Multilingual Symptom Detection on Social Media: Enhancing Health-related Fact-checking with LLMs

Sa'idah Zahrotul Jannah, Elyanah Aco, Shaowen Peng, Shoko Wakamiya, and **Eiji Aramaki**
Nara Institute of Science and Technology, Japan
saidah.zahrotul_jannah.rw9@naist.ac.jp
elyanah_marie_cariaga.aco.ef1@naist.ac.jp
peng.shaowen@naist.ac.jp
{wakamiya, aramaki}@is.naist.jp

Abstract

Social media has emerged as a valueable source for early pandemic detection, as repeated mentions of symptoms by users may signal the onset of an outbreak. However, to be a reliable system, validation through fact-checking and verification against official health records is essential. Without this step, systems risk spreading misinformation to the public. The effectiveness of these systems also depend on their ability to process data in multiple languages, given the multilingual nature of social media data. Yet, many NLP datasets and disease surveillance system remain heavily English-centric, leading to significant performance gaps for low-resource languages. This issue is especially critical in Southeast Asia, where symptom expression may vary culturally and linguistically. Therefore, this study evaluates the symptom detection capabilities of LLMs in social media posts across multiple languages, models, and symptoms to enhance health-related fact-checking. Our results reveal significant language-based discrepancies, with European languages outperforming under-resourced Southeast Asian languages. Furthermore, we identify symptom-specific challenges, particularly in detecting respiratory illnesses such as influenza, which LLMs tend to overpredict. The overestimation or misclassification of symptom mentions can lead to false alarms or public misinformation when deployed in real-world settings. This underscores the importance of symptom detection as a critical first step in medical fact-checking within early outbreak detection systems.

1 Introduction

Social media can be used for early pandemic detection (Shi et al., 2024). When many users repeatedly mention or complain about a certain symptom, it may indicate the potential onset of an outbreak. Gour et al. (2022) conducted a study on the COVID-19 outbreak and found that social media activity can reflect the state of an outbreak. Specifically, the study revealed that negative tweets posted during a crisis tend to align with the scale of the disease outbreak. However, to ensure the reliability of these detections, it is essential to validate them by fact-checking and verifying against the official health records. This transition from detection to health-related fact-checking and verification forms the foundation for building reliable public health monitoring systems. If a system detects a potential pandemic that does not correspond to official health records, it may contribute to the spread of misinformation to the public.

Nevertheless, the effectiveness of such systems relies on their ability to process data in multiple languages, as social media users come from all over the world. Yet, a comprehensive study on Natural Language Processing (NLP) datasets revealed a significant bias towards English, resulting in better performance than other languages for many tasks (Brown et al., 2020; Yu et al., 2022; Lai et al., 2023). In the field of disease surveillance, most existing epidemiological datasets and detection systems have also been developed primarily in English, with only limited support for other languages (Parekh et al., 2024a).

The performance gap is potentially wider for languages with little labeled or even unlabeled data, such as the majority of languages in Southeast Asia (SEA), a linguistically diverse region home to over 1300 languages (Joshi et al., 2020; Lovenia et al., 2024). These factors also pose a challenge in developing automatic symptom detection due to cultural (Anggoro and Jee, 2021) and linguistic variations (Wang et al., 2010), such as the use of idiomatic expressions and colloquial terms for common symptoms. A notable gap persists in addressing symptom identification and health-related data processing for languages in this region and under-resourced languages as a whole.

Addressing those challenges requires technolo-

Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), pages 54–68

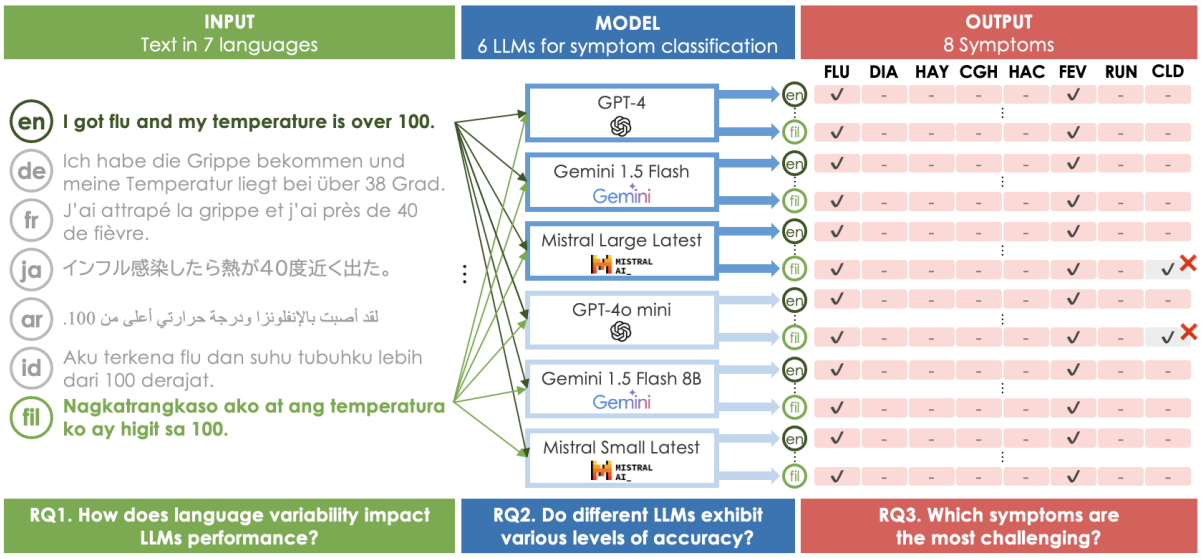


Figure 1: Overview of this study. We evaluate the symptom detection capabilities of LLMs on social media posts across seven languages (English: en, German: de, French: fr, Japanese: ja, Arabic: ar, Indonesian: id, Filipino: fil), six models (large-parameter models, e.g., GPT-4, Gemini 1.5 Flash, and Mistral Large Latest, and small-parameter models, e.g., GPT-4o mini, Gemini 1.5 Flash 8B, and Mistral Small Latest), and eight symptoms (influenza: FLU, diarrhea: DIA, hay fever: HAY, cough: CGH, headache: HAC, fever: FEV, runny nose: RUN, and cold: CLD). On the right side (Output), '✓' and '-' mean positive and negative for a symptom, respectively. Labels with a pink background indicate correct predictions, while labels with a gray background indicate incorrect predictions.

gies that can generalize across diverse linguistic contexts. In this regard, LLMs offer promising capabilities for improving symptom detection system, supporting health-related fact-checking. This study aims to investigate how language variability affects symptom identification using LLMs, highlighting the importance of developing practical systems with multilinguality in building reliable health-related fact-checking and disease surveillance systems. Specifically, the contribution of this paper is by answering the following research questions.

RQ1. How does language variability affect the LLMs performance for detecting symptom mentions that support health-related fact-checking?

RQ2. Do different LLMs exhibit various levels of accuracy when classifying symptoms?

RQ3. Which symptoms are the most challenging for LLMs to detect accurately, potentially impacting factuality assessment?

This study used European (English, German, and French) and Asian (Japanese, Arabic, Indonesian, and Filipino) languages as shown in Figure 1. Using an extended version of the NTCIR-13 MedWeb test dataset, we evaluated two model sizes (large- and small-parameter) from three general LLM providers: OpenAI, Google Gemini, and Mistral AI. Each model performs zero-shot multilabel symptom classification, categorizing posts as positive or negative for eight symptoms including influenza, diarrhea, hay fever, etc.Performance is measured through F1-score (standard NLP evaluation) and Relative Distance (disease-surveillance perspective) to assess estimation bias.

2 Related Work

2.1 Health-related Fact-Checking

Health-related fact-checking often involves addressing misinformation and infodemics. Social media, while being a rapid channel for the spread of misinformation, also serves as a valuable platform to counter false information, particularly during disease outbreaks, by disseminating content grounded in scientific evidence and supported by collaborations with local health authorities (Bayani et al., 2025; Vázquez-Gestal et al., 2024; Purnat et al., 2024). Approaches to fact-checking typically include manual verification, automated claim detection, and evidence retrieval (Sarrouti et al., 2021; Sharifpoor et al., 2025; Vladika et al., 2024).

Existing studies have primarily focused on verifying complete health claims. In contrast, this

paper aims to explore symptom mention detection as a critical first step within the broader framework of health-related fact-checking, especially in the context of disease outbreaks.

Accurately identifying symptoms from user-generated content is essential for improving both the speed and reliability of outbreak response. Recent research has demonstrated the potential of social media as an early detection system for pandemics, identifying signs of an outbreak before it is officially declared. Parekh et al. (2024b) conducted research on epidemic prediction using event detection from social media data. Their framework was able to generate warnings 4 to 9 weeks earlier than the official epidemic declaration by the WHO for Monkeypox. The study demonstrated an alignment between the predicted outbreaks and the actual epidemic cases later confirmed by the official sources.

2.2 Multilingual Medical LLMs

Several studies have shown that language and cultural barriers between patients and healthcare providers can lead to unequal health outcomes, such as misdiagnoses, inadequate treatment, and lower patient safety and satisfaction (Ohtani et al., 2015; Schouten et al., 2020; Shamsi et al., 2020). This is especially the case in low-resource settings and for ethnic minorities, where intermediaries such as qualified interpreters and comprehensive translation resources may not easily be available.

The introduction of LLMs has opened up possibilities for addressing these barriers by enabling real-time translation and enhancing diagnostic accuracy, especially when fine-tuned for medical applications. While most medical corpora and language models are primarily in and designed for English, recent advancements have expanded their capability to support multiple languages. Models such as Medical mT5, Apollo, and BiMediX, which were trained on medical datasets for languages other than English, demonstrate higher average performance across different languages compared to commercial models (García-Ferrero et al., 2024; Pieri et al., 2024; Wang et al., 2024). Additionally, multilingual medical benchmarks such as in Qiu et al. (2024) have been developed to evaluate LLMs on tasks such as biomedical academical question-answering and diagnosis assessment.

However, resource constraints and ethical issues can hinder the development of truly inclusive multilingual medical LLMs. Building medical corpora for low-resource languages to train models may require substantial effort, such as collecting and transcribing hand-written health records and building local data dictionaries for medical terminologies (Wahl et al., 2018). Bias and misinformation within training data can be reproduced in LLM-generated content, posing significant risks in medical decision-making and reinforcing health outcome inequities (Omiye et al., 2023; Poulain et al., 2024; Yang et al., 2024).

3 Corpus

The NTCIR-13 MedWeb task test dataset was used and extended for this study. The dataset was crowdsourcing-generated short posts as detailed in Wakamiya et al. (2017). It consists of 640 posts, with no personal identifiers, related to systemic (fever and headache), digestive (diarrhea) and respiratory symptoms (cold, cough, hay fever, influenza, and runny nose).

The posts were labeled as positive or negative for each symptom based on whether the user indicated experiencing that symptom at the time, following the annotation guideline for NTCIR-13 MedWeb task (MedWeb, 2017). As seen in Table 1, only a small number of posts are classified as positive for each symptom. Table 2 shows examples of post labeled for each symptom.

Symptoms	# of Positive label	% of Positive label
Influenza (FLU)	24	3.75%
Diarrhea (DIA)	64	10.00%
Hayfever (HAY)	46	7.19%
Cough (CGH)	80	12.50%
Headache (HAC)	77	12.03%
Fever (FEV)	93	14.53%
Runnynose (RUN)	123	19.22%
Cold (CLD)	90	14.06%

Table 1: Positive label count and percentage per symptom.

Aside from English and Japanese which were included in the original task, the dataset was expanded using human translation service to five different languages: German, French, Modern Standard Arabic, Indonesian, and Filipino. This represents a mix of languages from European (specifically Indo-European) and Asian language families, all of which are considered high-resource except for the SEA languages Indonesian and Filipino (Joshi et al., 2020; Hammarström et al., 2024). These languages exhibit substantial linguistic di-

versity due to differences in scripts, phonological structures and other linguistic features, which lead to varying levels of complexity in text processing. The use of translation service is to minimize translation bias and to ensure that the symptom expressions were naturally and appropriately conveyed in the target language. The service had experienced some translation tasks related to the author's research previously and complied with the institution's financial procedures. Payment is made according to the agreed terms between the service provider and the authors. We provided the instruction to do the translation and the delivery format as seen in Appendix A.1.

4 Experimental Setup

4.1 Large Language Models

Two models of different parameter sizes (large and small) were chosen from each of three LLM families. The six chosen models are GPT-4 and GPT-4o mini by OpenAI, Gemini 1.5 Flash and Gemini 1.5 Flash 8B by Gemini Team Google, and Mistral Large and Small by Mistral AI. GPT-4, GPT-4o mini, Gemini 1.5 Flash, and Gemini 1.5 Flash 8B are under their respective licenses. Mistral Large is under Mistral Research License while Mistral Small is under Apache 2 License.

Not all models are provided with the parameter size information. OpenAI has not disclosed the exact parameter size of GPT-4 and GPT-4o mini. However, as an advancement of GPT-3, which has 175 billion parameters (Dale, 2021), GPT-4 is believed to have a larger parameter size with its ability to comprehend natural language in more complex and nuanced contexts (OpenAI et al., 2024). Meanwhile, OpenAI has described GPT-4o-mini as its most cost-efficient small model (OpenAI, 2024). The Gemini Team Google has not revealed the parameter size of Gemini 1.5 Flash as well. However, it is known that the model has a larger parameter size than Gemini 1.5 Flash 8B, which, as its name suggests, contains 8 billion parameters (Kilpatrick and Mallick, 2024). In contrast, Mistral AI has announced the parameter sizes for Mistral Large Latest at 123 billion and Mistral Small Latest at 22 billion (Mistral, 2025).

Pretraining data for these models, nor the languages within them, are not publicly released. However, previous studies have demonstrated these LLM families' proficiency on multiple languages, including low-resourced ones. For example, Lovenia et al. (2024) showed that GPT-4 and Mistral LLMs generally matched or outperformed multilingual or language-specific models for various NLP tasks on SEA languages. In Ahuja et al. (2024), larger commercial models such as GPT-4 and Gemini performed better than smaller ones across various multilingual evaluation benchmarks, although the possibility of data contamination in pretraining is not ruled out. Despite this, Zhang et al. (2023) and Jin et al. (2024) found that commercial models consistently perform better on English prompts than their translations in other languages.

Three trials were done for each model, using default model parameters (temperature, maximum tokens, etc.) to perform zero-shot multilabel symptom classification on posts using the following prompt. See Appendix A.2 for the full set of rules, which follow the original NTCIR-13 task.

The prompt used for the study

Instruction:
Determine if the creator of this post is exhibiting symptoms for each of the following: influenza, diarrhea, hay fever, cough, headache, fever, runny nose, cold. For each symptom, only answer either 0 or 1 for negative (no symptoms) or positive (has symptoms) respectively.

Determination of symptoms is carried out based on the following rules:
• Cases where the symptom is expressed directly, including mild symptoms, are considered positive;
• A symptom can be labeled positive with indirect expressions of having a symptom;
. . .
• Other cases, like symptoms belonging to blog friends, should be labeled as negative since it is difficult to determine their location.

Post:
{A post in one of the seven studied languages.}

Return the result as a JSON object with the symptoms as keys and the values as either 0 or 1.

4.2 Evaluation Methods

In this study, results were evaluated from two perspectives. From the standard NLP perspective, models are evaluated using the average F1-score. From the disease surveillance perspective, we propose a new metric more suitable for capturing model estimation bias. Both metrics are discussed in detail in the following subsections.

4.2.1 NLP Perspective: F1-score

The F1-score is a weighted average between precision, or accuracy of positive predictions, and recall, or ability to capture positive instances. This

Post	Symptom							
	FLU	DIA	HAY	CGH	HAC	FEV	RUN	CLD
I got flu and my temperature is over 100.	✓	-	-	-	-	✓	-	-
It was diarrhea that woke me up in the middle of the night.	-	✓	-	-	-	-	-	-
My wife's allergies are acting up, it seems rough.	-	-	✓	-	-	-	✓	-
It's almost the flu season.	-	-	-	-	-	-	-	-
I think I coughed too much. My stomach muscles hurt.	-	-	-	✓	-	-	-	-
This cold is rough. I've got a headache too, it might not be an ordinary cold. Yikes.	-	-	-	-	✓	-	✓	✓

Table 2: Sample English posts with multi-symptom labels. '✓' and '-' mean positive and negative for a symptom, respectively. A post may be positive for multiple symptoms.

makes it a preferred evaluation metric for machine learning and NLP tasks, especially on imbalanced datasets. We used `scikit-learn` to calculate the F1-score for this evaluation.

4.2.2 Surveillance Perspective: Relative Distance

While the F1-score is commonly used within NLP, it is more common and practical to evaluate disease surveillance models based on how closely the total predicted positive cases match the actual figures (Xiang-Sheng and Zhong, 2015; Samui et al., 2020; Bhatia et al., 2021). We propose a new metric called relative distance (RD) for measuring estimation bias, which reflects a model's tendency to overestimate or underestimate positive predictions. This metric provides added practical relevance for public health applications beyond what standard NLP metrics can offer.

We define RD as the ratio change between predicted and actual positives as described in the following formula:

$$\mathrm{RD} = \frac{(\mathrm{TP}+\mathrm{FP}) - (\mathrm{TP}+\mathrm{FN})}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{FP}-\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \quad (1)$$

where TP indicates the number of true positives, FP the number of false positives, and FN the number of false negatives.

In a binary classification task, 0 serves as the gold standard or baseline against which prediction values are compared. A positive RD indicates overestimation of positives while a negative RD indicates underestimation.

5 Results and Discussion

We analyze LLM performance in multilingual symptom prediction across three dimensions: language-based, LLM-based, and symptom-based. The results provide insights into the applicability of LLMs for disease surveillance task, revealing current strengths and remaining challenges for real-world implementation and decision-making.

5.1 NLP Perspective: F1-Score Insights

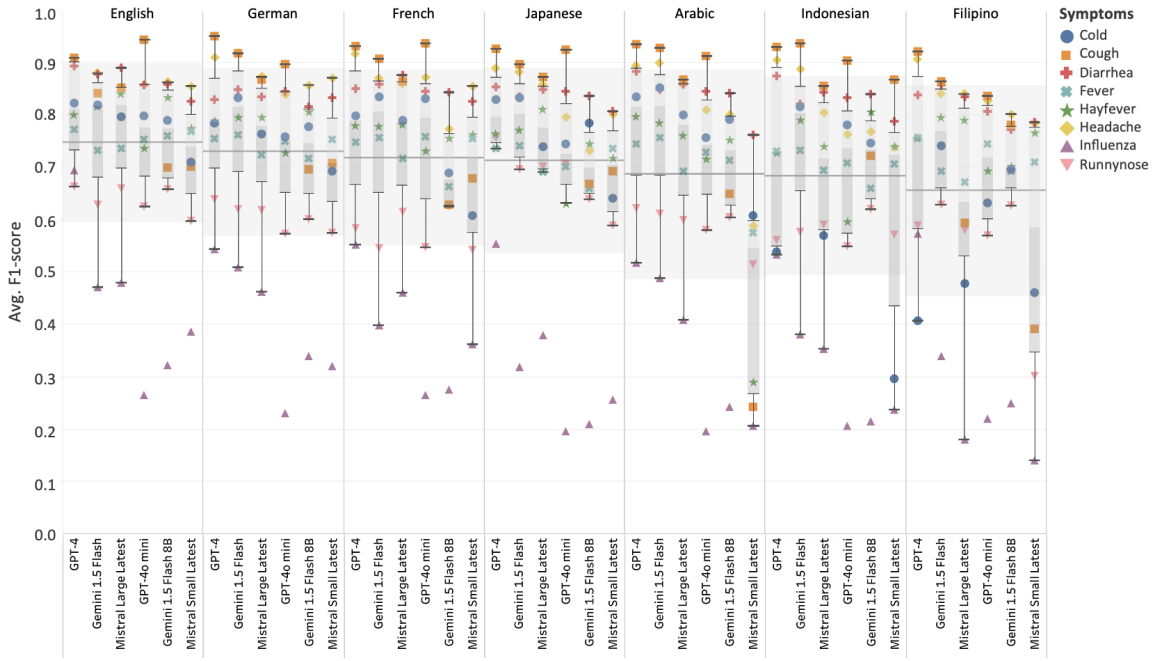
The distribution of F1-scores across the three dimensions are shown in Figure 2. Appendix A.3 and A.4 provide evaluation details.

5.1.1 Language-based Analysis

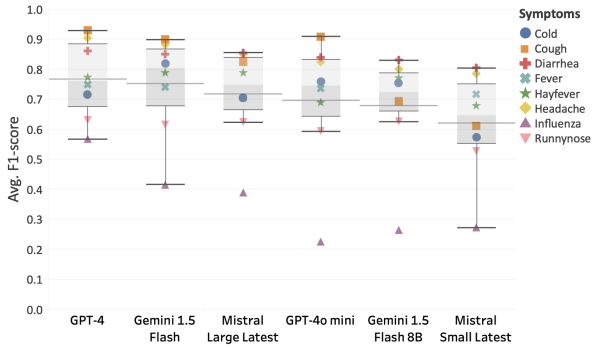
Figure 2(a) shows the distribution of F1-scores for each language, where each point corresponds to a specific symptom.

Scores range from moderate to high in all languages. The average F1-scores for English, German, and French are 0.748, 0.730, 0.719 while for Asian languages (Japanese, Arabic, Indonesian, and Filipino) are 0.714, 0.687, 0.685, and 0.656 respectively. The scores showed that LLMs performed better in the three European languages (English, German, and French) than the Asian ones based on average F1-scores, with the mid-resourced Indonesian and Filipino ranking the lowest. Furthermore, a comparison based on language categories as shown in Appendix A.5, European and Asian, reveals a significant difference in mean F1 scores. The average F1 scores for European and Asian languages are 0.73 and 0.68, respectively, with a p-value of 0.0000. Moreover, variability across European languages is also lower than all Asian languages except Japanese.

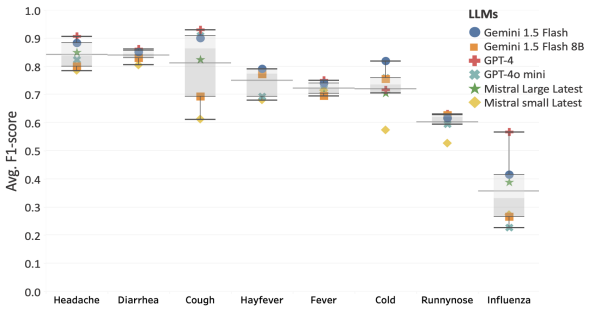
Vocabulary sharing and cultural nuances between languages may explain these results. Indo-European languages were found to improve model performance on unseen languages in the same family, but the same was not observed on other language families (Yuan et al., 2024). Addition-



(a) Score distribution per language



(b) Score distribution per LLM



(c) Score distribution per symptom

Figure 2: Comparative performance of LLMs in several perspectives. Horizontal lines within each shaded area represent group averages, while shaded areas denote scores within one standard deviation of this average. From language perspective, (a) suggests that English, German, and French have better performance than Asian languages. In LLMs' view, (b) showing large-parameter models are better. Moreover, in language-based (c) presents that Influenza is harder for LLM to predict accurately.

ally, languages with cultural contexts and idiomatic expressions that differ from those predominantly found in training data had lower and less consistent model performance (Tao et al., 2024).

Numerous outliers with F1-scores below 0.4 are observed for all languages, indicating possible challenges in symptom prediction regardless of language used. The presence of outliers is crucial in real-world applications, where low-performance instances can lead to critical errors, especially in sensitive tasks such as disease symptom detection. The following sections examine model-specific and symptom-specific performance to identify factors contributing to these outliers.

5.1.2 LLM-based Analysis

Figure 2(b) shows the distribution of F1 scores for each LLM, with points representing individual symptoms. Large-parameter models—GPT-4, Gemini 1.5 Flash, and Mistral Large Latest—achieved higher average F1 scores. The average performance of small-parameter models was at least 0.02 lower than their larger counterparts, with the largest gap observed between GPT-4 and Mistral Small Latest. Furthermore, statistical testing (Appendix A.6) confirmed a significant difference between large- and small-models, with mean scores of 0.74 and 0.66 and a p-value of 0.0000.

Outliers were observed in some models, indi-

cating potential challenges in predicting certain symptoms. Additionally, Mistral Small Latest exhibited outlier behavior for certain symptoms in Figure 2(c), suggesting difficulties in accurately identifying those symptoms.

In detailed, as shown in Appendix A.3, Gemini 1.5 Flash has the smallest variance across different languages among the larger parameter models, although it does not reach the higher F1-scores achieved by GPT-4. All models perform better on European languages than Asian ones, and English continues to have the highest F1 scores for most models. The performance difference between European and Asian languages is more pronounced in Mistral LLMs. This may be explained by the number of languages officially supported by each family. Only English, French, Spanish, German, and Italian are officially supported in the Mistral models (Mistral AI, 2024), compared to the 98 supported in OpenAI's speech-to-text Whisper models (Radford et al., 2023) and over 100 supported in Google's Gemini models (Barkley, 2024).

These results highlight two key insights. First, selecting between large- and small-parameter models for complex tasks such as multilingual symptom detection involves a significant trade-off between performance and cost-effectiveness. Second, even large-parameter models can struggle on lower-resourced languages like Indonesian and Filipino. Addressing these challenges is crucial to improving the reliability and applicability of LLMs in disease surveillance.

5.1.3 Symptom-based Analysis

Figure 2(c) shows the distribution of F1-scores for each symptom, with the points representing the models. Respiratory symptoms generally exhibit high variability in scores, with RUN and FLU particularly standing out due to their lower averages compared to other symptoms. Furthermore, a significant difference was observed among digestive, systemic, and respiratory symptoms. Their mean F1 scores were 0.84, 0.78, and 0.62, respectively, with all pairwise p-values below 0.000 (See Appendix A.7).

Outliers in Figure 2(a) and 2(b) were from FLU predictions, which occurred across all languages and most models. Low outliers in multiple languages and models suggest that this symptom poses challenges, leading to sharp drops in performance for these instances.

As detailed in Appendix A.4, symptom scores vary across different languages and language groups. All European languages scored above average and all Asian languages scored below average for HAC and FLU. A similar pattern is observed for RUN except for Japanese. For other symptoms, scores for at least one language fell outside the standard deviation range. Notably, CLD scores for Indonesian and Filipino and were significantly lower than those of other languages, while RUN scores were significantly higher. CGH scores for Arabic and Filipino were also lower, while scores for Indonesian were higher.

Score differences like these may be due to cultural variations in how symptoms are named or described in different languages. Some symptoms may have no direct counterparts in some languages, resulting in the use of catch-all terms that can apply to multiple symptoms depending on the context. In Filipino, then term *sipon* can refer to either having a cold or a runny nose. Additionally, expressions and colloquial terms may instead be used, such as *meler* which can be referred as Runny nose, *meriang* and *masuk angin* in Indonesian which describe feeling unwell, including having a cold (Anggoro and Jee, 2021). This ambiguity makes it challenging for LLMs to identify specific symptoms, especially if pretraining was done primarily on formal texts.

Thus, to address these challenges, expanding training datasets to include informal and colloquial expressions is crucial for enhancing model robustness across diverse linguistic contexts, especially in digital disease surveillance, where social media data often contains local terms used by the public to describe symptoms. Additionally, fine-tuning models for underrepresented languages and cultural contexts can help bridge performance gaps and improve the accuracy of symptom detection across languages.

5.2 Surveillance Perspective: Relative Distance Insights

Table 3 shows that models tend to overestimate or label symptoms as positive for most languages, likely due to the dataset being imbalanced for all symptoms. Even so, RD scores for FLU are much higher than other symptoms for all languages, especially Asian ones. For example, the RD score for FLU in Japanese is 5.019, indicating that the predicted positive labels are five times higher than the actual positive cases. When combined with the low F1-scores for this symptom (Figure 2(c)), this suggests that LLMs are overly cautious by overpre-

Symptom	Language							Average
	English	German	French	Japanese	Arabic	Indonesian	Filipino	
FEV	-0.047	-0.002	0.011	0.084	0.032	0.090	0.038	0.030
RUN	-0.139	-0.005	0.218	-0.023	-0.071	0.113	0.161	0.036
CGH	0.268	0.246	0.257	0.324	0.178	0.173	0.462	0.273
DIA	0.235	0.309	0.290	0.302	0.253	0.323	0.394	0.301
CLD	0.290	0.270	0.491	0.457	0.356	0.075	0.249	0.313
HAC	0.293	0.271	0.301	0.426	0.498	0.470	0.362	0.374
HAY	0.442	0.582	0.568	0.682	0.437	0.564	0.539	0.545
FLU	**2.852**	**3.250**	**3.313**	**5.019**	**4.519**	**4.764**	**3.977**	**3.596**
Average	0.524	0.615	0.684	0.909	0.775	0.821	0.773	0.729

Table 3: Relative distance (RD) scores by language and symptom. Shaded scores are beyond ± 0.2, showing that LLMs overestimate positive cases for most symptoms and languages, while bold numbers presenting the highest overestimation of symptom prediction in each language.

dicting at the cost of performance. On the other hand, RD scores for FEV are close to 0 for all languages, indicating minimal bias for this symptom.

Traditional case-based surveillance systems are generally affected by some degree of underestimation, such as individuals attempting to self-treat their symptoms or institutions underreporting the cases. Final estimates usually have to be adjusted to capture a more accurate picture of disease incidence (Gibbons et al., 2014). Our findings suggest that text-based digital epidemiological systems, especially LLMs which are trained on large amounts of data, may have an advantage over traditional systems in this regard.

However, overestimation can lead to a loss of public trust or factuality in digital epidemiological disease surveillance, especially when underlying algorithms are not replicable or are hard to interpret. The validity of the now-defunct Google Flu Trends, which used search query data for predictions, was questioned after it overestimated peak flu levels during the 2012/2013 epidemic season by nearly double the actual figures (Butler, 2013; Olson et al., 2013). Mitigating overestimation bias by fine-tuning for specific symptoms or languages is recommended when deploying LLM-based surveillance systems.

In summary, there are two insights that can be drawn. First, Overestimation Tendency: LLMs exhibit a tendency to overestimate symptoms. It means that they are inclined to label symptoms as positive. From this observation, there are two key lessons: (1) if an LLM is deployed as a symptom identification and its results indicate a critical or dangerous situation, this may not necessarily reflect the actual case. This highlights that overestimation or misclassification of symptom mentions could lead to false alarms or public misinformation; however (2) if the system identifies a situation as safe, this can be considered reliable and trustworthy. These insights contribute to the practical application of LLM as a symptom identification system. Second, Influenza Detection Challenges: Detecting diseases similar to Influenza, such as COVID-19, using a symptom-based disease surveillance system with LLMs can result in poor performance. One possible contributing factor is the limited capabilities of LLMs in handling multilingual task, particularly in medical-related content in underrepresented languages. This limitation may lead to inaccurate symptom identification which can affect the factual accuracy of detected disease signal.

6 Conclusion

This paper evaluates LLM performance in symptom detection across different languages, LLMs and symptoms as the first crucial steps in health-related fact-checking data. In terms of (1) languages, our experiments show that European languages outperform Asian languages, particularly the SEA languages Indonesian and Filipino. Then, (2) LLMs achieve moderate to high performance overall, but varies significantly across languages and symptoms especially for small parameter models. As for (3) symptoms, respiratory symptoms are notably challenging for LLMs to predict accurately, with influenza being significantly overpredicted across all languages. These findings underscore the

potential of LLMs in digital epidemiology, while at the same time highlighting the need to address performance gaps in lower-resourced languages before practical implementation. We acknowledge that commercial LLMs are helpful, but adapting them to the public health field is likely needed for high-risk tasks like disease surveillance which can be explored in the future research. Moreover, symptom detection for medical fact-checking becomes critical to ensure that the early outbreak detection system align with real-world health conditions and are not based on misclassified or incomplete symptom data.

Limitations

This paper evaluates multiple languages, models, and symptoms in assessing LLMs for symptom detection for enhancing medical fact-checking, with a particular focus on the performance of Southeast Asian languages compared to high-resource ones. However, many key Southeast Asian languages remain unaddressed, including highly under-resourced languages such as Khmer, Burmese, and Lao. While this study examines LLM performance in symptom identification across languages, it does not propose new methods to enhance LLM performance. Instead, it aims to highlight the potential applications of LLMs in fact-checking for disease surveillance systems. Additionally, our approach to disease surveillance relies on identifying common symptoms from social media, which may be self-diagnosed by users. The models used also do not incorporate a large language model (LLM) specifically designed for the target language or fine-tuned for languages within the region. Furthermore, as the posts analyzed were translated rather than directly sourced from online platforms, they may not fully capture the linguistic and cultural nuances of how native speakers communicate in their own language online. To address these limitations, future iterations of this study will expand the evaluation by covering more languages and models, providing a more comprehensive assessment of multilingual LLM performance in digital epidemiology.

Ethics Statement

This study is an extension of the NTCIR-13 Task, utilizing its test dataset with the consent from the task organizers. The dataset used was pseudotweets since the tweet data obtained via the Twitter API cannot be publicly shared due to Twitter's developer policy on data redistribution. The dataset, originally in Japanese, was generated through crowdsourcing and subsequently translated into six other languages by human translators to minimize bias in machine translation. Additionally, all resources used in this study comply with their respective licenses. We have authorized API access to the resources, strictly for research purposes, and have fully complied with all terms and conditions. No personally identifiable information (PII) was included, and the research does not involve human subjects requiring IRB approval.

References

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2637.

Florencia K. Anggoro and Benjamin D. Jee. 2021. The substance of cold: Indonesians' use of cold weather theory to explain everyday illnesses. *Frontiers in Psychology*, 12.

Warren Barkley. 2024. New strides in making ai accessible for every enterprise. Accessed: 19 May 2025.

Azadeh Bayani, Alexandre Ayotte, and Jean Noel Nikiema. 2025. Transformer-based tool for automated fact-checking of online health information: Development study. *JMIR Infodemiology*, 5:e56831.

Sangeeta Bhatia, Britta Lassmann, Emily Cohn, Angel N Desai, Malwina Carrion, Moritz U G Kraemer, Mark Herringer, John Brownstein, Larry Madoff, Anne Cori, and Pierre Nouvellet. 2021. Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread. *npj Digital Medicine*, 4(1):73.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Declan Butler. 2013. When Google got flu wrong. *Nature*, 494:155–156.

Robert Dale. 2021. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Gemini Team Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Cheryl L. Gibbons, Marie-Josée J. Mangen, Dietrich Plass, Brooke Russell John Havelaar, Arie H., Piotr Kramarz, Karen L. Peterson, Anke L. Stuurman, Alessandro Cassini, Eric M. Fèvre, and Mirjam EE. Kretzchmar. 2014. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*, 14.

Alekh Gour, Shikha Aggarwal, and Subodha Kumar. 2022. Lending ears to unheard voices: An empirical analysis of user-generated content on social media. *Production and Operations Management*, 31(6):2457–2476.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. Accessed: 19 May 2025.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference*, pages 2627–2638. Association for Computing Machinery.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Chodhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Logan Kilpatrick and Shrestha Basu Mallick. 2024. Gemini 1.5 flash-8b is now production ready. Accessed: 19 May 2025.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Jann Railey Montalan, Ryan Ignatius Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Chandra Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

MedWeb. 2017. Ntcir-13 medweb annotation corpus guideline. Accessed: 19 May 2025.

Mistral. 2025. Model weights. Accessed: 19 May 2025.

Mistral AI. 2024. Au large. Accessed: 19 May 2025.

Ai Ohtani, Takefumi Suzuki, Hiroyoshi Takeuchi, and Hiroyuki Uchida. 2015. Language barriers and access to psychiatric care: A systematic review. *Psychiatric Services*, 66(8):798–805.

Donald R. Olson, Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. 2013. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLOS Computational Biology*, 9(10).

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propogate race-based medicine. *npj Digital Medicine*, 6(195).

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 19 May 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin,

Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Tanmay Parekh, Jeffrey Kwan, Jiarui Yu, Sparsh Johri, Hyosang Ahn, Sreya Muppalla, Kai-Wei Chang, Wei Wang, and Nanyun Peng. 2024a. SPEED++: A multilingual event extraction framework for epidemic prediction and preparedness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12936–12965, Miami, Florida, USA. Association for Computational Linguistics.

Tanmay Parekh, Anh Mac, Jiarui Yu, Yuxuan Dong, Syed Shahriar, Bonnie Liu, Eric Yang, Kuan-Hao Huang, Wei Wang, Nanyun Peng, and Kai-Wei Chang. 2024b. Event detection from social media for epidemic prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5758–5783, Mexico City, Mexico. Association for Computational Linguistics.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMediX: Bilingual medical mixture of experts LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002, Miami, Florida, USA. Association for Computational Linguistics.

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. 2024. Bias patterns in the application of LLMs for clinical decision support: A comprehensive study. *Preprint*, arXiv:2404.15149.

T Purnat, M Kajimoto, J Kalinic, A Stevanovic, S Mandic-Rajcevic, and E Wilhelm. 2024. How factchecking organizations can partner within public health for a healthier internet. *European Journal of Public Health*, 34.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and

Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. JMLR.org.

Piu Samui, Jayanta Mondal, and Subhas Khajanchi. 2020. A mathematical model for covid-19 transmission dynamics with a case study of india. *Chaos, Solitons Fractals*, 140:110173.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara C. Schouten, Antoon Cox, Gözde Duran, Koen Kerremans, Leyla Köseoğlu Banning, Ali Lahdidioui, Maria van den Muijsenbergh, Sanne Schinkel, Hande Sungur, Jeanine Suurmond, Rena Zendedel, and Demi Krystallidou. 2020. Mitigating language and cultural barriers in healthcare communication: Towards a holistic approach. *Patient Education and Counseling*, 103(12):2604–2608.

Hilal Al Shamsi, Abdullah G. Alumtairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of language barriers for healthcare: A systematic review. *Oman medical journal*, 35(2).

Elham Sharifpoor, Maryam Okhovati, Mostafa Ghazizadeh-Ahsaee, and Mina Avaz Beigi. 2025. Classifying and fact-checking health-related information about COVID-19 on Twitter/X using machine learning and deep learning models. *BMC Medical Informatics and Decision Making*, 25(1):73.

Boyang Shi, Weixiang Huang, Yuanyuan Dang, and Wenhui Zhou. 2024. Leveraging social media data for pandemic detection and prediction. *Humanities and Social Sciences Communications*, 11(1):1075.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

Montse Vázquez-Gestal, Jesús Pérez-Seoane, and Ana-Belén Fernández-Souto. 2024. Disinformation and health: fact-checking strategies of spanish health public institutions through youtube. *Frontiers in Communication*, Volume 9 - 2024.

Brian Wahl, Aline Cossy-Gantner, Stefan Germann, and Nina R Schwalbe. 2018. Artificiall intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Global Health*, 3(4).

Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2017. Overview of the NTCIR-13 MedWeb Task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-13)*, pages 40–49.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *Preprint*, arXiv:2403.03640.

Xin Shelley Wang, Charles S. Cleeland, Tito R. Mendoza, Young Ho Yun, Ying Wang, Toru Okuyama, and Valen E. Johnson. 2010. Impact of cultural and linguistic factors on symptom reporting by patients with cancer. *JNCI: Journal of the National Cancer Institute*, 102(10):732–738.

Wang. Xiang-Sheng and Luoyi Zhong. 2015. Ebola outbreak in West Africa: real-time estimaion and multiple-wave prediction. *Mathematical Biosciences and Engineering*, 12(5).

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, 4(176).

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2024. How vocabulary sharing facilitates multilingualism in LLaMA? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12111–12130, Bangkok, Thailand. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

A Appendix

A.1 Instruction to the Human Translation Services

Method

The text written in cells B2 to B641 of the attached

Excel file will be translated into the following languages. Text written in cells other than column B will not be translated.
• Indonesian
• Filipino
• Arabic
• German
• French

The target text consisted of 640 sentences with a total of 8,276 words. The manuscript tweet data is available in two files, one written in Japanese and one written in English, with the same content.

Delivery form

Excel data (any data format can be delivered via email attachment or file sharing system) Please create Excel data according to the following procedure.
• Separate files for each language
• Edit only columns A and B, and copy the values from the manuscript to columns C to J.
• Column A (ID)
• Enter the same number + language code as the original
• Example: If the row in column A of a Japanese manuscript with "1921ja" is to be translated into German, the ID of column A in the German Excel file should be "1921de". For language codes, see below https://mt-auto-minhon-mlt.ucri.jgn-x.jp/content/help/detail.html?q_pid=FAQ_ETC • Column B (Tweet)
• Enter the translated text.

Delivery Date

Friday, November 29, 2024

A.2 Complete Prompts

Instruction:

Determine if the sender of this Twitter message is exhibiting symptoms for each of the following: influenza, diarrhea, hay fever, cough, headache, fever, runny nose, cold. For each symptom, only answer either 0 or 1 for negative (no symptoms) or positive (has symptoms) respectively. Determination of symptoms is carried out based on the following rules:
• Cases where the symptom is expressed directly including mild symptoms are considered positive;
• A symptom can be labeled positive with indirect expressions of a symptom;
• If a symptom is mentioned but then also dismissed or denied, this information is regarded as positive;
• It is considered positive if someone or the user is still affected with such mild symptoms during recovery. However, if the symptoms are completely gone, it is considered negative;
• A symptom is positive even if the user expresses uncertainty regarding its cause;
• Since it is generally presumed that many patients may overlook symptoms or diseases due to insufficient medical knowledge, even suspicion of symptoms and diseases are recognized and labeled positive;
• Symptoms that disappeared completely are recognized and labeled negative. Note that we regarded and labeled positive when a user took medicine that could cause temporary recovery from a symptom;
• For cases that express expectation or process, indicated with words such as “if,” “going,” “if it is,” etc., these should be labeled as negative;
• If the disease is mentioned merely as a topic rather than someone having it, these tweets should be labeled as negative. These include news, general theories, and advertisements;
• If the disease is mentioned in the context of a joke, these should be labeled as negative;
• The symptoms are only for humans;
• Symptoms are within 24 hours including today;
• The label for symptoms that occurred yesterday are dependent on the disease or symptom;
• Past Symptoms Including Two or More Days Ago considered as negative;
• Recent occurrence and Recurring Symptom that Still Persists considered as positive;
• We regard as a symptom in the vicinity and label positive regardless of living together or not (i.e. family members). We also label positive when a symptom was observed from hearsay;
• As for symptoms of people belonging to a specified group in the vicinity (school, club, etc.), we labeled them positive.
• Other cases, like symptoms belonging to blog friends, should be labeled as negative since it is difficult to determine their location.

Post:

A post in one of the seven studied languages.

Return the result STRICTLY as a JSON object with the symptoms as keys and the values as either

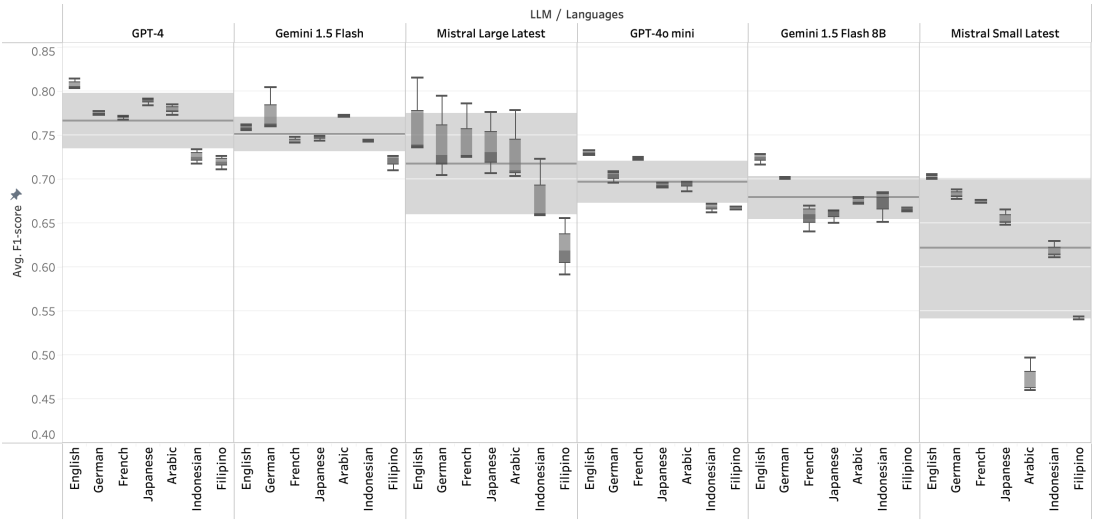


Figure 3: F1-Score Distribution of Each LLM Across Different Language.

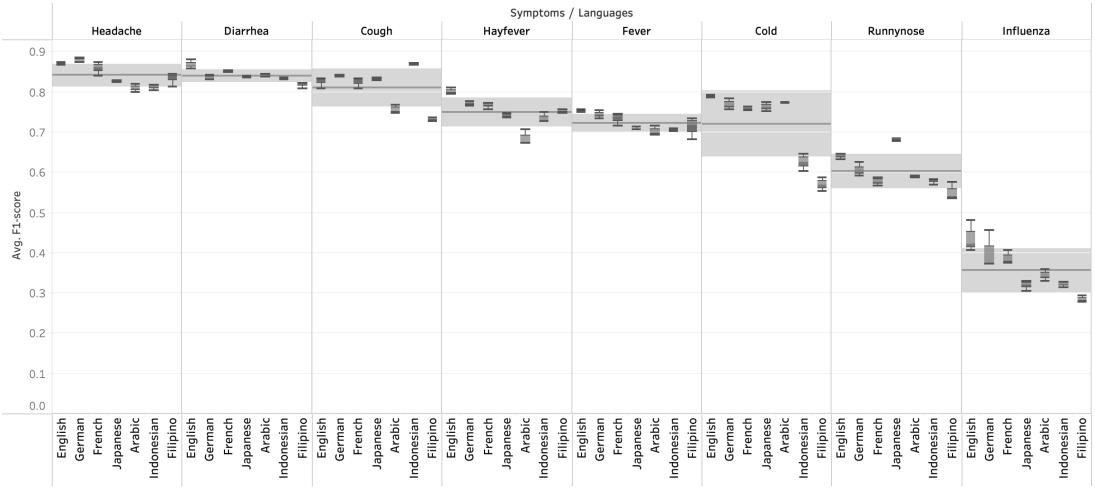


Figure 4: F1-Score Distribution of Each Symptom Across Different Language

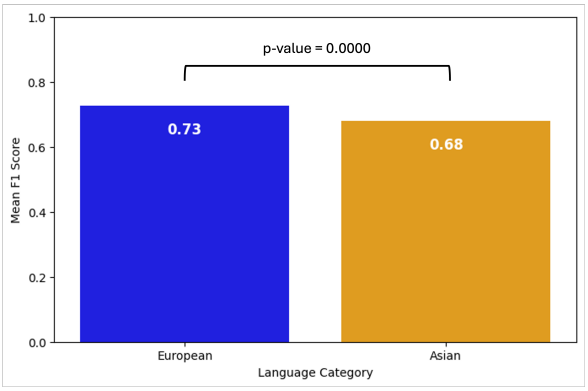


Figure 5: One-way ANOVA Test in Language Category

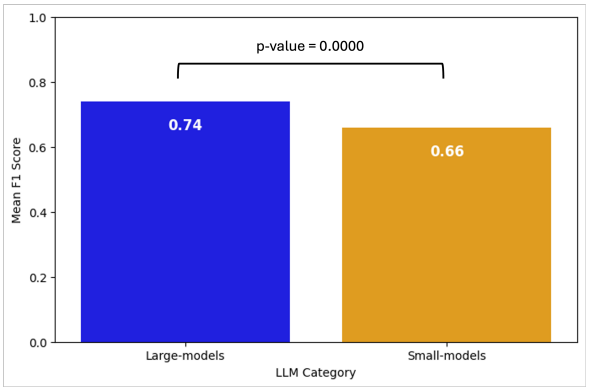


Figure 6: One-way ANOVA Test in LLM Category

0 or 1.

A.3 F1-Score Distribution of Each LLM Across Different Language

Figure 3 illustrates the F1 scores of each LLM across different languages, aiming to analyze per-

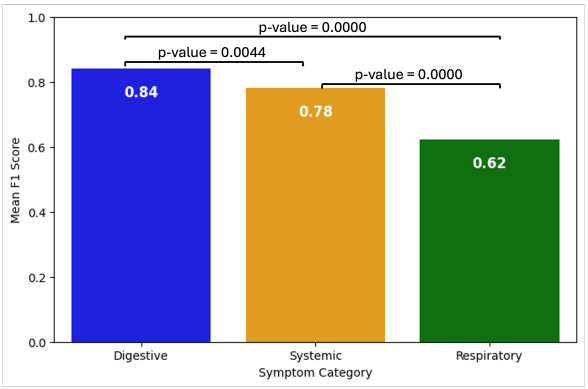


Figure 7: One-way ANOVA Test in Symptom Category

formance variations across languages.

A.4 F1-Score Distribution of Each Symptom Across Different Language

Figure 4 provides the F1-scores of each symptom in various languages, with the goal of examining performance differences across languages.

A.5 Statistical Testing on F1-Score for Language Categories

Figure 5 shows the mean comparison of F1-score for language categories to determine significant statistical differences.

A.6 Statistical Testing on F1-Score for LLM Categories

Statistical testing was presented in Figure 6 to compare average F1-score for LLM categories.

A.7 Statistical Testing on F1-Score for Symptom Categories

The comparison of mean F1 scores across symptoms was analyzed to identify statistically significant differences as presented in Figure 7.