extraction	✗	✓	✓
Gemini 2.0 Flash	Semantic Match	✓	✓	✓
	Normalization	✓	✓	✓
	Categorization	✓	✓	✓

Legend: ✓ = Prompt used; ✗ = No prompt used

Note: Gemini 2.0 Flash was used for all after extraction steps, regardless of the original extraction model.

10.5.1 Prompt Development and Deployment

Initial prompt design was iterative and experimental. Prompts for the extraction stage were developed through multiple rounds of testing using web interfaces for GPT-4 Turbo and Gemini 2.0 Flash. During this phase, different formulations were explored for clarity, specificity, and alignment with expected outputs in a JSON structure. Once optimal prompts were identified, they were reused consistently across all 635 tweets using API-based execution for each experimental model. This controlled reuse ensured that variations in model behavior could be attributed to the model itself, rather than inconsistencies in prompt design.

For the extraction task, GPT-4 Turbo and Gemini 2.0 Flash both received the same system instruction and user-level prompt, which instructed them to extract symptoms and align them with the referenced day. In contrast, GPT-4-0613 was constrained via a schema passed through LangChain, without a manually authored prompt.

Following extraction, the next three tasks—semantic match, normalization, and categorization—were all performed using Gemini 2.0 Flash. For these post-extraction tasks, the same prompt and system instruction were applied across all datasets, regardless of whether the original extracted outputs came from GPT-4-0613, GPT-4 Turbo, or Gemini 2.0 Flash. This use of Gemini 2.0 Flash as a unifying post-processing engine was designed to ensure consistency in semantic interpretation and classification. At each step, the model was guided using carefully worded prompts to produce structured, interpretable outputs aligned with ground truth annotations.

10.5.2 Worked Example: Semantic Match Prompting

To illustrate how semantic matching was applied after extraction, Table 3 shows the process for a representative tweet. Here, Gemini 2.0 Flash was used to compare extracted symptom expressions from all three models against the gold-standard human annotation.

System Instruction

You are a great linguist capable of finding semantically similar terms. Output your result in valid JSON format with the similar terms in the first list as the keys and the corresponding terms in the second list as the values.

Prompt Template

Without using your own knowledge, please find semantically similar terms that ONLY appear in both lists below: {ground truth list} and {extracted symptom list}

Raw Tweet

"Self isolation day 1. Cough is productive and nasty. Sore throat wish it would just calm down. Temp keeps going up and down. Not confirmed as covid19 as I haven't been swabbed. 6 more days of self isolation. Worst thing is, I HAVE BRITTLE ASTHMA."

Ground Truth (Human Annotation)

```
{  
  "Day 1": ["cough", "sore throat", "temp keeps going up and down"]  
}
```

Table 3. Semantic Match Results for a Single Tweet Across Models

Model	Extracted Symptoms	Semantic Match Output
GPT-4-0613 (LangChain)	["Cough is productive and nasty", "Sore throat", "Temp keeps going up and down", "BRITTLE ASTHMA"]	{ "cough": "Cough is productive and nasty", "sore throat": "Sore throat", "temp keeps going up and down": "Temp keeps going up and down" }
GPT-4 Turbo	["Cough", "Sore throat", "Temp keeps going up and down"]	{ "cough": "cough", "sore throat": "sore throat", "temp keeps going up and down": "temp keeps going up and down" }
Gemini 2.0 Flash	["cough is productive", "brittle asthma", "temp keeps going up and down", "sore throat"]	{ "cough": "cough", "sore throat": "sore throat", "temp keeps going up and down": "temp keeps going up and down" }

As shown above, the prompt-driven semantic match model was able to correctly align extracted expressions to the ground truth in each case. Gemini 2.0 Flash identified semantically equivalent pairs even when the phrasing was informal or slightly varied, such as recognizing “Cough is productive and nasty” and “cough is productive” as valid expressions of the normalized symptom

“cough.” In all three cases, the model successfully excluded non-health-related or irrelevant terms like “BRITTLE ASTHMA,” which was not included in the gold-standard annotation.

The success of this strategy highlights the importance of consistent, well-engineered prompts not only for extraction but also for semantic reasoning and interpretation tasks. It also reinforces the value of using a common downstream model—Gemini 2.0 Flash—to normalize post-extraction behavior and ensure fair comparison across all input sources.

All prompts and system instructions used at each stage of the pipeline are included in Appendix A for full transparency and reproducibility.

10.6 Multi-Run Execution and Consensus Selection

One of the challenges inherent to large language models (LLMs) is their non-deterministic behavior—identical prompts can produce different outputs across multiple executions. This randomness, driven by sampling parameters and internal decoding variability, poses a challenge for reproducibility, especially in sensitive applications like public health informatics. When using models such as GPT-4 Turbo and Gemini 2.0 Flash, variability can result in inconsistencies in extracted symptom mentions, even when the input tweet and prompt remain unchanged.

To address this, the framework adopted a manual ensemble and consensus strategy. Each tweet was processed three separate times by both GPT-4 Turbo and Gemini 2.0 Flash using the same prompt and system instruction. This was done intentionally to surface the range of possible outputs generated by each model. The output sets were then manually compared, and a consensus output was selected based on semantic alignment and frequency. This step was not automated but instead performed manually to maintain precise control over judgment criteria, allowing for nuanced evaluation of model behavior across runs.

Importantly, GPT-4-0613 via LangChain did not undergo this multi-run execution. As a schema-driven baseline model, its outputs were deterministic and did not require repetition for stabilization. In contrast, the experimental models required consensus selection to mitigate the unpredictability

of generative outputs and to ensure downstream tasks were built on the most representative extractions.

Table 4 presents an illustrative example of this process. It shows how Gemini 2.0 Flash produced three slightly different outputs for a single tweet. Although the core symptoms are consistent, the example also demonstrates capitalization drift and the occasional omission of a symptom ("sleep"), underscoring the need for manual consensus.

Raw Tweet 2:

"It started with a headache, feeling achy and tired. Coughing and a sore throat. Fever spiked to 103. Hard to breathe. So tired all I want to do is sleep.. -COVID19? JOURNAL DAY 1"

Human Annotation (Ground Truth):

```
{
  "Day 1": ["headache", "achy", "tired", "coughing", "sore throat", "fever", "hard to breathe",
"sleep"]
}
```

Table 4. Example of Multi-Run Output and Manual Consensus

Tweet #	Day #	Symptoms				
		Gemini2.0 Output 1	Gemini2.0 Output 2	Gemini2.0 Output 3	Gemini2.0 Consensus	Human Annotation
2	1	headache	headache	headache	headache	headache
2	1	feeling achy	achy	achy	achy	achy
2	1	tired	tired	tired	tired	tired
2	1	coughing	coughing	Coughing	coughing	coughing
2	1	sore throat	sore throat	sore throat	sore throat	sore throat
2	1	Fever	Fever	Fever	Fever	fever
2	1	Hard to breathe	Hard to breathe	Hard to breathe	Hard to breathe	Hard to breathe
2	1	sleep	sleep		sleep	

As the table 3 shows, while the model captured nearly all symptoms consistently, minor variations such as inconsistent casing (e.g., “Fever” vs. “fever”) and omission (absence of “sleep” in Output 3) demonstrate the subtle inconsistencies that can affect downstream reliability. The consensus output—assembled manually—represented the most semantically complete and accurate representation for downstream tasks such as semantic matching, normalization, and categorization.

By incorporating multi-run output generation and consensus refinement, this study ensured that evaluations were grounded in stable, contextually validated outputs. This manual approach, though time-intensive, added transparency and interpretive rigor to the use of generative models in health symptom extraction pipelines.

10.7 Performance Evaluation

The goal of the performance evaluation phase was to assess how accurately each model in the framework extracted symptom mentions from Twitter data compared to the human-annotated ground truth. This evaluation was conducted after the extraction stage and used a semantic similarity model—Gemini 2.0 Flash—with a consistent prompt to compare extracted outputs to annotated symptoms. By performing semantic rather than surface-level comparisons, the evaluation accommodated minor differences in phrasing (e.g., "burning up" vs. "fever") and accounted for informal or variable expressions typical of user-generated content.

Each model’s output was evaluated using standard confusion metrics: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). True Positives referred to the number of symptom mentions that were correctly extracted and matched the ground truth. False Positives were those symptom mentions generated by the model that were not part of the annotated ground truth. Conversely, False Negatives were those symptoms present in the ground truth but missing from the model’s output. In this task, True Negatives could not be directly measured because the domain is unbounded—there is no comprehensive list of "non-symptoms" in each tweet. In traditional classification settings, True Negatives represent correctly identified absences of a class. However, in this symptom extraction setting, it is infeasible to determine all the symptom expressions that were rightly ignored by the model. Therefore, while True Negatives are

included in the standard metric definitions for completeness, they were not used in calculations for this task.

The following standard formulas were applied to quantify model performance:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1 Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Since TN was undefined in this context, accuracy was interpreted with caution, and greater emphasis was placed on precision, recall, and F1 score—metrics that better reflect model performance in unbalanced, span-based information extraction tasks.

During evaluation, a small but important phenomenon—hallucination—was considered. Hallucinations refer to symptom mentions generated by the model that are not present in the tweet, either explicitly or implicitly. These outputs are typically the result of the model generalizing from prior knowledge rather than tweet-specific evidence. Although no quantitative count is reported here, such hallucinations were occasionally observed during manual review. When identified, they were counted as false positives in the confusion matrix, helping to reflect the impact of generative overreach in the overall performance metrics.

To ensure reliable evaluation, this performance assessment was applied to each model’s final output. For the experimental models (GPT-4 Turbo and Gemini 2.0 Flash), the outputs used were those selected via the consensus mechanism after three separate runs. For GPT-4-0613 via LangChain, a single deterministic output was used, given its schema-enforced extraction behavior. Each tweet’s symptoms were compared against the gold-standard annotations on a day-wise basis, and TP, FP, and FN counts were aggregated across all examples before metric computation.

Together, this performance evaluation framework enabled rigorous, transparent, and fair assessment of model behavior, accounting for both variability in LLM outputs and the complexity of symptom expression in informal social media content. Subsequent analysis and visualizations—including confusion matrices and 3D symptom timelines—are presented in the Results chapter.

10.8 Symptom Normalization

After consensus selection, the next step in the pipeline involved symptom normalization, which aimed to standardize informal, creative, or ambiguous symptom expressions into consistent medical terminology. This process was essential because social media posts often describe symptoms using non-clinical language—for instance, “my chest feels like it’s on fire” or “temp going wild again”—which, without normalization, would be treated as distinct from canonical terms like “chest pain” or “fever.”

Normalization was performed using Gemini 2.0 Flash, which processed the consensus outputs of GPT-4 Turbo and Gemini 2.0 Flash, as well as the baseline outputs from GPT-4-0613 via LangChain. The same system instruction and prompt were used across all three sets of model outputs to ensure consistency. The prompt used for normalization was:

Normalize the following symptoms to standard medical terms. Return ONLY a JSON dictionary (no code) where keys are the original symptoms, and values are normalized medical terms. Do not include any code or explanation.

Symptoms:

{unique_symptoms}

This prompt was specifically designed to generate clean, structured output in JSON format. Each key in the dictionary corresponded to an extracted symptom phrase, while the value represented its normalized medical equivalent. For example, a symptom like "Cough is productive and nasty" would be mapped to "cough", and "Body hurts everywhere" to "body ache".

Following normalization, a term replacement step was implemented in Python. This process replaced the original extracted phrases with their normalized counterparts in each model's output. No LLM was used in this step. This ensured that the downstream categorization model would operate on a clean and consistent vocabulary across all inputs.

By normalizing informal expressions to standard terminology and applying these mappings consistently, the pipeline established a uniform dataset that improved interpretability and enabled structured symptom categorization in the following stage.

10.9 Symptom Categorization

Following the normalization step, each symptom was categorized into broader health-related groups using Gemini 2.0 Flash. The purpose of this stage was to structure the extracted symptom data into clinically interpretable clusters, enabling high-level trend analysis and category-based visualization.

The model was guided by a prompt designed to return structured JSON output, with health categories as dictionary keys and lists of corresponding symptoms as values. The same prompt was applied uniformly across the normalized outputs of GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash:

Categorize the following symptoms into common health conditions. Return ONLY a JSON dictionary (no code) where the keys are health categories and the values are lists of symptoms.

Symptoms:

{normalized_symptoms}

Importantly, the categories themselves were not manually specified by anyone but were generated by the LLM based on its understanding of medical terminology and relationships between symptoms. This allowed the model to flexibly define categories such as “Respiratory Infections” or “Systemic Illness/General Symptoms,” depending on the input list.

Because categorization was applied after normalization, the input to this step consisted entirely of consistent and de-duplicated symptom terms. Each model’s output was categorized independently using the same instruction and prompt, ensuring consistency in the task while allowing the LLM to determine category boundaries.

10.10 Data Visualization and Analysis

After the stages of symptom extraction, normalization, and categorization, the final step in the framework involved visualizing the processed data to support interpretation of temporal patterns, symptom category trends, and model-level behavior. This phase was designed not only to present results in an interpretable form but also to enable visual diagnostics that could guide comparative evaluation and inform future public health applications.

The categorized dataset was first structured into a machine-readable format that allowed symptom mentions to be tracked across time. Each normalized symptom was associated with a tweet index and a specific numeric day, as extracted from the text (e.g., “day one,” “day 1,” or “yesterday”). All extracted symptoms were mapped to Days 1 through 14, based on the annotation guidelines used in the original dataset. This time frame aligns with widely accepted clinical and epidemiological understanding of COVID-19, which recognized that symptoms typically emerge and evolve within a 14-day window. The same two-week period was used in public health guidance for quarantine, testing, and recovery timelines. As a result, this window provided a consistent and medically relevant scope for evaluating symptom progression.

Using this structured timeline, occurrences of normalized symptom mentions were computed for each day and grouped by symptom category. This facilitated a longitudinal view of how specific symptoms emerged, peaked, or declined across the 14-day span, as observed through self-reported Twitter data.

To visualize these trends, an interactive three-dimensional cube was constructed. The cube represented the relationship between three variables: the day of symptom reporting (x-axis), the normalized symptom category (y-axis), and the frequency of mentions (z-axis). This design allowed for an intuitive, multidimensional view of how symptoms evolved over time and how their

prevalence varied across categories. Viewers could easily identify clusters of core COVID-19 symptoms (such as fever, cough, or fatigue) and trace how these symptoms shifted in distribution from early to later days.

In addition to the 3D cube, a suite of two-dimensional visualizations was developed to further examine model performance and temporal dynamics. These included heatmaps to highlight the density of symptom mentions per day and category, line graphs tracking symptom trends over time, and bar charts showing model-wise comparisons of symptom counts. Each model's outputs were visualized separately to allow for side-by-side comparison, highlighting differences in extraction patterns across GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash.

To evaluate model accuracy in more detail, confusion matrix plots were generated. These plots captured the distribution of true positives, false positives, and false negatives by symptom category, offering insight into which types of symptoms each model handled well or struggled with. For example, while some models consistently recognized respiratory symptoms like cough or shortness of breath, others were prone to overlooking or hallucinating certain symptom types, particularly under ambiguous phrasing.

These visualizations collectively provided a static but comprehensive overview of symptom reporting patterns across the 14-day annotated timeline. The combination of the 3D symptom cube, category-specific heatmaps, line graphs, and confusion matrix plots enabled in-depth analysis of how symptoms were distributed, how models differed in their extraction patterns, and where performance varied by symptom type. Although this framework was not implemented as a real-time or streaming system, the structured outputs and visualization techniques developed in this study establish a strong foundation for potential integration into future public health monitoring dashboards or syndromic surveillance platforms.

10.11 Implementation Notes and Runtime Considerations

The end-to-end implementation of this research framework was conducted entirely within Google Colab, a cloud-based development environment that offered sufficient flexibility for prototyping and executing Python-based natural language processing pipelines. Colab's hosted runtime

environment provided built-in support for essential libraries, easy integration with APIs, and access to Google Drive for data persistence. Its compatibility with Python 3.10 allowed for seamless development across all components of the pipeline, which included preprocessing, prompt engineering, model inference, semantic matching, normalization, categorization, and visualization.

The framework relied on widely used Python libraries: pandas and NumPy were used for data manipulation and numerical operations, Matplotlib and Plotly supported visualization, and requests, json, and time modules handled API interaction. LangChain was specifically used to manage schema-based prompt execution for GPT-4-0613, allowing for structured output parsing and task alignment with a defined JSON schema. The LLM inference workload was distributed across direct API calls to OpenAI's GPT-4-0613 and GPT-4 Turbo, as well as Google's Gemini 2.0 Flash, each accessed using securely stored API keys loaded from Google Drive during runtime.

The framework followed a modular and function-based design, enabling tweets to be processed independently. Each tweet was passed through both GPT-4 Turbo and Gemini 2.0 Flash three times to account for output variability and allow for consensus generation. In total, 5,715 inference calls were made across all stages of evaluation. These outputs were stored in .json format immediately after generation to prevent redundant calls and facilitate downstream processing without re-invoking LLM APIs. After extraction, Gemini 2.0 Flash was consistently used for semantic matching, normalization, and categorization tasks, leveraging its fast response times and consistent output structure.

Runtime resource demands were low. All processing tasks—including multi-run inference, semantic comparison, and prompt-based transformations—were completed using Colab's default CPU allocation, with no GPU acceleration required. Peak memory usage remained under 1 GB per session throughout the pipeline, confirming the framework's feasibility for lightweight academic or public health deployments in environments with limited computational resources. Each complete processing cycle for a tweet, from inference to categorization, took approximately 18 to 22 seconds depending on the model and response latency.

Costs were incurred primarily through OpenAI’s API services. GPT-4 Turbo calls averaged between \$0.003 and \$0.006 per request, depending on the length and complexity of the prompts and outputs. GPT-4-0613 requests were approximately 20–30% more expensive per call, due to the inclusion of longer system instructions and schema-enforced formatting. Gemini 2.0 Flash was accessed through a free-tier usage plan and did not incur additional charges but was subject to quota limitations based on request volume.

Overall, Google Colab provided a reproducible, cost-effective, and scalable environment for conducting all experiments. Its integration with cloud-based language models, combined with effective caching and modular code design, made it a practical and efficient platform for managing large-scale inference experiments in a structured research workflow.

10.12 Summary

This chapter detailed the methodology used to evaluate large language models for extracting, normalizing, and categorizing COVID-19 symptoms from Twitter data. Beginning with a pre-annotated dataset derived from prior research, the study implemented a modular pipeline incorporating schema-based extraction, prompt-driven inference, multi-run consensus generation, semantic similarity evaluation, normalization, categorization, and visualization.

Three LLM configurations—GPT-4-0613 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash—were assessed across multiple stages of the pipeline. A multi-run strategy was used for the experimental models to address output variability, while structured prompts and instructions ensured consistency across downstream processes. All tasks after extraction, including semantic matching, normalization, and categorization, were executed using Gemini 2.0 Flash with same prompt templates to maintain comparability.

The pipeline’s implementation in Google Colab supported repeatability, efficient model querying, and structured result caching. Visualization tools such as a 3D data cube, heatmaps, and confusion matrices provided insights into symptom trends and model behavior across the 14-day temporal scope of the dataset.

By combining modern LLM capabilities with structured design and careful prompt engineering, this methodology lays the foundation for scalable, reproducible health symptom extraction from informal, user-generated content. The results and analysis derived from this pipeline are presented in the following chapter.

11. RESULTS

11.1 Model Performance Overview

To evaluate model performance in extracting symptom mentions from informal COVID-19-related tweets, outputs were compared against human-annotated ground truth using semantic matching. For GPT-4 Turbo and Gemini 2.0 Flash, each model was executed three times per tweet, and a manual consensus was selected to improve consistency. In contrast, GPT-4 via LangChain, used as the baseline model, was executed once using a structured schema and returned deterministic outputs. After semantic alignment between model outputs and annotations, confusion matrix components—true positives (TP), false positives (FP), and false negatives (FN)—were identified. These values were then used to compute accuracy, precision, recall, and F1 for each model.

Table 5. Confusion Matrix Components (TP, FP, FN)

Model	True Positives (TP)	False Positives (FP)	False Negatives (FN)
GPT-4 via LangChain	1957	379	182
GPT-4 Turbo	1627	361	478
Gemini 2.0 Flash	1077	913	872

This table 5 summarizes the number of symptoms mentions that were semantically matched between the model consensus outputs and human annotations. True positives represent symptom mentions that both the model and the human annotators aligned on, while false positives were symptoms extracted by the model but not present in the ground truth. False negatives refer to annotated symptoms that the model failed to detect.

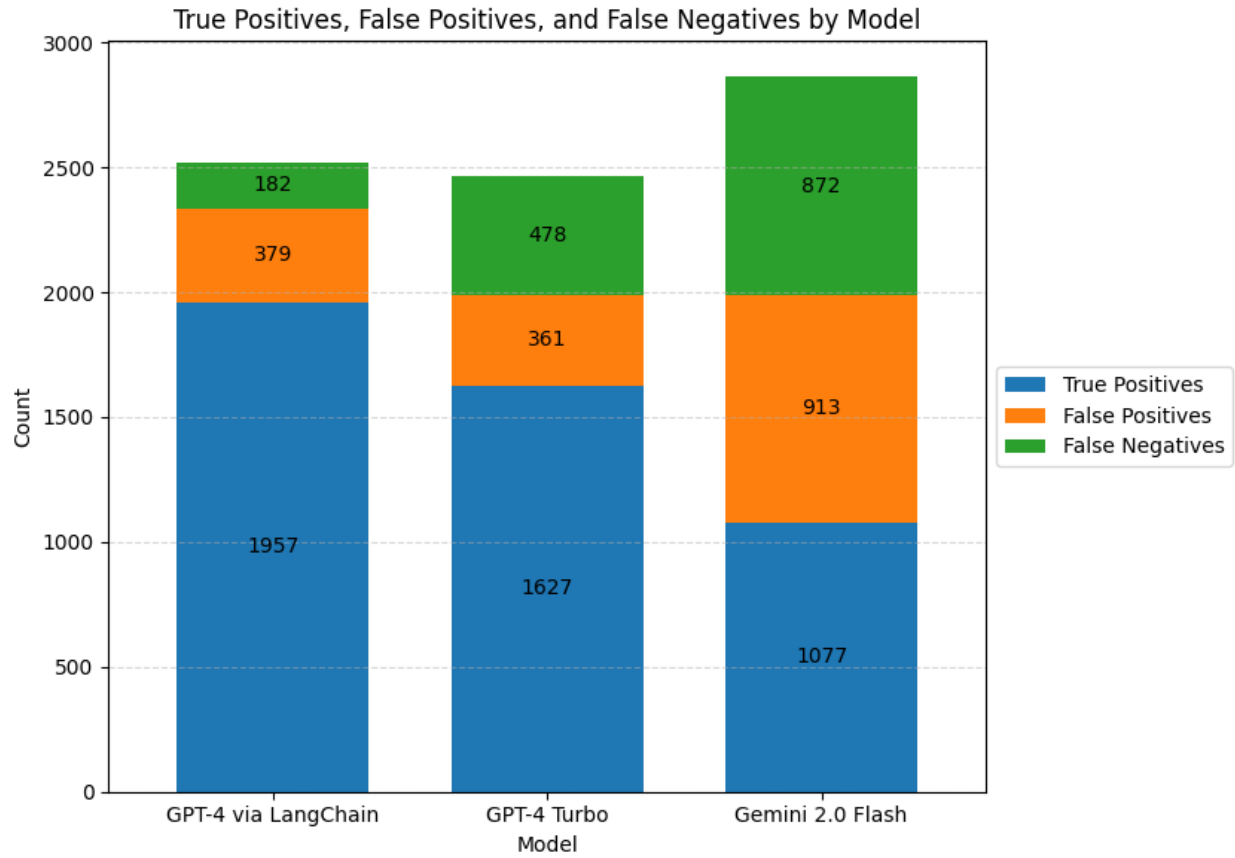


Figure 3. Stacked bar chart showing Confusion Matrix components produced by each model during symptom extraction.

True Positives, False Positives, and False Negatives by Model. This stacked bar chart fig. 3 visualizes the diagnostic contributions of each model to overall performance, revealing significant disparities in false positive and false negative rates.

Table 6. Evaluation Metrics

Model	Accuracy	Precision	Recall	F1 Score
GPT-4 via LangChain	0.78	0.84	0.91	0.87
GPT-4 Turbo	0.66	0.82	0.77	0.80
Gemini 2.0 Flash	0.38	0.54	0.55	0.55

GPT-4 via LangChain delivered the most consistent and accurate results, as reflected in its high evaluation metrics score. Its structured schema-based extraction process likely contributed to its superior performance in identifying symptom mentions that aligned with human annotations. GPT-4 Turbo, while prompt-based, also performed reasonably well, though it missed more true mentions than LangChain. Gemini 2.0 Flash showed weaker performance in extraction, despite being effective in downstream tasks like normalization and categorization. Its relatively low precision and recall suggest difficulty in accurately identifying relevant symptom expressions from informal tweet content.

To better illustrate model-wise differences across individual evaluation metrics, we present separate bar charts for each metric below.

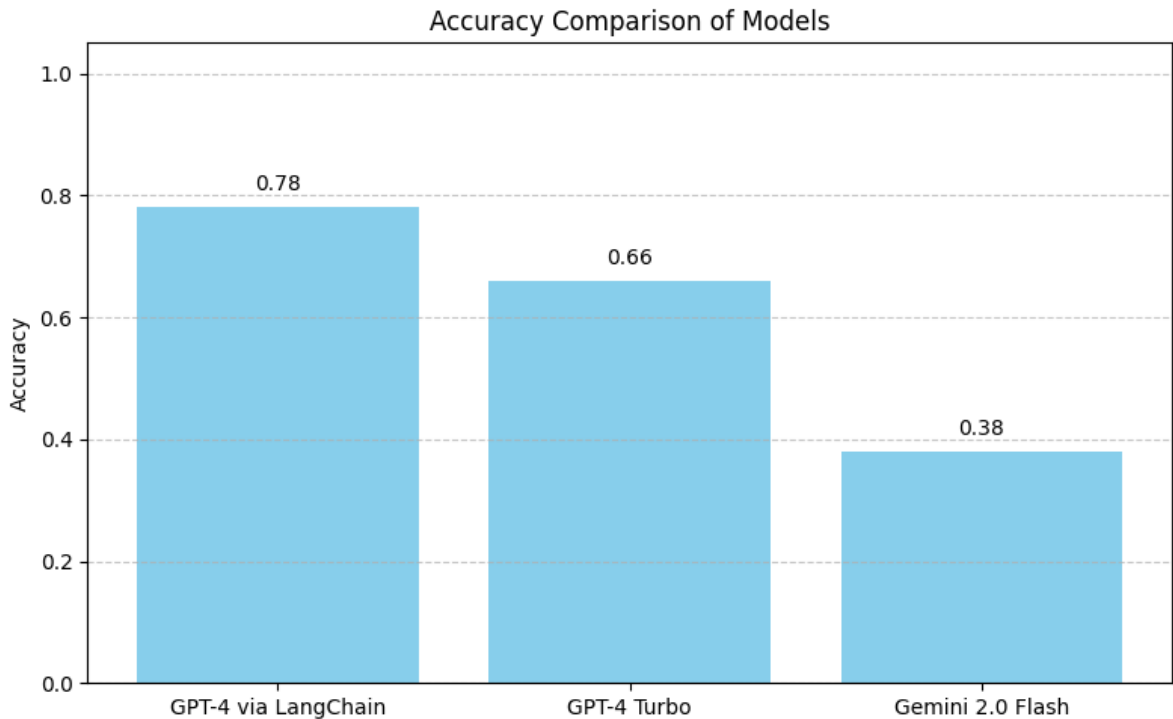


Figure 4. Accuracy Comparison Across Models

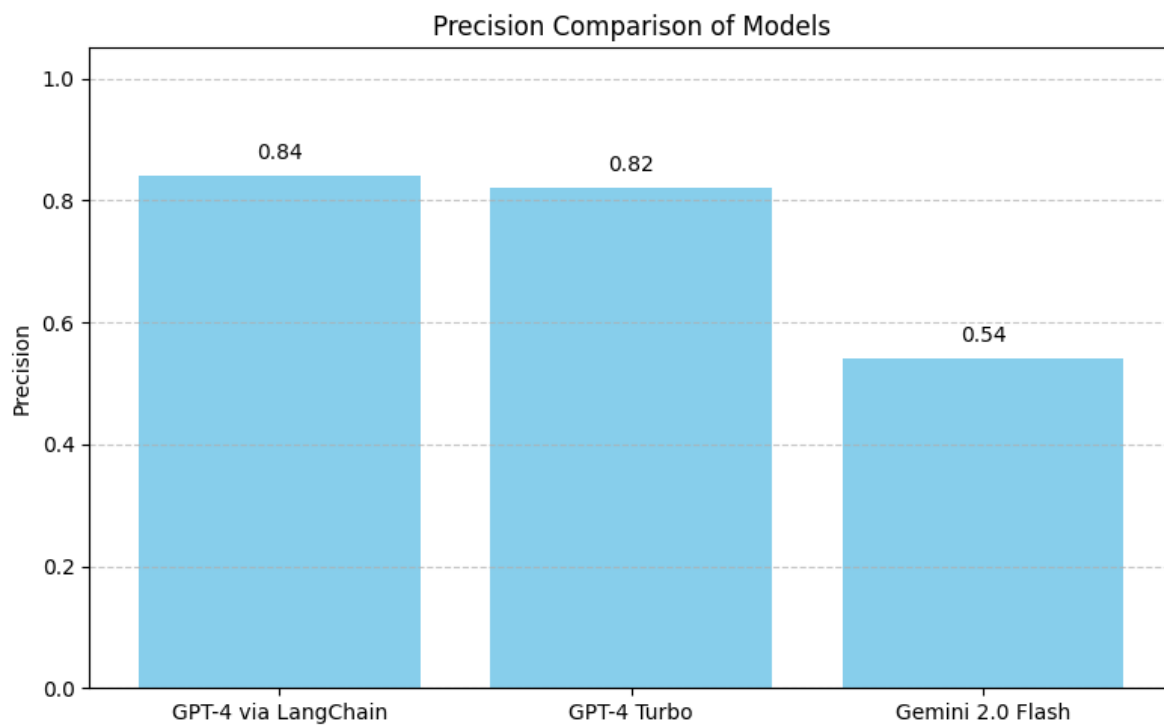


Figure 5. Precision Comparison Across Models

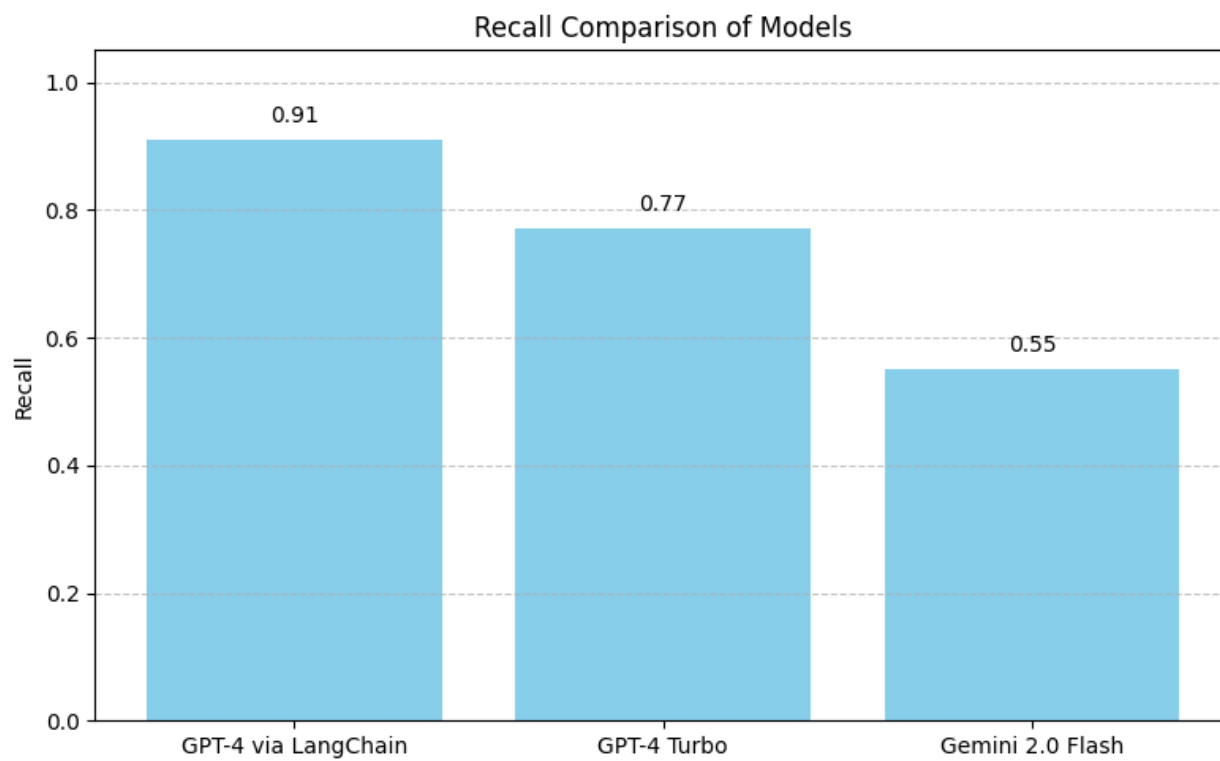


Figure 6. Recall Comparison Across Models

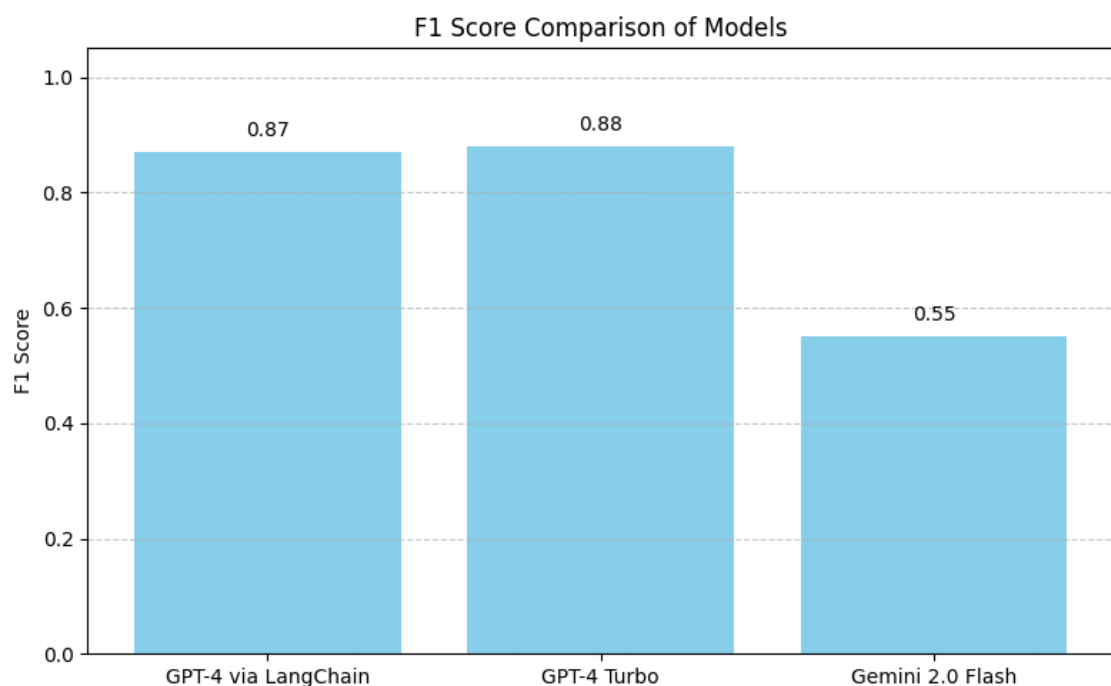


Figure 7. F1 Score Comparison Across Models

Finally, for compact comparison across all metrics, the figure 8 below combines all four scores into a single comparative chart.

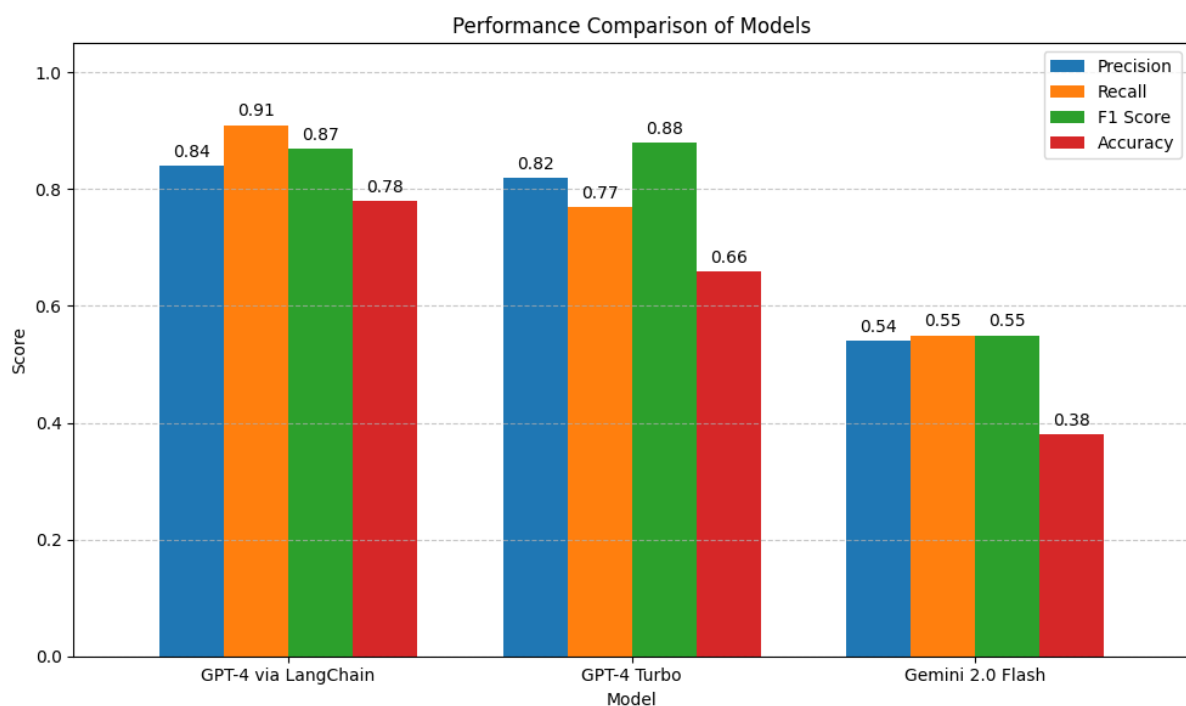


Figure 8. Combined Performance Comparison of All Models

11.2 Symptom Normalization Insights

Following the extraction and semantic alignment stages, each model’s output was passed through a normalization phase to standardize diverse symptom expressions into medically meaningful terms. This process was crucial for enabling structured analysis and cross-model comparability. The normalization was performed using Gemini 2.0 Flash, which mapped colloquial expressions like “cough w phlegm” or “bad stomach ache” to formal terms such as “productive cough” and “abdominal pain,” respectively.

The purpose of normalization was to reduce redundancy and ambiguity across expressions referring to the same clinical symptom. This was particularly important because each model exhibited varying levels of abstraction in their raw symptom mentions. For example, while GPT-4 Turbo might extract “dry cough” and “so tired,” Gemini 2.0 Flash tended to identify similar phrases as “non-productive cough” and “fatigue,” respectively.

To illustrate the diversity and convergence of symptom interpretation across models, Table 7 provides a compact sample of normalized symptom mappings. The examples demonstrate both consistency across models and points of divergence, especially in nuanced expressions where interpretation may vary depending on phrasing and model behavior.

Table 7. Compact Comparison of Normalized Symptom Expressions Across Models

Expression (Example)	GPT-4 via LangChain	GPT-4 Turbo	Gemini 2.0 Flash
“fevers”	Fever	Fever	Fever
“slight dry cough”	Mild cough	Mild dry cough	Mild cough
“cough w phlegm”	cough	—	Productive Cough
“tighter chest”	Chest tightness	Chest tightness	Chest tightness
“feeling achy” / “achy”	Myalgia	Myalgia	Myalgia
“struggled to breath”	Dyspnea	Dyspnea	Dyspnea
“bleary eyes”	Blurred vision	—	Blurred vision
“brittle asthma”	Severe asthma	—	Severe asthma

“—” indicates the expression was not present in that model’s output.

Note: The full list of normalization mappings generated by Gemini 2.0 Flash across all three model outputs is included in Appendix B.

This normalization stage ensured that all subsequent analyses—particularly categorization and visualization—were conducted on a consistent representation of symptoms, eliminating variability introduced by model phrasing or token formatting.

11.3 Categorization Distribution

After the normalization phase, the final step in the framework involved categorizing each symptom into broader health categories to support structured interpretation and downstream analysis. This categorization was performed using Gemini 2.0 Flash, applying a consistent prompt across the normalized outputs of all three models: GPT-4 via LangChain, GPT-4 Turbo, and Gemini 2.0 Flash. Each symptom was mapped to a high-level category such as Respiratory, General, Neurological, or Digestive, based on semantic context.

Figure 9 visualizes the distribution of categorized symptoms generated from the LangChain (GPT-4-0613) output. As shown, the most frequently assigned categories included Respiratory, General, and Fever, followed by Pain and Neurological. Less common categories, such as Ophthalmologic, Integumentary, and Skin, were rarely used, reflecting both the nature of the dataset and the specificity of symptom mentions. The distribution offers insight into the dominant symptom

classes present in COVID-19-related tweets and highlights the capacity of the categorization model to structure noisy health data into interpretable groups.

A summary of the number of categories generated and symptom counts per model is shown below in Table 8:

Table 8. Symptom Category Comparison

Categorize	GPT-4 via LangChain	GPT-4 Turbo	Gemini 2.0 Flash
Respiratory	143	131	156
Pain	89	87	109
Digestive	41	36	37
Fever	61	48	42
Mental Health	31	26	32
Neurological	35	11	26
Cardiovascular	15	4	7
General	75	41	37
Musculoskeletal	8	3	18
Skin	2	7	10
Sensory	5	11	2
Infection	7	0	1
Other	42	38	34
Integumentary	12	0	9
Ophthalmologic	4	0	0
Chest	9	9	0
Urinary	1	0	0
ENT	0	16	0
Temperature	0	2	0
Cardiac	0	2	0
Taste/Smell	0	5	0
Headache	0	1	0
Infectious	0	0	7
Pulmonary	0	0	2
Ophthalmological	0	0	1
General/Systemic	0	0	20
Gynecological	0	0	1
General Symptoms	0	0	3

This variation demonstrates how LLM-driven categorization can be both flexible and dependent on subtle differences in normalized expressions. While the majority of symptoms fell under Respiratory and Systemic categories across all models, finer distinctions such as Vision Problems, Asthma, or Condition Status appeared more frequently in the Gemini-based output, suggesting higher category granularity.

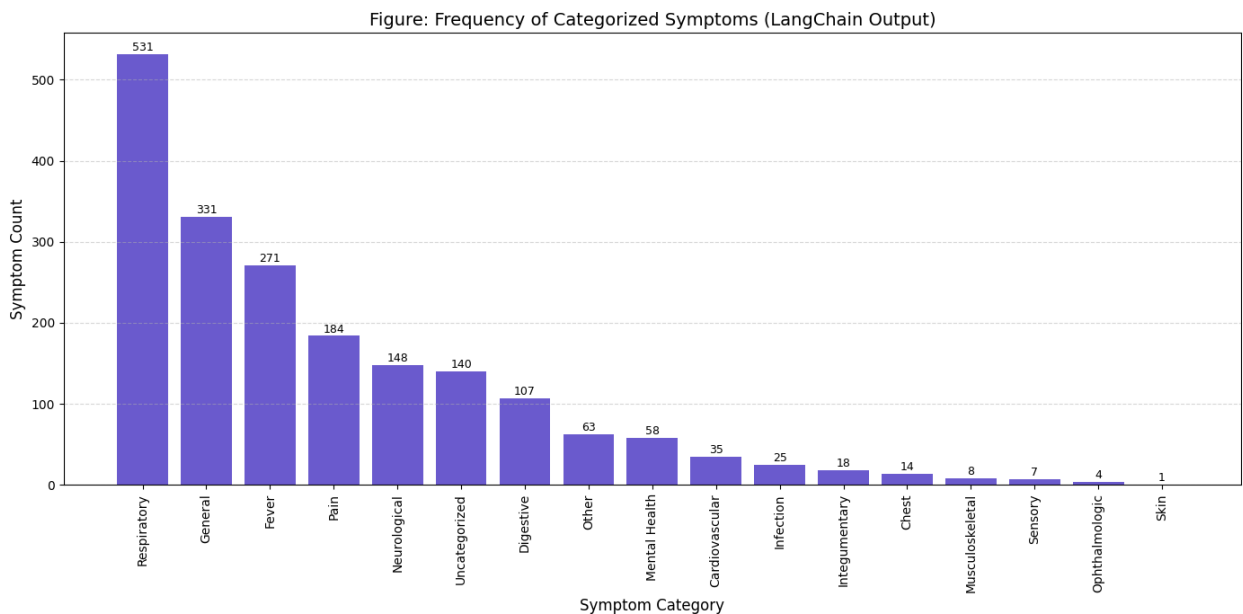


Figure 9. Bar chart showing the total frequency of each symptom category after normalization. Respiratory and systemic symptoms were the most prevalent, consistent with the COVID-19.

Further granularity is shown in Figure 10, a heatmap that visualizes how symptom categories varied over a 14-day timeline. Systemic symptoms were highly concentrated in the initial days (Days 1–3), consistent with the onset phase of many COVID-19 cases. Respiratory symptoms appeared more uniformly throughout the timeline, while neurological symptoms—including “loss of smell,” “headache,” or “brain fog”—tended to peak around Days 4 to 7. Psychological symptoms, although less frequent overall, emerged sporadically and often co-occurred with systemic indicators such as fatigue and low energy.

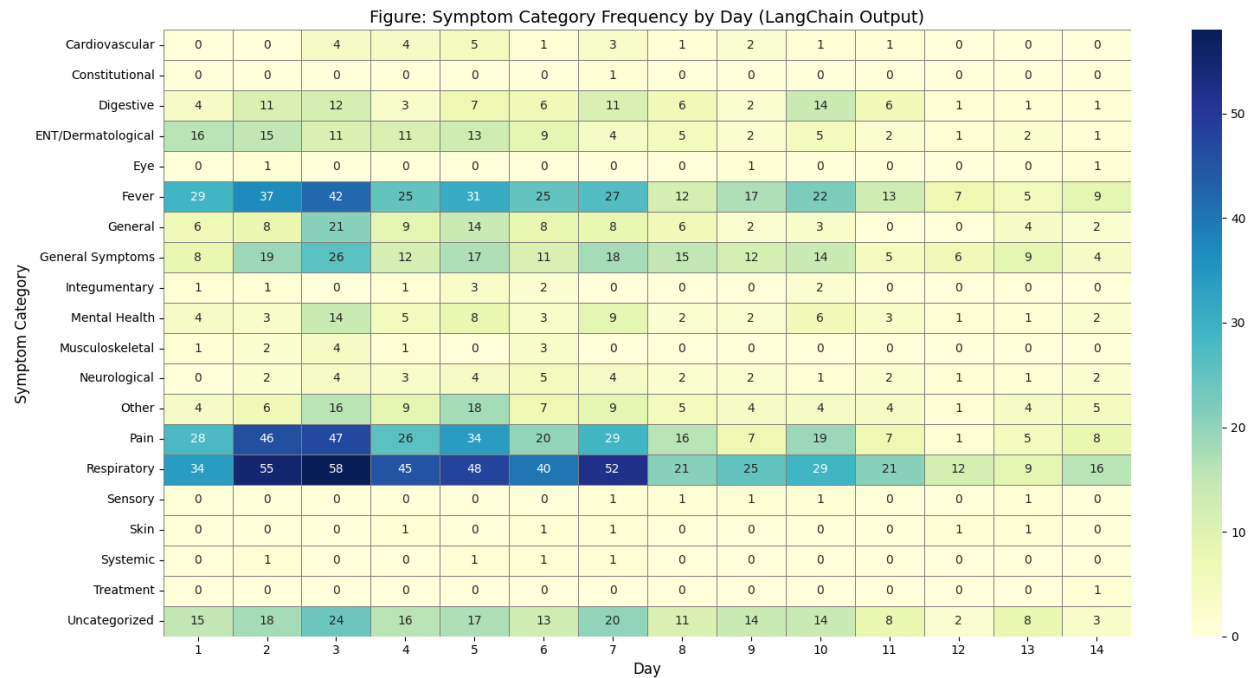


Figure 10. Heatmap of symptom category frequency by day (Day 1 to Day 14). Each cell represents the count of normalized symptoms within a category for that specific day.

These visual and semantic groupings enabled deeper clinical abstraction, allowing for pattern recognition beyond raw symptom terms. For instance, tweets that mentioned both psychological and systemic symptoms often spanned longer timelines and included more complex phrasing, suggesting a potential link between cognitive-emotional stress and symptom reporting behavior. While this behavioral hypothesis falls outside the scope of this study, it introduces opportunities for future interdisciplinary research at the intersection of public health, psychology, and social media analytics.

By transforming noisy, informal tweets into structured symptom categories, the proposed framework demonstrates its potential to support timely health monitoring, pattern detection, and epidemiological insights—especially in contexts where rapid understanding of symptom clusters is critical.

11.4 Multidimensional Visualization – 3D Data cube

The 3D data cube visualization revealed distinct temporal and categorical patterns in the distribution of symptoms across the 14-day observation window. Systemic symptoms such as fever, fatigue, and chills consistently appeared with high frequency during the early days—particularly Days 1 through 3—highlighting their prominence at the onset of illness. Respiratory symptoms, including cough, sore throat, and chest tightness, showed a sustained presence, often peaking between Days 4 and 6, indicating their progression over time.

Gastrointestinal symptoms and neurological indicators like headache and digital pain were less frequently observed but displayed more dispersed timelines, suggesting variability in how and when they were experienced or reported. Certain symptoms, such as wheezing or nasal congestion, appeared sporadically, often concentrated in specific categories and confined to narrower time ranges.

Across all models, the cube highlighted that core COVID-19 symptoms tended to cluster in the first week, with diminishing symptom mentions in the latter half of the 14-day window. This pattern aligns with known clinical symptom trajectories for COVID-19 and supports the relevance of using the 14-day frame. The visualization also showed that models varied in how symptoms were assigned across categories, with some producing broader or more granular groupings—reflected in the density and distribution of the cube’s vertical bars.

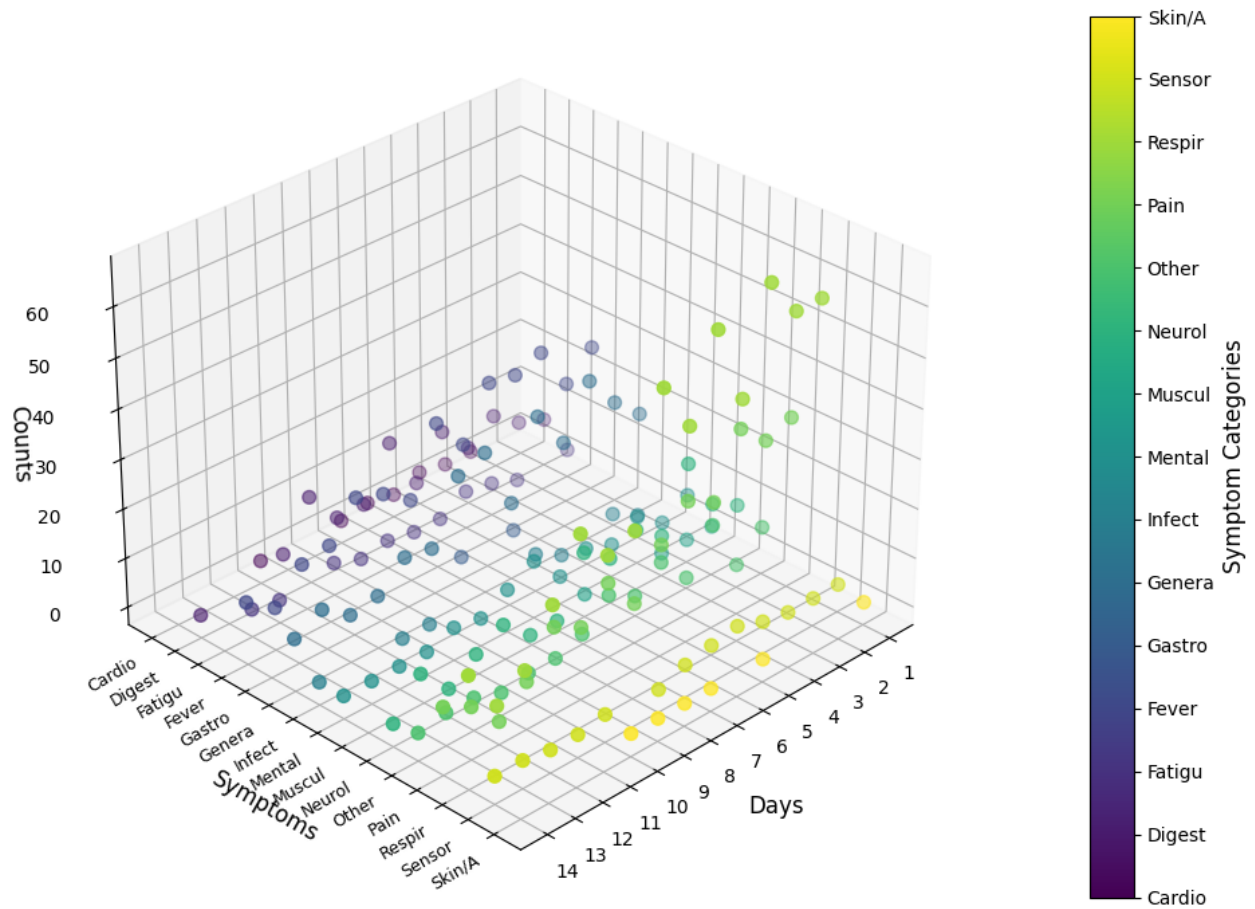


Figure 11. 3D Symptom Cube

Overall, the 3D cube provided a clear, interpretable view of symptom evolution, category overlap, and frequency trends, reinforcing the strengths and limitations observed in earlier model-level comparisons.

12. DISCUSSIONS

This chapter discusses the outcomes of the study in relation to the research objectives and questions. It reflects on model behaviors, the impact of prompt-based design, and the challenges and opportunities in using large language models (LLMs) for health symptom extraction from Twitter data.

12.1 Addressing Research Questions

RQ1: To what extent can large language models accurately extract symptom mentions from informal Twitter data when compared to human-annotated ground truth?

→ GPT-4 via LangChain achieved the highest F1-score (0.87) and recall (0.91), demonstrating strong alignment with human annotations. GPT-4 Turbo showed competitive precision but slightly lower recall, while Gemini 2.0 Flash underperformed in both.

RQ2: How does output variability across multiple runs of the same model affect the consistency and reliability of symptom extraction?

→ GPT-4 Turbo and Gemini 2.0 Flash exhibited variability in extracted symptoms across runs, confirming non-deterministic behavior. The consensus mechanism (manual in this study) effectively stabilized outputs and improved alignment with ground truth.

RQ3: Can post-extraction processing using semantic matching and schema-based normalization improve the interpretability and usability of LLM outputs for public health symptom analysis?

→ Yes. Normalization converted varied expressions like "cough like crazy" and "dry cough" into standardized terms. Categorization further grouped these into health domains, enabling structured interpretation and longitudinal visualizations (e.g., 3D symptom cube).

12.2 Reassessing Limitations in Context

While the methodological limitations of this study were outlined in Chapter 7, it is important to revisit them in light of the empirical findings. Some constraints—such as the modest dataset size, language scope, and use of zero-shot prompting—did not prevent meaningful insights from emerging but did shape the strength and generalizability of the conclusions. Others, such as reliance on a single model (Gemini 2.0 Flash) for post-extraction processing, may have introduced bias in downstream tasks like normalization and categorization. Recognizing these factors helps clarify how the observed model behavior should be interpreted and highlights areas for improvement in future implementations.

12.3 Implications for Public Health Informatics and LLM Applications

The results of this study demonstrate that Large Language Models (LLMs) can be effectively harnessed for health symptom extraction and categorization from informal social media text, supporting several areas of public health informatics. In particular, the ability of GPT-4 and Gemini 2.0 Flash to interpret diverse symptom expressions, align them with standardized categories, and visualize symptom trajectories over time suggests their potential for supplementing syndromic surveillance systems. Platforms such as Twitter, though informal and noisy, offer early insight into self-reported symptoms from users who may not be represented in formal health systems (Sinnenberg et al., 2017; Sarker et al., 2022). LLM-driven pipelines such as the one developed in this study can help structure this information, providing timely input to decision-making processes during outbreaks.

From an application standpoint, this work highlights the need for consistency and structure when integrating LLMs into public health workflows. Prompt-based extraction combined with post-hoc normalization and categorization enables alignment with epidemiological vocabularies, which is essential for cross-study comparability and downstream analysis. While traditional methods often require retraining or heavy annotation, the zero-shot and few-shot capabilities of LLMs offer a flexible and lower-barrier alternative—particularly useful during the early stages of emerging public health crises.

However, reliance on a single model (e.g., Gemini 2.0 Flash) for all post-extraction tasks also introduces potential homogeneity and bias in downstream outputs. As public health tools become increasingly AI-driven, maintaining model diversity and ensuring transparency in model selection and prompt design will be essential (Guo et al., 2024). Furthermore, ethical considerations must be integrated into system design, particularly around privacy, consent, and the risk of amplifying misinformation from noisy data sources (Chancellor et al., 2019; Charles-Smith et al., 2015).

Ultimately, the framework proposed in this study can serve as a foundation for timely, LLM-assisted monitoring of health trends across social media platforms. With further refinement, scaling, and validation on multilingual and diverse datasets, such systems could be integrated into broader digital epidemiology efforts to enhance responsiveness and granularity in public health surveillance.

While the current study demonstrates practical feasibility and clear utility in health surveillance settings, several avenues remain open for refining and extending this work. The following section outlines key directions for future research in symptom extraction, prompt engineering, and large language model evaluation.

12.4 Opportunities for Future Research

Although this framework demonstrates the potential of large language models (LLMs) for symptom extraction and analysis from social media data, several areas warrant further investigation to improve consistency, adaptability, and applicability.

First, future studies could explore integrating multi-lingual and cross-platform datasets to expand the generalizability of the approach. The current work is limited to English-language tweets during the early COVID-19 period, but similar pipelines could be adapted for other languages and sources such as Reddit, health forums, or WhatsApp public channels. Leveraging multilingual LLMs, such as mT5 or XLM-R, may enable broader population coverage in public health surveillance.

Second, while this research used zero-shot and few-shot prompting strategies, further gains may be achieved through fine-tuning or instruction tuning on health-specific corpora. Domain-adapted

LLMs—such as those fine-tuned on PubMed, MIMIC, or health-focused social media data—have shown improved performance in classification and entity extraction tasks (Guo et al., 2024; Tinn et al., 2023). Combining fine-tuning with prompt engineering could yield more reliable outputs, especially for ambiguous or indirect symptom expressions.

Third, future work should examine real-time deployment potential. While this study used a static, pre-annotated dataset, integrating live social media feeds would enable ongoing syndromic surveillance. This would require the development of robust filtering mechanisms for identifying relevant personal health narratives in high-noise environments. Previous efforts in real-time event detection from Twitter (Sinnenberg et al., 2017; Signorini et al., 2011) offer valuable starting points.

Fourth, the framework could be extended to support temporal progression modeling beyond discrete day mappings. Integrating LLM outputs with time series models or neural sequence tagging frameworks may help capture symptom trajectories or identify deviations indicative of long COVID, relapse, or comorbidity onset. Efforts like MedTime (Lin et al., 2013) demonstrate how temporal parsing can be enhanced with hybrid models.

Finally, from an evaluation standpoint, future research should move beyond token-level metrics to incorporate clinical relevance and downstream utility. This includes user-centric metrics like interpretability, system latency in real-time use, and robustness to evolving language trends. Additionally, incorporating human-in-the-loop evaluation with public health experts could help validate outputs beyond automated metrics like F1-score or semantic similarity.

Collectively, these directions reflect an ongoing need to build more adaptive, interpretable, and real-world-ready tools for extracting health intelligence from informal online discourse.

13. CONCLUSION

This thesis set out to explore how large language models (LLMs) can be used to extract, normalize, and categorize health-related symptom information from informal Twitter data. Recognizing the limitations of traditional natural language processing (NLP) methods in handling unstructured, colloquial, and context-dependent text, the study aimed to develop and evaluate a robust pipeline that integrates prompt engineering, semantic validation, and structured post-processing for public health informatics.

The research successfully demonstrated that LLMs, particularly GPT-4-0614 via LangChain, offer a promising solution for symptom extraction from social media. Through multi-run execution and consensus-based decision-making, the system addressed the inherent variability in generative model outputs. The application of semantic matching and normalization, followed by categorization into clinically meaningful symptom groups, enabled more consistent and interpretable analysis. Comparative evaluation using accuracy, precision, recall, and F1 score metrics confirmed that the LangChain-based GPT-4 model outperformed GPT-4 Turbo and Gemini 2.0 Flash in aligning with human-annotated ground truth.

Beyond model performance, the study contributed a methodological framework that incorporates multi-stage validation, schema-guided prompting, and temporal mapping of extracted data. The 3D data cube visualization introduced in this research provided an intuitive and multidimensional perspective on symptom trends over a 14-day period, enhancing the interpretability of findings and supporting epidemiological insights.

The implications of this work extend into multiple domains. For public health, the pipeline offers a scalable, adaptable, and language-aware approach for monitoring disease symptoms in real-time digital environments. For the field of NLP, it showcases the value of structured prompting, semantic consensus, and downstream categorization in converting noisy language into actionable information. The study also highlights how LLMs can move beyond passive text generation to serve as active participants in analytical workflows, provided they are guided by clear schemas and decision logic.

Several limitations were acknowledged and addressed throughout the thesis, including reliance on English-only data, constraints in model interpretability, and the static nature of the dataset. These limitations suggest important directions for future research, including multilingual support, live streaming analysis, hybrid rule-based and neural systems for normalization, and deeper exploration of comorbid symptom clusters.

In conclusion, this research provides strong evidence that LLMs can play a pivotal role in next-generation public health surveillance systems, particularly when deployed through thoughtfully engineered pipelines that combine flexibility with structure. By bridging the gap between informal digital expression and formal clinical categorization, the framework proposed in this study opens up new possibilities for leveraging artificial intelligence to support health monitoring, crisis response, and epidemiological research in a connected world.

APPENDIX A. PROMPT TEMPLATES

This appendix presents the exact prompt templates used to guide large language models (LLMs) in the tasks of entity extraction, semantic matching, normalization, and categorization from social media data (Twitter). These prompts were specifically crafted to ensure structured outputs, reduce model hallucination, and maintain alignment with human annotation guidelines.

Entity Extraction Prompt for GPT-4-Turbo and Gemini 2.0 Flash

Extract the relevant entities mentioned in the following passage together with their properties in JSON format.

Only extract the properties mentioned in the schema of:

```
{
  "properties": {
    "day": {"type": "string"},
    "symptom_list": {
      "type": "array",
      "items": {
        "type": "string"
      }
    }
  }
}
```

For day entity, extract and convert the property value as a cardinal number.

For symptoms, extract the exact but shortest phrases in the passage.

Passage: {tweet}

Semantic Match Prompt

Without using your own knowledge, please find semantically similar terms that **ONLY** appear in both lists below:

['ground truth symptom 1', 'ground truth symptom 2', ...]

and

['extracted symptom 1', 'extracted symptom 2', ...]

Normalization Prompt

Normalize the following symptoms to standard medical terms.

Return **ONLY** a JSON dictionary (no code) where keys are the original symptoms, and values are normalized medical terms. Do not include any code or explanation.

Symptoms:

["temp", "feverish", "my head hurts", "can't breathe"]

Categorization Prompt

Categorize the following symptoms into simple health categories (e.g., Respiratory, Pain, Digestive, Fever, Mental Health, etc).

Provide only a JSON dictionary where the keys are categories and the values are lists of symptoms:

Symptoms:

["fever", "chest pain", "diarrhea", "anxiety", "shortness of breath"]

APPENDIX B. FULL SYMPTOM NORMALIZATION MAPPINGS

This appendix presents the complete set of normalized symptom expressions across the outputs of the three evaluated models: GPT-4 via LangChain (Baseline), GPT-4 Turbo, and Gemini 2.0 Flash. The normalization was performed using Gemini 2.0 Flash with a structured prompt to map informal, colloquial, or noisy symptom mentions into standardized clinical terms. Each table below lists the original expressions extracted by each model and their corresponding normalized terms.

GPT-4 via LangChain (Baseline)

```
{
  "bad stomach ache": "Abdominal pain",
  "bleary eyes": "Blurred vision",
  "body aches everywhere": "Myalgia",
  "bone aches": "Ostealgia",
  "brittle asthma": "Severe asthma",
  "chest pain": "Chest pain",
  "cold": "Upper respiratory tract infection",
  "colds": "Upper respiratory tract infections",
  "cough": "Cough",
  "cough is productive and nasty": "Productive cough",
  "cough like crazy": "Severe cough",
  "cough w phlegm": "Productive cough",
  "coughing": "Cough",
  "dry cough": "Nonproductive cough",
  "extreme fatigue": "Fatigue",
  "extremely fatigued": "Fatigue",
  "eyes on fire": "Eye irritation",
  "feeling achy": "Myalgia",
  "feeling nauseous": "Nausea",
  "feels terrible": "Malaise",
  "fever": "Fever",
```

```

"fevers": "Fever",
"fingers feeling pain/ tired": "Digital pain",
"hard to breathe": "Dyspnea",
"headache": "Headache",
"heavy lungs": "Pulmonary congestion",
"hot/cold spells": "Chills",
"lethargy": "Lethargy",
"nasal congestion": "Nasal congestion",
"neck, back, and chest pain": "Neck pain, back pain, and chest
pain",
"persistent cough": "Chronic cough",
"pyrexia": "Fever",
"runny nose": "Rhinorrhea",
"sciatica": "Sciatica",
"slight cough": "Mild cough",
"slight dry cough": "Mild nonproductive cough",
"sneezing": "Sneezing",
"sore throat": "Pharyngitis",
"still d same": "No change in condition",
"struggle to breath": "Dyspnea",
"stuffy nose": "Nasal congestion",
"symptoms may disappear": "Remission of symptoms",
"temp": "Body temperature",
"temp keeps going up and down": "Fluctuating body temperature",
"throat constricted": "Throat constriction",
"tight chest": "Chest tightness",
"tired": "Fatigue",
"want to sleep": "Somnolence",
"whooping": "Whooping",
"you may feel better": "Symptom improvement"
}

```

GPT-4 Turbo

```
{
  "back pain": "Back pain",
  "chest pain": "Chest pain",
  "cough": "Cough",
  "coughing": "Cough",
  "dry cough": "Dry cough",
  "extreme fatigue": "Fatigue",
  "extremely fatigued": "Fatigue",
  "feeling achy": "Myalgia",
  "fever": "Fever",
  "fevers": "Fever",
  "hard to breathe": "Dyspnea",
  "headache": "Headache",
  "heavy lungs": "Pulmonary congestion",
  "hot/cold spells": "Chills",
  "lethargy": "Lethargy",
  "nasal congestion": "Nasal congestion",
  "neck pain": "Neck pain",
  "painful": "Pain",
  "persistent cough": "Chronic cough",
  "pyrexia": "Fever",
  "runny nose": "Rhinorrhea",
  "sciatica": "Sciatica",
  "slight cough": "Mild cough",
  "slight dry cough": "Mild dry cough",
  "sneezing": "Sneezing",
  "so tired": "Fatigue",
  "sore throat": "Sore throat",
  "struggled to breath": "Dyspnea",
  "temp keeps going up and down": "Intermittent fever",
  "tighter chest": "Chest tightness",
```

```
"tired": "Fatigue",
"unproductive": "Nonproductive cough",
"wheezing": "Wheezing"
}
```

Gemini 2.0 Flash

```
{
  "achy": "Myalgia",
  "anxiety": "Anxiety",
  "back pain": "Back pain",
  "bad stomach ache": "Abdominal pain",
  "bleary eyes": "Blurred vision",
  "body aches": "Myalgia",
  "bone aches": "Ostealgia",
  "brittle asthma": "Severe asthma",
  "chest pain": "Chest pain",
  "cold": "Upper respiratory tract infection",
  "colds": "Upper respiratory tract infections",
  "cough": "Cough",
  "cough like crazy": "Severe cough",
  "cough til you wretch": "Cough with emesis",
  "cough w phlegm": "Productive cough",
  "coughing": "Cough",
  "coughing is annoyingly frequent": "Frequent cough",
  "dry cough": "Non-productive cough",
  "extreme fatigue": "Fatigue",
  "eyes on fire": "Eye irritation",
  "fatigued": "Fatigue",
  "feeling like poop": "Malaise",
  "feeling nauseous": "Nausea",
  "fever": "Fever",
  "fevers": "Fever",
}
```

```

"fingers feeling pain/ tired": "Digital pain/fatigue",
"hacking": "Cough",
"hard to breathe": "Dyspnea",
"headache": "Headache",
"heavy lungs": "Pulmonary congestion",
"hot/cold spells": "Chills",
"lethargy": "Lethargy",
"mostly unproductive": "Non-productive",
"nasal congestion": "Nasal congestion",
"neck pain": "Neck pain",
"no fever": "Afebrile",
"painful": "Pain",
"persistent cough": "Chronic cough",
"pyrexia": "Fever",
"runny nose": "Rhinorrhea",
"sciatica": "Sciatica",
"sleep": "Sleep disturbance",
"slight cough": "Mild cough",
"slight dry cough": "Mild non-productive cough",
"sneezing": "Sneezing",
"sore throat": "Pharyngitis",
"stress": "Stress",
"struggled to breath": "Dyspnea",
"stuffy nose": "Nasal congestion",
"temp keeps going up and down": "Intermittent fever",
"throat constricted": "Throat constriction",
"tighter chest": "Chest tightness",
"tired": "Fatigue",
"wheezing": "Wheezing"
}

```

APPENDIX C. 3D CUBE SLICING

This appendix presents 3D slices of the full symptom cube, with each plot isolating one symptom category. These slices highlight the daily frequency of each symptom, derived from large language model (LLM) extractions. Each slice provides a focused view of how a particular symptom was distributed across a 14-day observation window.

Visualization Details:

- X-axis: Day (1–14)
- Y-axis: Symptom category index (fixed per slice)
- Z-axis: Count (frequency of predicted mentions)
- Each figure showcases a vertical slice of the full cube, isolating one symptom for clear temporal analysis

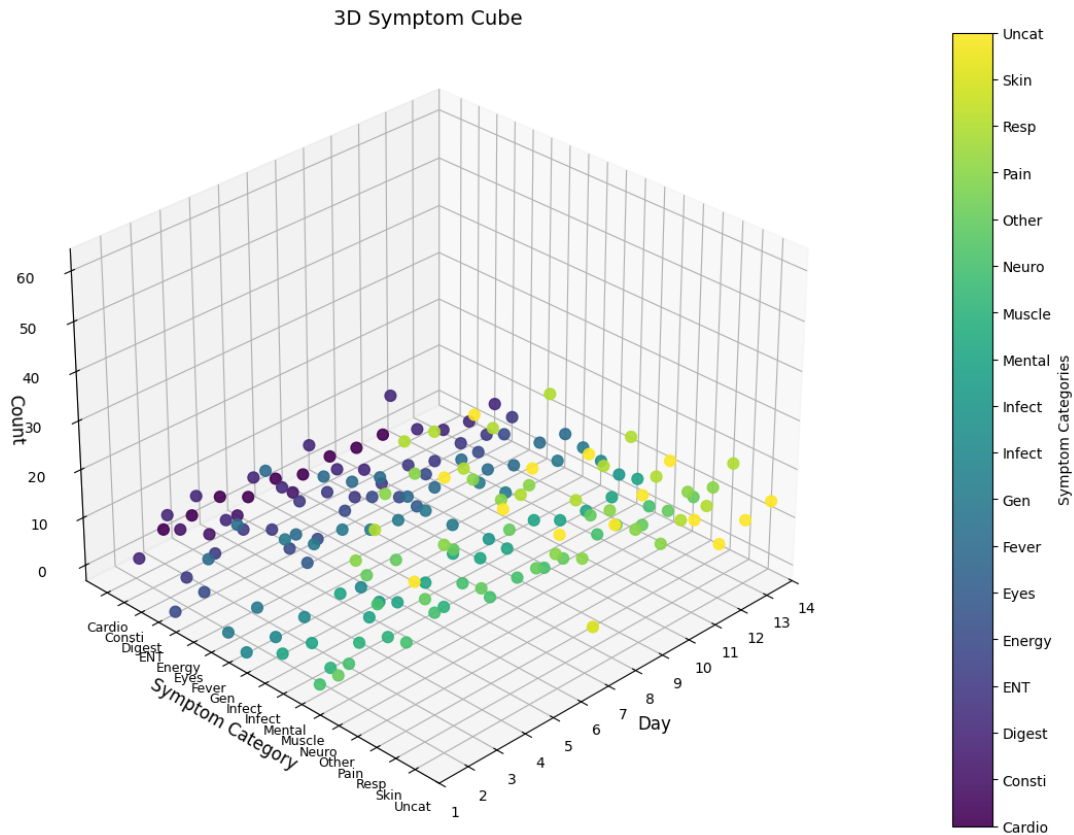


Figure C.1. 3D Symptoms cube of GPT 4-0613 via LangChain

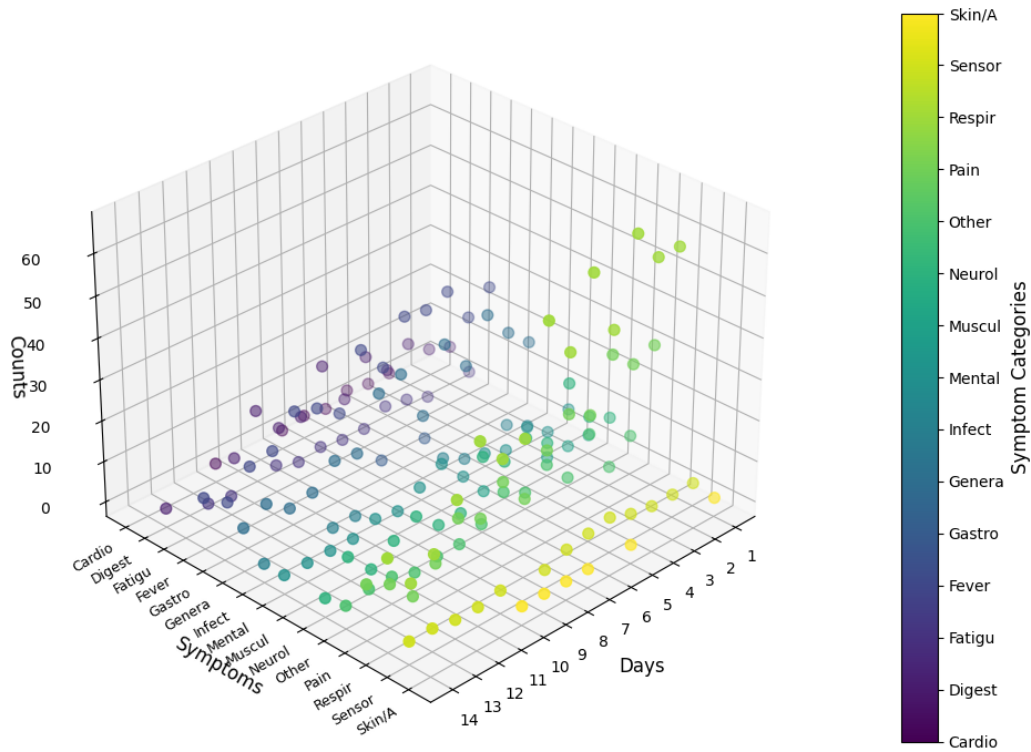


Figure C.2. 3D Symptoms cube of GPT 4-Turbo

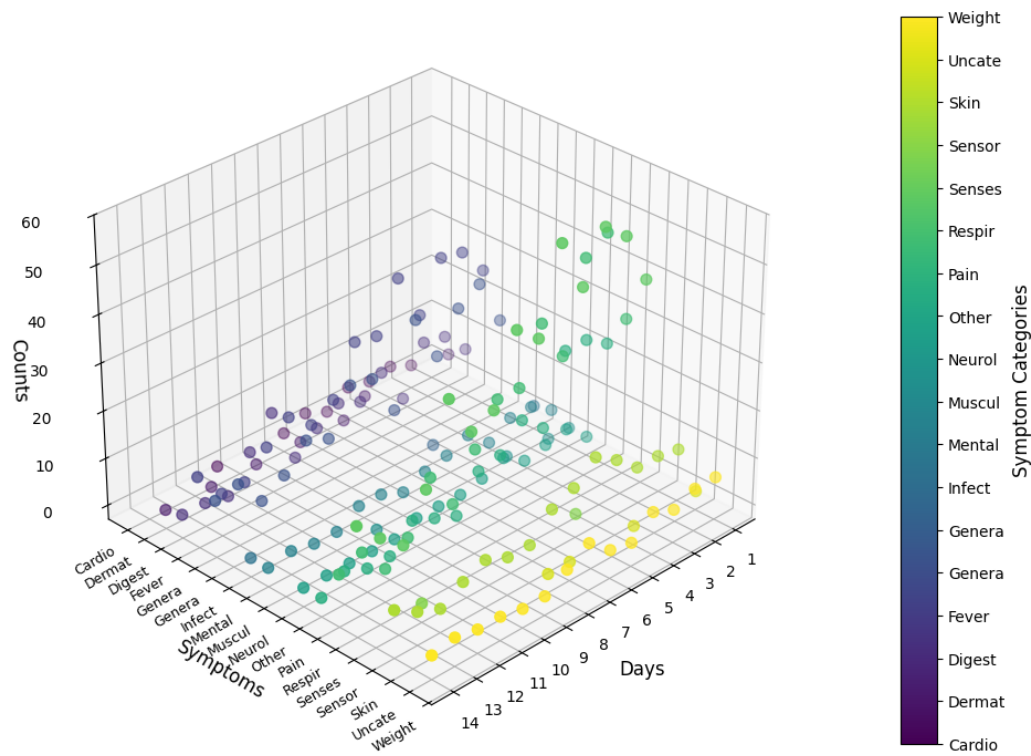


Figure C.3. 3D Symptoms cube of Gemini 2.0 Flash

The following figures present individual line plots showing the daily distribution of symptom categories extracted using Gemini 2.0 Flash across a 14-day illness timeline. Each chart captures the frequency of tweets containing symptoms associated with a specific category, highlighting temporal patterns in how self-reported experiences were distributed across different stages of illness. These category-wise trends complement the 3D cube visualization presented in the main text and provide additional granularity for interpreting model behavior over time.

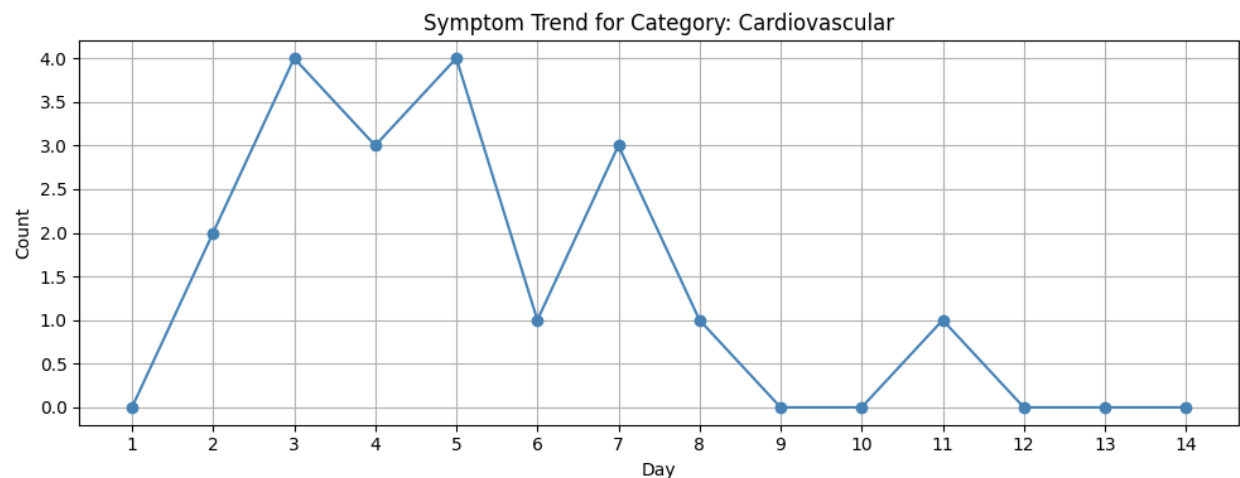


Figure C.4. Daily trend of Cardiovascular symptoms over 14 days.

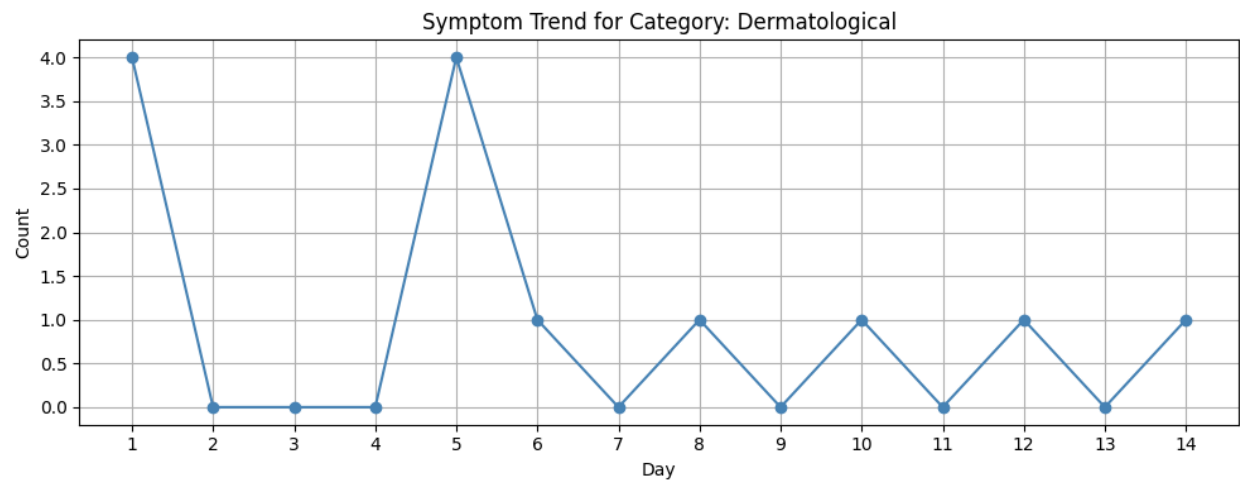


Figure C.5. Daily trend of Dermatological symptoms over 14 days.

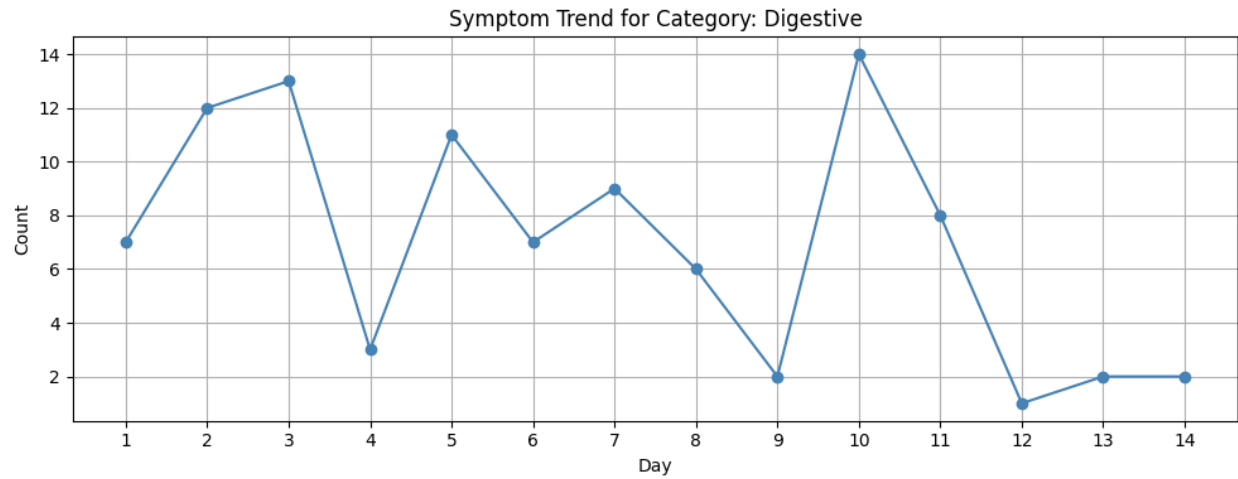


Figure C.6. Daily trend of Digestive symptoms over 14 days.

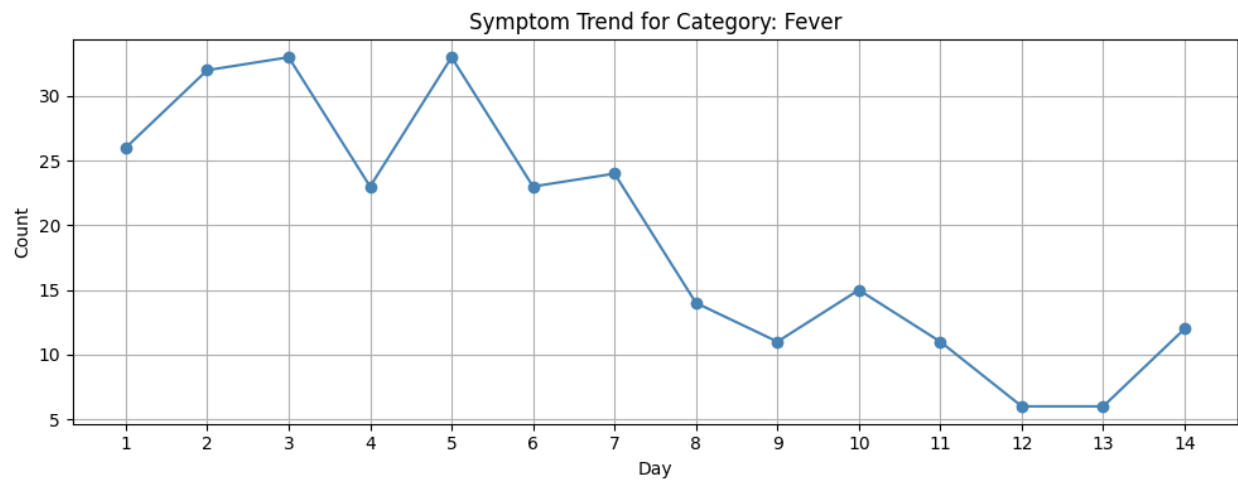


Figure C.7. Daily trend of Fever symptoms over 14 days.

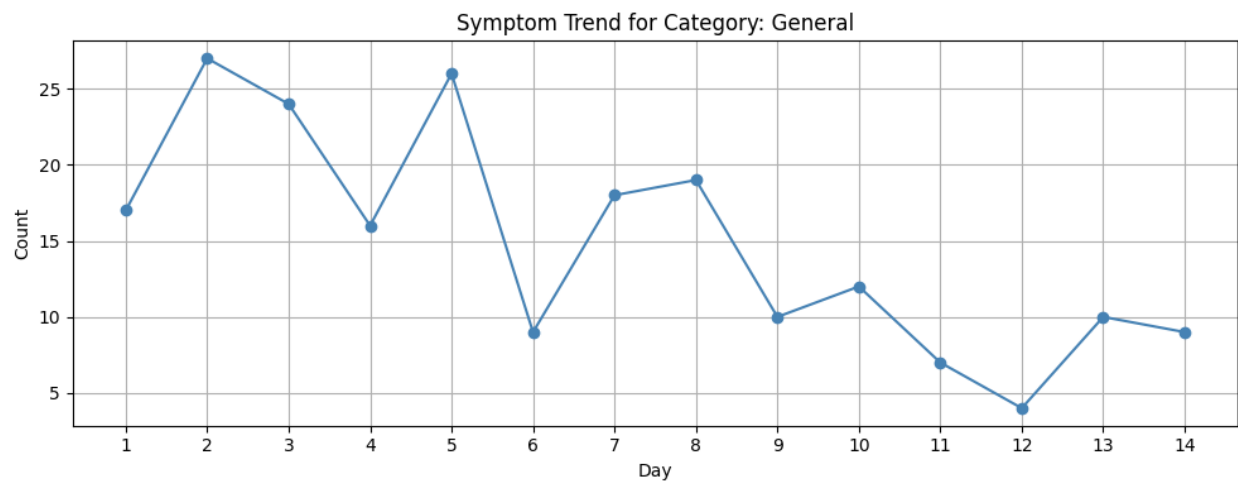


Figure C.8. Daily trend of General symptoms over 14 days.

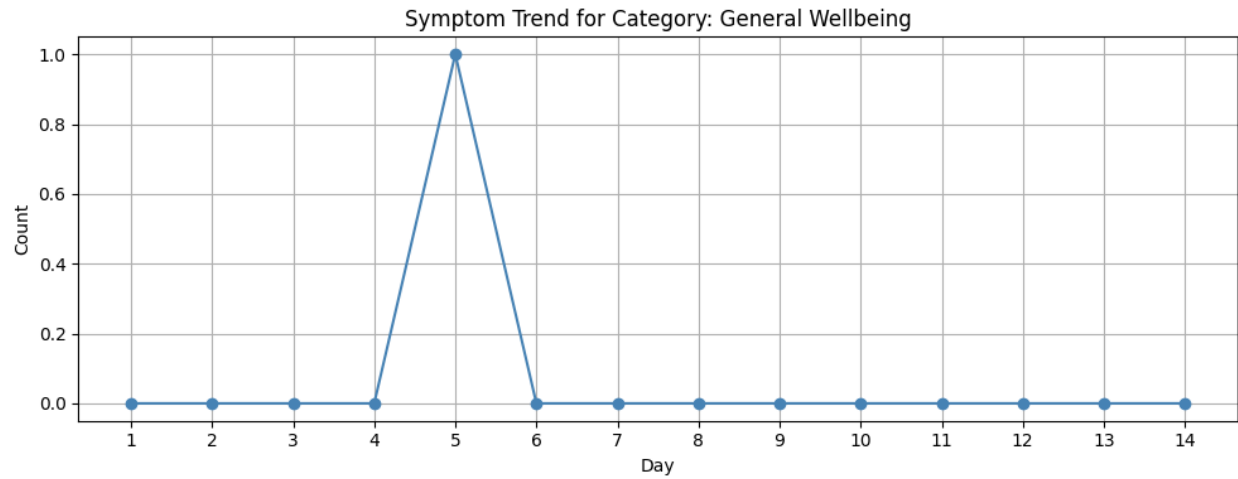


Figure C.9. Daily trend of General Wellbeing symptoms over 14 days.

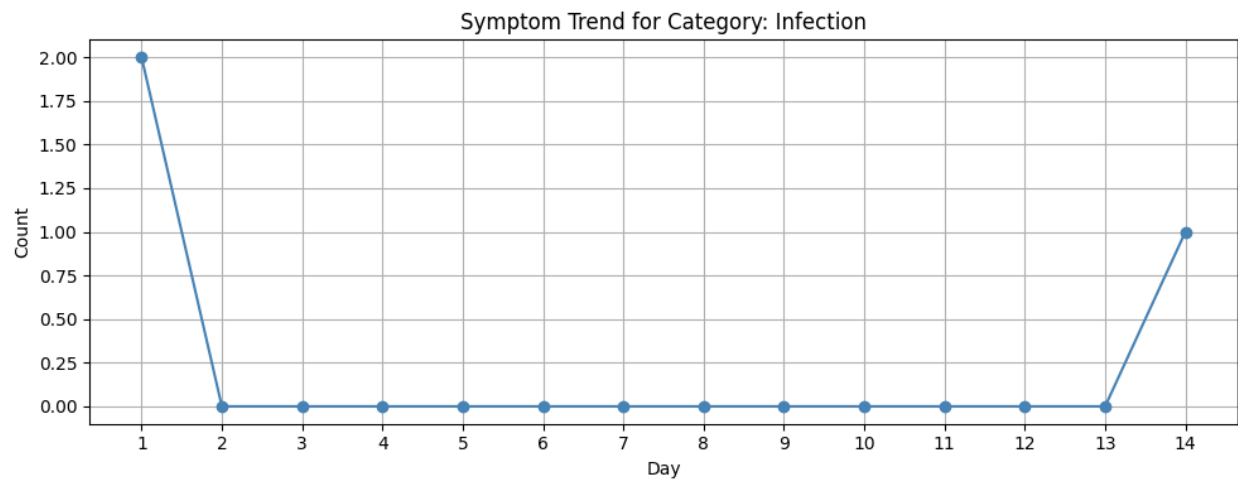


Figure C.10. Daily trend of Infection symptoms over 14 days.

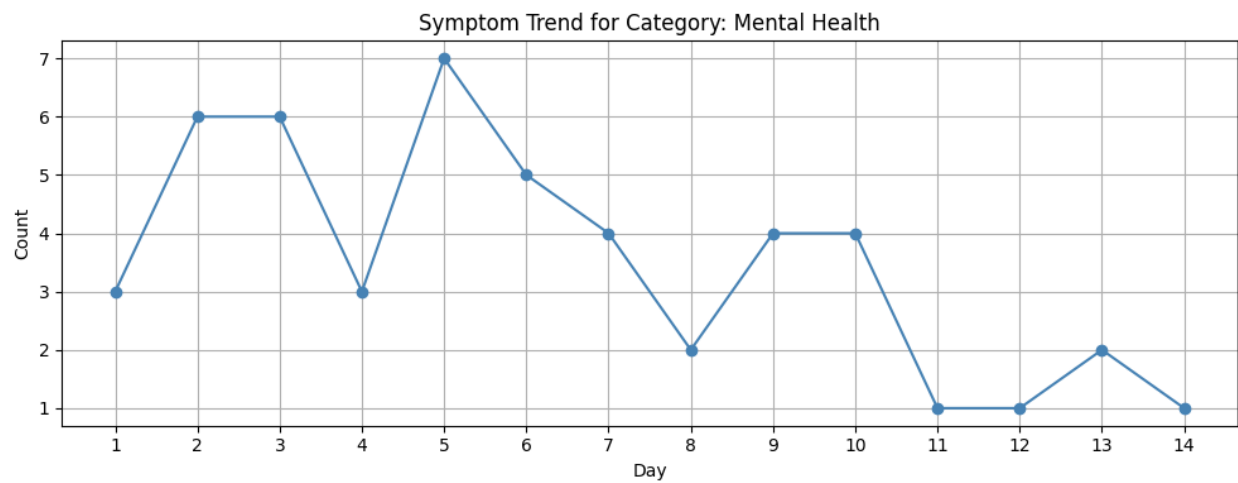


Figure C.11. Daily trend of Mental Health symptoms over 14 days.

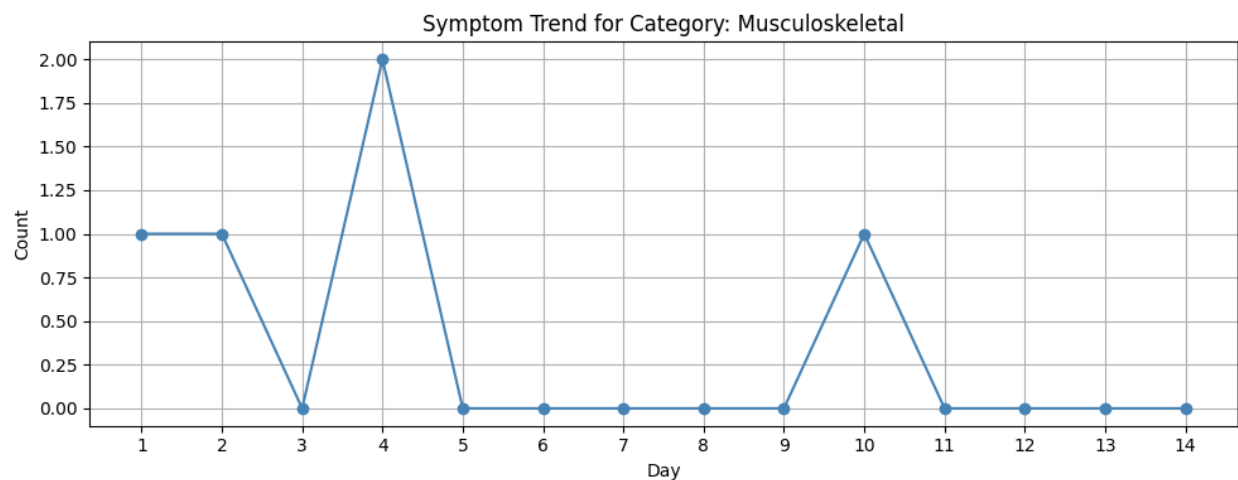


Figure C.12. Daily trend of Musculoskeletal symptoms over 14 days.

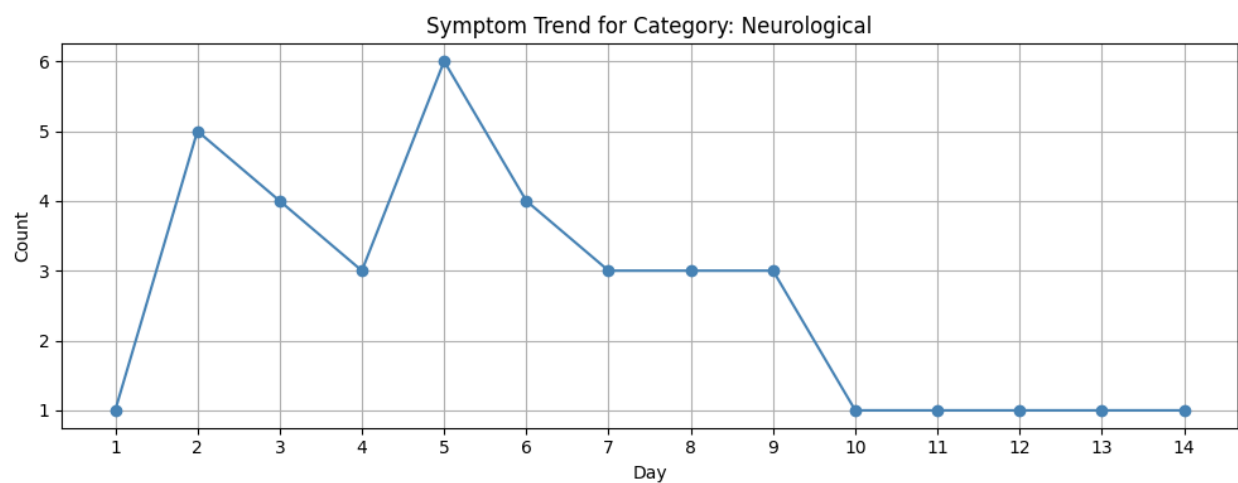


Figure C.13. Daily trend of Neurological symptoms over 14 days.

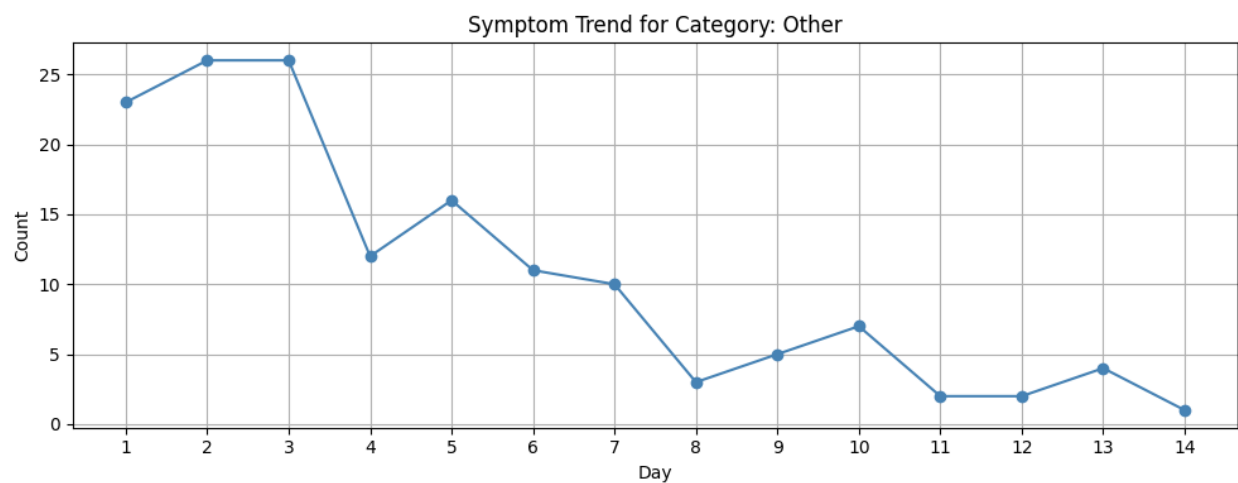


Figure C.14. Daily trend of Other symptoms over 14 days.

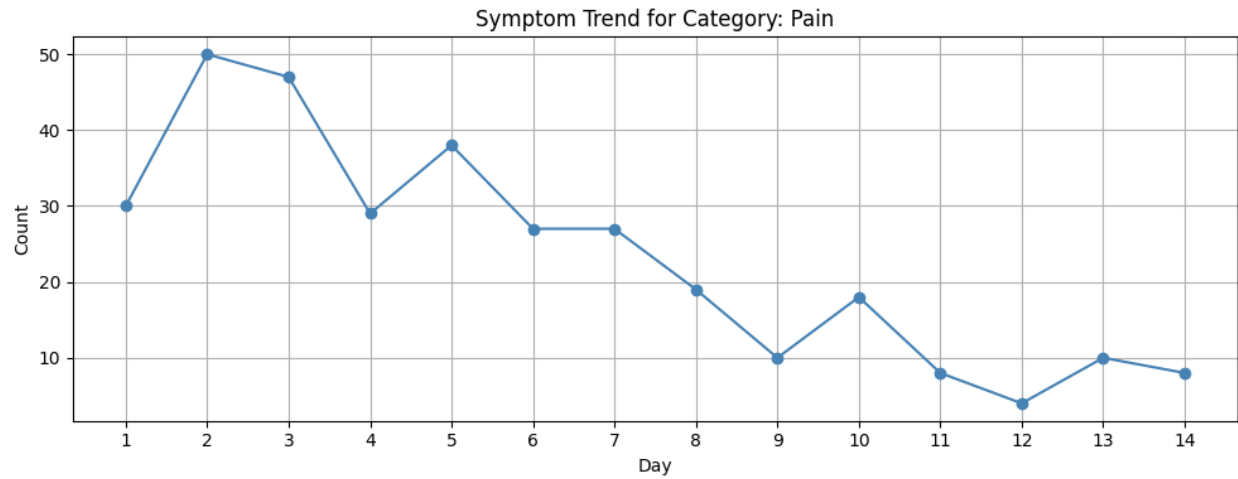


Figure C.15. Daily trend of Pain symptoms over 14 days.

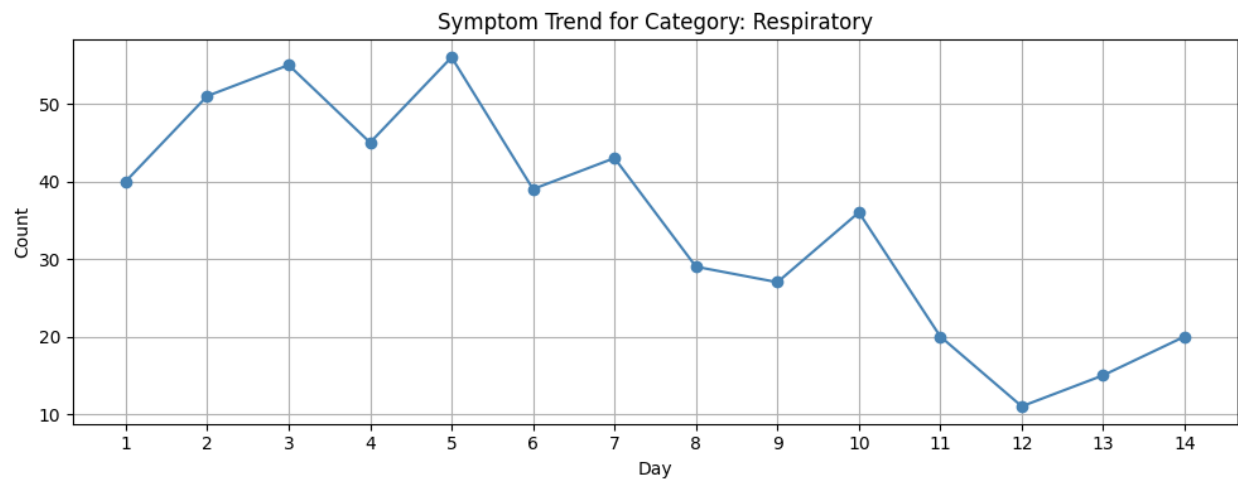


Figure C.16. Daily trend of Respiratory symptoms over 14 days.

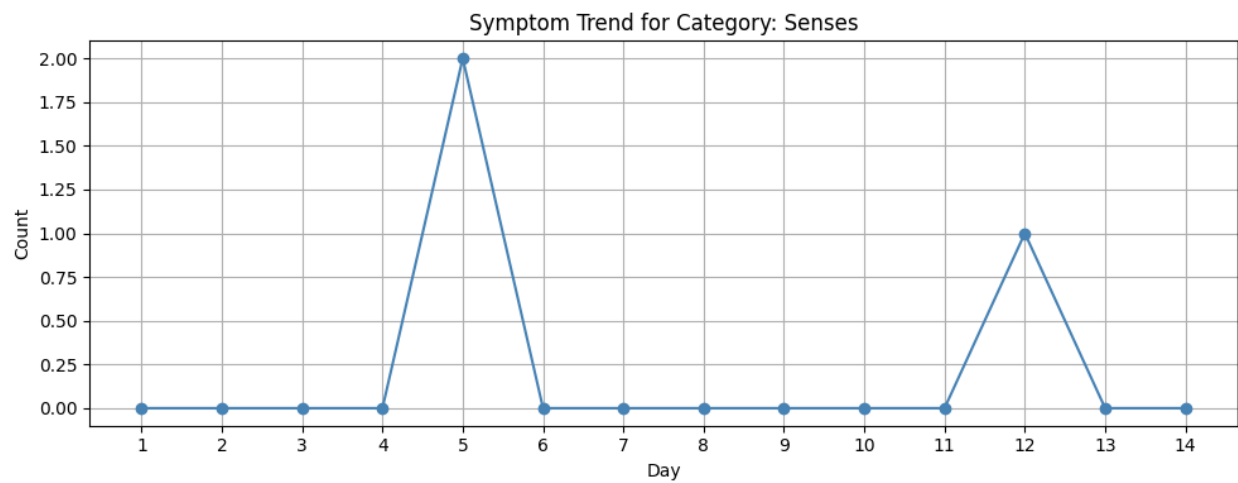


Figure C.17. Daily trend of Senses symptoms over 14 days.

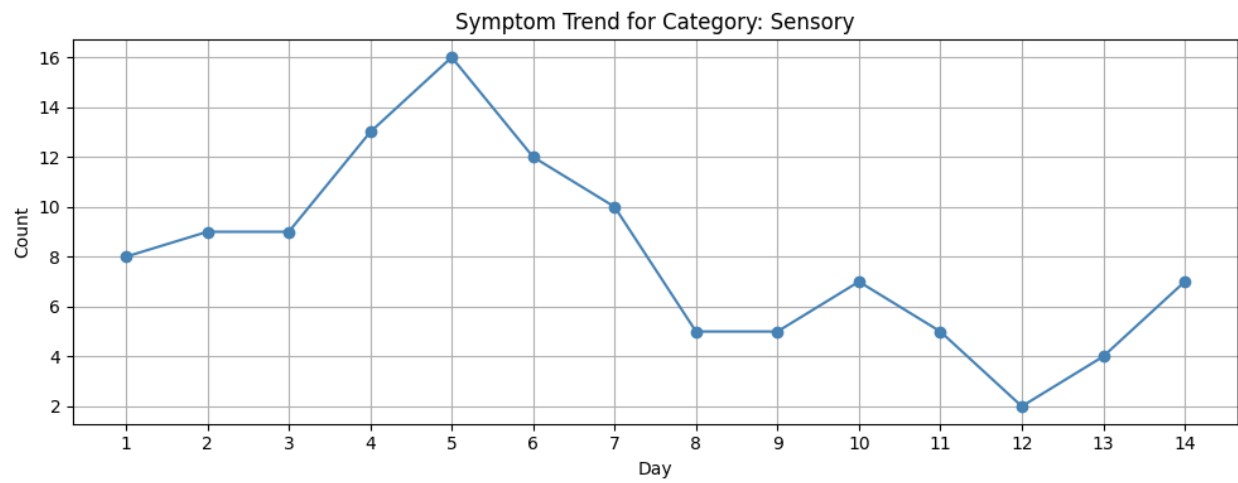


Figure C.18. Daily trend of Sensory symptoms over 14 days.

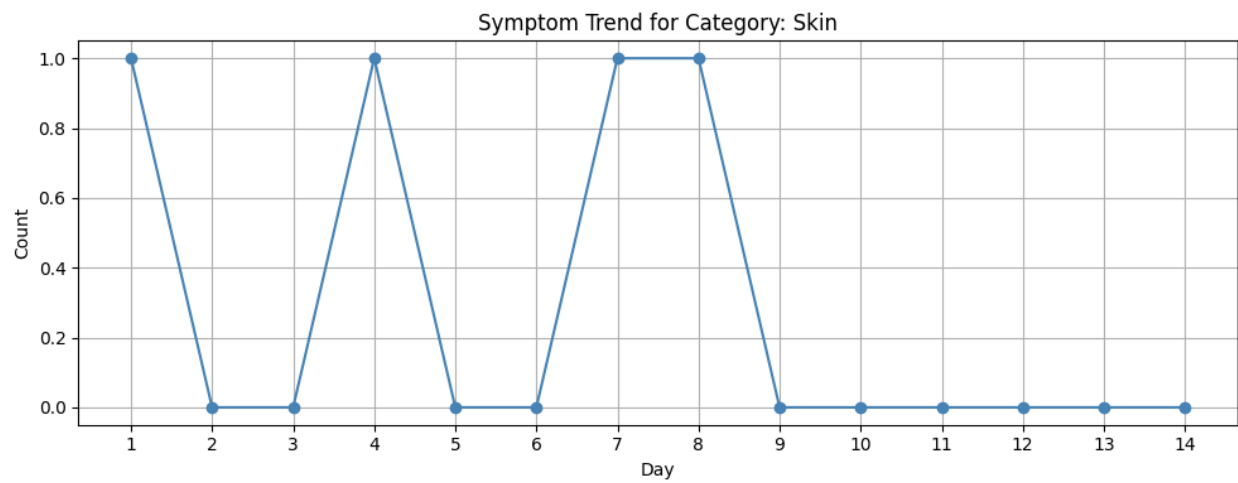


Figure C.19. Daily trend of Skin symptoms over 14 days.

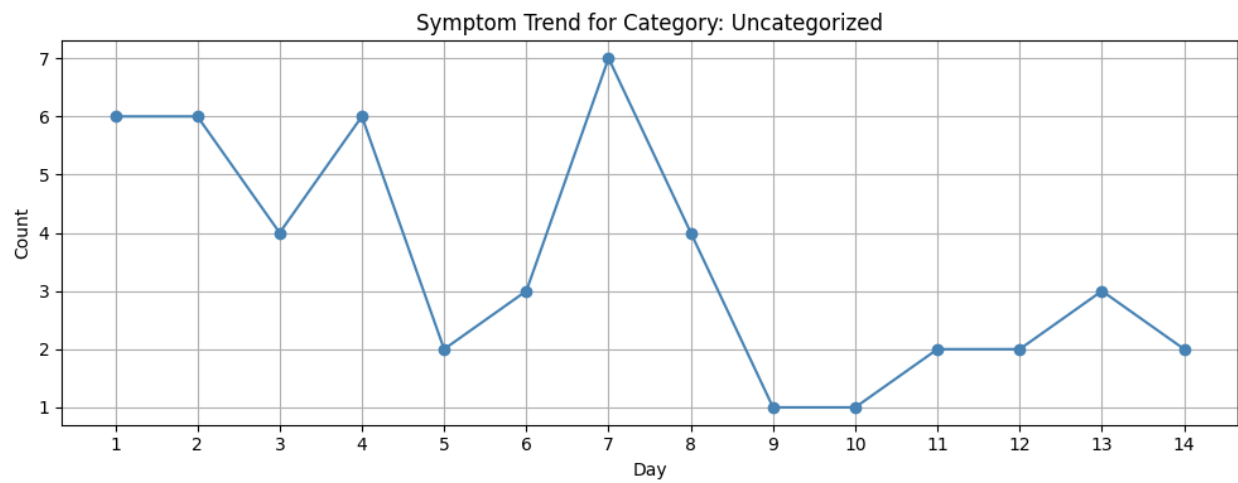


Figure C.20. Daily trend of Uncategorized symptoms over 14 days.

REFERENCES

- Anggrainingsih, R., Hassan, G. M., & Datta, A. (2021). BERT based classification system for detecting rumours on Twitter. *arXiv preprint arXiv:2109.02975*.
- Arias, F., Núñez, M. Z., Guerra-Adames, A., Tejedor-Flores, N., & Vargas-Lombardo, M. (2022). Sentiment analysis of public social media as a tool for health-related topics. *Ieee Access*, 10, 74850-74872.
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H., Kader, R., Ortiz-Prado, E., ... & Bressem, K. K. (2025). Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1), 26.
- Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016, February). Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 1171-1184).
- Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen, J. M., ... & Corley, C. D. (2015). Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10), e0139701.
- Correia, R. B., Wood, I. B., Bollen, J., & Rocha, L. M. (2020). Mining social media data for biomedical signals and health-related behavior. *Annual review of biomedical data science*, 3(1), 433-458.
- Denecke, K. (2014). Extracting medical concepts from medical social media with clinical NLP tools: a qualitative study. In *Proceedings of the fourth workshop on building and evaluation resources for health and biomedical text processing* (pp. 54-60).
- Ding, X., Carik, B., Gunturi, U. S., Reyna, V., & Rho, E. H. R. (2024, May). Leveraging prompt-based large language models: predicting pandemic health decisions and outcomes through social media language. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-20).
- Dunn, A. G., Mandl, K. D., & Coiera, E. (2018). Social media interventions for precision public health: promises and risks. *NPJ digital medicine*, 1(1), 47.

- Guo, Y., Ovadje, A., Al-Garadi, M. A., & Sarker, A. (2024). Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, 31(10), 2181-2189.
- Hermawan, A. R., Hafidz, I., Rangkuti, R. Y., Latiffianti, E., & Rakhmawati, N. A. (2024, December). Early Detection of Long COVID Symptoms from Social Media Using BERT. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1-5). IEEE.
- Jiang, K., Calix, R., & Gupta, M. (2016, August). Construction of a personal experience tweet corpus for health surveillance. In *Proceedings of the 15th workshop on biomedical natural language processing* (pp. 128-135).
- Jiang, K., Devendra, V., Chavan, S., & Bernard, G. R. (2023, December). Detection of Day-Based Health Evidence with Pretrained Large Language Models: A Case of COVID-19 Symptoms in Social Media Posts. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 4208-4212). IEEE.
- Jiang, K., & Bernard, G. R. (2024, December). The Ability of Pretrained Large Language Models in Understanding Health Concepts in Social Media Posts. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 5494-5498). IEEE.
- Jiang, Y., Qiu, R., Zhang, Y., & Zhang, P. F. (2023, November). Balanced and explainable social media analysis for public health with large language models. In *Australasian database conference* (pp. 73-86). Cham: Springer Nature Switzerland.
- Khan, P. I., Asim, M. N., Dengel, A., & Ahmed, S. (2023). A Unique Training Strategy to Enhance Language Models Capabilities for Health Mention Detection from Social Media Content. *arXiv preprint arXiv:2310.19057*.
- Lamsal, R. (2021). Design and analysis of a large-scale COVID-19 tweets dataset. *applied intelligence*, 51, 2790-2804.
- LangChain. (2024). *Introduction to LangChain*. LangChain Documentation.
- Lee, K., Agrawal, A., & Choudhary, A. (2015, August). Mining social media streams to improve public health allergy surveillance. In *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015* (pp. 815-822).
- Lee, K., Hasan, S. A., Farri, O., Choudhary, A., & Agrawal, A. (2017, August). Medical concept normalization for online user-generated texts. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 462-469). IEEE.
- Lin, Y. K., Chen, H., & Brown, R. A. (2013). MedTime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics*, 46, S20-S28.

- Lu, Z., Peng, Y., Cohen, T., Ghassemi, M., Weng, C., & Tian, S. (2024). Large language models in biomedicine and health: current research landscape and future directions. *Journal of the American Medical Informatics Association*, 31(9), 1801-1811.
- Martínez, P., Martínez, J. L., Segura-Bedmar, I., Moreno-Schneider, J., Luna, A., & Revert, R. (2016). Turning user generated health-related content into actionable knowledge through text analytics services. *Computers in Industry*, 78, 43-56.
- Moradi, M., & Samwald, M. (2022). Improving the robustness and accuracy of biomedical language models through adversarial training. *Journal of Biomedical Informatics*, 132, 104114.
- Ntinopoulos, V., Biefer, H. R. C., Tudorache, I., Papadopoulos, N., Odavic, D., Risteski, P., ... & Dzemali, O. (2025). Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. *BMJ Health & Care Informatics*, 32(1), e101139.
- OpenAI. (2023). *GPT-4 technical specifications*. OpenAI.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35, 27730-27744.
- Prieto, V. M., Matos, S., Alvarez, M., CACHEDA, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. *PloS one*, 9(1), e86191.
- Qin, H., & Tong, Y. (2025). Opportunities and Challenges for Large Language Models in Primary Health Care. *Journal of Primary Care & Community Health*, 16, 21501319241312571.
- Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53, 196-207.
- Schmidt, D. C., Spencer-Smith, J., Fu, Q., & White, J. (2024). Towards a catalog of prompt patterns to enhance the discipline of prompt engineering. *ACM SIGAda Ada Letters*, 43(2), 43-51.
- Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: a systematic review. *American journal of public health*, 107(1), e1-e8.
- Smolyak, D., Bjarnadóttir, M. V., Crowley, K., & Agarwal, R. (2024). Large language models and synthetic health data: progress and prospects. *JAMIA open*, 7(4), ooae114.
- Sun, W., Rumshisky, A., & Uzuner, O. (2013). Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, 46, S5-S12.

- Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., ... & Poon, H. (2023). Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4).
- Tseng, Y. C., Kuo, C. W., Peng, W. C., & Hung, C. C. (2024). al-BERT: a semi-supervised denoising technique for disease prediction. *BMC medical informatics and decision making*, 24(1), 127.
- Velasco, E., Agheneza, T., Denecke, K., Kirchner, G., & Eckmanns, T. (2014). Social media and internet-based data in global systems for public health surveillance: a systematic review. *The Milbank Quarterly*, 92(1), 7-33.
- Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., & Liu, H. (2020). MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54, 57-72.
- Wen, B., Norel, R., Liu, J., Stappenbeck, T., Zulkernine, F., & Chen, H. (2024, July). Leveraging Large Language Models for Patient Engagement: The Power of Conversational AI in Digital Health. In *2024 IEEE International Conference on Digital Health (ICDH)* (pp. 104-113). IEEE.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Wiest, I. C., Ferber, D., Zhu, J., van Treeck, M., Meyer, S. K., Juglan, R., ... & Kather, J. N. (2024). Privacy-preserving large language models for structured medical information retrieval. *NPJ Digital Medicine*, 7(1), 257.
- X.com. (2024). *Twitter API: Data dictionary*. X.com Developer Documentation.
- Xie, J., Zhang, Z., Zeng, S., Hilliard, J., An, G., Tang, X., ... & Xu, D. (2025). Leveraging Large Language Models for Infectious Disease Surveillance—Using a Web Service for Monitoring COVID-19 Patterns From Self-Reporting Tweets: Content Analysis. *Journal of Medical Internet Research*, 27(1), e63190.
- Xu, H., Stetson, P. D., & Friedman, C. (2007). A study of abbreviations in clinical notes. In *AMIA annual symposium proceedings* (Vol. 2007, p. 821).
- Xu, Y., & Cheng, Y. (2023, July). Spontaneous gestures encoded by hand positions improve language models: An information-theoretic motivated study. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 9409-9424).
- Zhang, K., Meng, X., Yan, X., Ji, J., Liu, J., Xu, H., ... & Tang, Y. D. (2025). Revolutionizing health care: The transformative impact of large language models in medicine. *Journal of Medical Internet Research*, 27, e59069.

Zheng, E. T., Fu, H. Z., Thelwall, M., & Fang, Z. (2024). Can tweets predict article retractions? A comparison between human and LLM labelling. *arXiv preprint arXiv:2403.16851*.