Open-source solution for evaluation and benchmarking of large language models for public health

Laura Espinosa[1*], Djilani Kebaili[1], Sergio Consoli[2], Kyriaki Kalimeri[3], Yelena Mejova[3], Marcel Salathé[1]

Affiliations:

1 Digital Epidemiology Lab, EPFL, Geneva, Switzerland.

2 Joint Research Centre, European Commission, Ispra, Italy.

3 ISI Foundation, Turin, Italy.

*Corresponding author: laura.espinosamontalban@epfl.ch

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing tasks, including text classification, information extraction, and sentiment analysis. Despite their increased usage and development, there are no standardized benchmarking frameworks for domain-specific applications. The objective of this study is to develop and test an open-source solution that enables an intuitive, easy and rapid benchmarking of open-source and commercial LLMs for public health use cases.

An LLM prediction prototype was developed, supporting multiple open-source and commercial LLMs. It enables users to upload datasets, apply prompts, and generate structured JSON outputs for LLM tasks. The application prototype built with R and the Shiny library facilitates automated LLM evaluation and benchmarking by computing key performance metrics for categorical and variables. We tested these prototypes on four public health use cases: stance detection towards vaccination in tweets and Facebook posts, detection of vaccine adverse reaction in BabyCenter forum posts, and extraction of epidemiological quantitative features from the World Health Organization Disease Outbreak News.

Results revealed high variability in LLM performance depending on the task, dataset, and model, with no single LLM consistently outperforming others across all tasks. While larger models generally excelled, smaller models performed competitively in specific scenarios, highlighting the importance of task-specific model selection.

This study contributes to the effective LLM integration in public health by providing a structured, user-friendly and scalable solution for LLM prediction, evaluation and

benchmarking. Our results underline the relevance of standardised and task-specific evaluation methods and model selection, and the use of clear and structured prompts to improve LLM performance in domain-specific use cases.

1. Introduction

Large language models (LLMs) are a type of language model that use neural networks with billions of parameters and are trained on a vast amount of unlabeled data using a self-supervised approach [1] based on Transformers [2]. This has revolutionised their application in sophisticated and general-purpose language tasks, where traditional natural language processing techniques and other language models struggled to achieve high performance without burdensome and costly fine-tuning approaches. LLMs excel in tasks such as text translation and summarisation, data extraction and sentiment analysis, which are relevant across various domains, including healthcare and public health [3–5]. Furthermore, previous studies have shown the increased performance of LLMs in comparison to other machine learning approaches [6–8].

Since 2022, the number of newly released LLMs has increased exponentially [9], requiring continuous benchmarking to ensure the adoption of cutting-edge technology. However, benchmarking standards are not always applicable to public health. Several studies have assessed the performance of specific LLMs for concrete health or public health use cases [4,6,10,11]. A systematic review on testing and evaluation of healthcare applications of LLMs between 2022 and 2024 concluded that current evaluations are fragmented and insufficient [9]. Additionally, Reddy (2023) proposed a framework for evaluating LLMs in healthcare, which included human assessment of results and emphasised the need to evaluate other aspects such as utility, ethics, and functionality [12].

For general benchmarking of LLMs performance in language skills, mathematical problem-solving and conversational abilities, standardised tests have been developed to compare different models [13]. While these tests promote a harmonised approach to LLMs evaluation, these are not always available for domain-specific application and are not always directly applicable to public health. Given these challenges, a user-friendly tool for LLMs prediction, evaluation and benchmarking is essential to ensure successful implementation in the public health field. Furthermore, when developing a prototype, testing it with real examples is crucial for assessing its performance before full deployment.

The objective of this study is to develop and test an open-source solution that enabled an intuitive, easy and rapid benchmarking of open-source and commercial LLMs for public health use cases. Following the key tasks that LLMs are expected to perform, we selected four public health use cases to evaluate the performance of specific LLMs in their respective objectives.

2. Materials and methods

This study consisted of three major pillars: (i) LLMs prediction prototype for categorical and numerical variables, (ii) Shiny app prototype for LLMs evaluation and benchmarking, and (iii) four different public health use cases for testing the first two pillars.

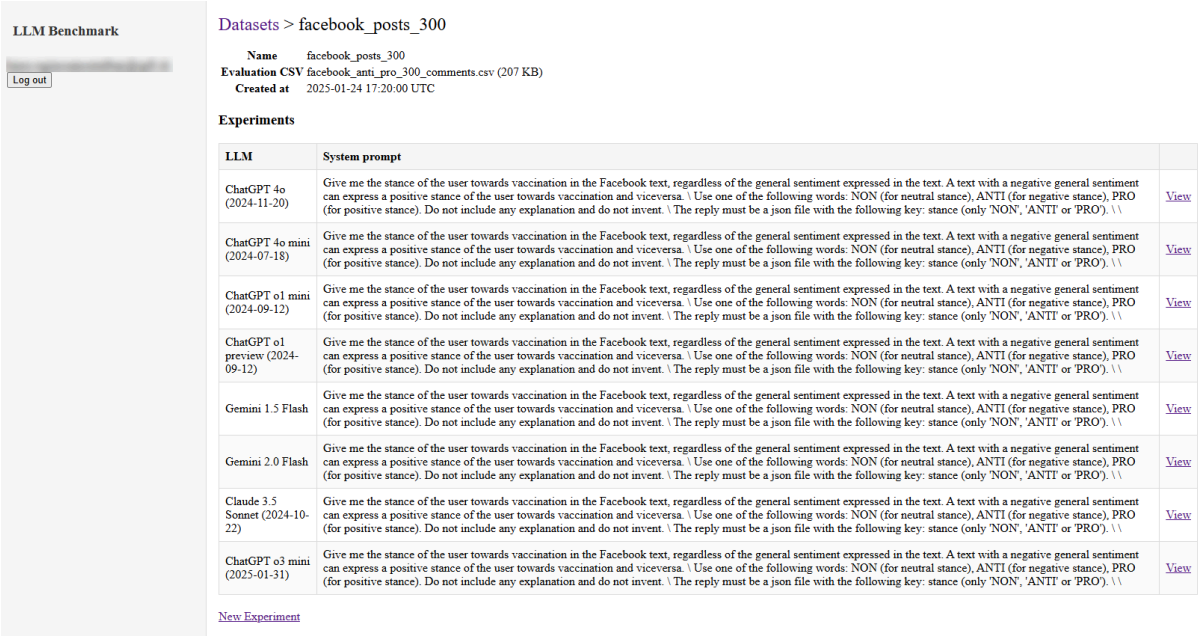
2.1. LLM prediction prototype

We built an LLM prediction prototype based on Ruby 3.4.2 as the scripting language and Ruby on Rails 8.0.1 and HTML for the web application. The application has three main components: datasets, prompts and LLMs.

The datasets must be CSV files with at least one column named "##_MSG_##_TEXT" which is the text from which the LLM will predict the desired categorical or numerical variable (e.g., stance towards vaccination from social media posts or number of cases reported in a short document). The file can contain other columns which will be ignored by the prototype. The users can upload as many datasets as needed. It will also allow images as input in the next releases.

Each dataset can have one or more experiments which is the combination of a specific prompt and single LLM (Fig. 1). The application loops for each row of the designated column in the dataset, adding the specified prompt (provided by the user in plain text) and sending these to the Application Programming Interface (API) endpoint of the specified LLM. The prototype forces the LLM to give a JSON file as output; however, it is recommended to include clear instructions as well in the prompt on the requirement to provide a JSON file as output and on the keys or variables that should be provided for the prediction.

To date, the prototype includes the following LLMs: Generative Pre-trained Transformer (GPT) 4o (endpoint 2024-11-20), GPT 4o mini (endpoint 2024-07-18), GPT o1 preview (endpoint 2024-09-12), GPT o1 mini (endpoint 2024-09-12) and GPT o3 mini (endpoint 2025-01-31) as commercial models, and Gemini 1.5 Flash, Gemini 2.0 Flash, Deepseek R1 (Distilled Qwen 32B via Hugging Face), and Claude 3.5 Sonnet (endpoint 2024-10-22) as open-source models. The specific cost associated with the use of a commercial LLM needs to be covered by each user, and the corresponding tokens for accessing the related model endpoints should be inserted in the application accordingly.



LLM Benchmark

Log out

Datasets > facebook_posts_300

Name facebook_posts_300
Evaluation CSV facebook_anti_pro_300_comments.csv (207 KB)
Created at 2025-01-24 17:20:00 UTC

Experiments

LLM	System prompt	
ChatGPT 4o (2024-11-20)	Give me the stance of the user towards vaccination in the Facebook text, regardless of the general sentiment expressed in the text. A text with a negative general sentiment can express a positive stance of the user towards vaccination and viceversa. \ Use one of the following words: NON (for neutral stance), ANTI (for negative stance), PRO (for positive stance). Do not include any explanation and do not invent. \ The reply must be a json file with the following key: stance (only 'NON', 'ANTI' or 'PRO'). \ \	View
ChatGPT 4o mini (2024-07-18)	Give me the stance of the user towards vaccination in the Facebook text, regardless of the general sentiment expressed in the text. A text with a negative general sentiment can express a positive stance of the user towards vaccination and viceversa. \ Use one of the following words: NON (for neutral stance), ANTI (for negative stance), PRO (for positive stance). Do not include any explanation and do not invent. \ The reply must be a json file with the following key: stance (only 'NON', 'ANTI' or 'PRO'). \ \	View
ChatGPT o1 mini (2024-09-12)	Give me the stance of the user towards vaccination in the Facebook text, regardless of the general sentiment expressed in the text. A text with a negative general sentiment can express a positive stance of the user towards vaccination and viceversa. \ Use one of the following words: NON (for neutral stance), ANTI (for negative stance), PRO (for positive stance). Do not include any explanation and do not invent. \ The reply must be a json file with the following key: stance (only 'NON', 'ANTI' or 'PRO'). \ \	View
ChatGPT o1 preview (2024-09-12)	Give me the stance of the user towards vaccination in the Facebook text, regardless of the general sentiment expressed in the text. A text with a negative general sentiment can express a positive stance of the user towards vaccination and viceversa. \ Use one of the following words: NON (for neutral stance), ANTI (for negative stance), PRO (for positive stance). Do not include any explanation and do not invent. \ The reply must be a json file with the following key: stance (only 'NON', 'ANTI' or 'PRO'). \ \	View
Gemini 1.5 Flash	Give me the stance of the user towards vaccination in the Facebook text, regardless of the general sentiment expressed in the text. A text with a negative general sentiment can express a positive stance of the user towards vaccination and viceversa. \ Use one of the following words: NON (for neutral stance), ANTI (for negative stance), PRO (for positive stance). Do not include any explanation and do not invent. \ The reply must be a json file with the following key: stance (only 'NON', 'ANTI' or 'PRO'). \ \	View
Gemini 2.0 Flash	Give me the stance of the user towards vaccination in the Facebook text, regardless of the general sentiment expressed in the text. A text with a negative general sentiment can express a positive stance of the user towards vaccination and viceversa. \ Use one of the following words: NON (for neutral stance), ANTI (for negative stance), PRO (for positive stance). Do not include any explanation and do not invent. \ The reply must be a json file with the following key: stance (only 'NON', 'ANTI' or 'PRO'). \ \	View
Claude 3.5 Sonnet (2024-10-22)	Give me the stance of the user towards vaccination in the Facebook text, regardless of the general sentiment expressed in the text. A text with a negative general sentiment can express a positive stance of the user towards vaccination and viceversa. \ Use one of the following words: NON (for neutral stance), ANTI (for negative stance), PRO (for positive stance). Do not include any explanation and do not invent. \ The reply must be a json file with the following key: stance (only 'NON', 'ANTI' or 'PRO'). \ \	View
ChatGPT o3 mini (2025-01-31)	Give me the stance of the user towards vaccination in the Facebook text, regardless of the general sentiment expressed in the text. A text with a negative general sentiment can express a positive stance of the user towards vaccination and viceversa. \ Use one of the following words: NON (for neutral stance), ANTI (for negative stance), PRO (for positive stance). Do not include any explanation and do not invent. \ The reply must be a json file with the following key: stance (only 'NON', 'ANTI' or 'PRO'). \ \	View

New Experiment

Figure 1. Screenshot of the LLM prediction prototype with the list of experiments for one of the public health use cases included in this study.

2.2. Shiny app prototype for LLM evaluation and benchmarking

We have developed a Shiny app for LLMs evaluation and benchmarking, after predicting the value of the variable(s). This app has been built with R Studio 2023.06.1 and R version 4.4.2, and it is available as open-access at https://github.com/digitalepidemiologylab/llm_benchmarking.

The app has three pages: instructions, and LLMs evaluation and benchmarking of categorical and numerical variables. The instructions page includes information on the tool, requirements and step-by-step instructions. The other two pages contain two main panels: left panel, to include all the parameters needed to evaluate the LLMs results and the steps required to process the data and calculate the performing metrics according to the variable type; and right panel, showing the outcome and results of each

step. Each page has unique parameters and results, thus information from one page is not automatically transferred to another page.

The app requires two files: a CSV file with the labelled variable and a column with a unique identifier for each entry, and a CSV or JSON file with the LLM predicted variable and the same unique identifier for each entry as the first file. The second file can be the output from the LLM prediction prototype or from any other tool.

The parameters required to be provided by the user are: paths to the dataset; LLM predictions; folder where to store the evaluation and benchmarking results; specific LLM that has been used; name of the target and LLM predicted columns; and name of the specific experiment.

The app includes a series of steps for evaluating and benchmarking LLMs. It begins by importing the labeled and the LLM prediction datasets. These datasets are then merged using a unique identifier for each entry. For categorical variables, the evaluation is done by calculating the confusion matrix which includes sensitivity, specificity, positive predictive value, negative predictive value, precision, recall, F1 score, detection rate, detection prevalence, and balanced accuracy for each class, as well as overall accuracy. Additionally, a 95% confidence interval is provided for the overall accuracy with a one-sided test to determine if the accuracy was better than selecting the majority class. For numerical variables, the evaluation includes accuracy, root-mean-square error (RMSE), and mean absolute error (MAE) as metrics.

2.3. Public health use cases

Four different public health use cases were selected to test the functionality and usability of the LLM prediction and Shiny app for LLM evaluation and benchmarking

prototypes. These use cases were: (i) predicting stance towards vaccination in English tweets, (ii) predicting stance towards vaccination in Facebook posts, (iii) detecting vaccine adverse reactions from the digital parenting forum BabyCenter United States (US), and (iv) extracting relevant epidemiological information from the World Health Organization (WHO) Disease Outbreak News (DON), The prompts used for each public health use case are available in the supporting information (S1 Text) and were extracted from previous publications using the same data when available.

2.3.1. Stance towards vaccination in English tweets

The objective of this use case is to predict the stance toward vaccination (neutral, negative or positive) in 1,000 random English-language tweets on vaccination collected between 2nd of December, 2019 and 11th of March, 2022. The stance towards vaccination may differ from the general sentiment of the tweet. The gold standard dataset was labelled by a convenience sample of four experts from the EPFL Digital Epidemiology Lab: two public health experts, one expert annotator, and one expert in administration as a representative of the general population [6].

2.3.2. Stance towards vaccination in Facebook comments

The objective of this use case is to predict the stance towards vaccination in Facebook comments from pages supportive or contrary to vaccination. Facebook pages containing the keywords 'vaccine', 'vaccines' or 'vaccination' were queried using the Facebook Graph API between January 2012 and June 2019 by Beiró *et al.*, 2023 [11]. The gold standard dataset is a random sample of 300 Facebook comments (150 from pages supportive to vaccination and 150 from pages contrary to vaccination) labelled by nine annotators familiar with the moral foundation's theory.

2.3.3. Detecting vaccine adverse reactions from BabyCenter US forum

The objective of this use case is to detect the mention of vaccine adverse reactions (yes, no) from BabyCenter US forum posts and comments. BabyCenter is a digital platform for new or expectant parents that has more than 34 million monthly users worldwide from its eight websites [14]. Posts and comments with the keyword 'vaccine' from 11th of March, 2008 to 26th of April, 2019 were queried in the site search function by Betti *et al.*, 2021 [10] in compliance with the Terms of Use. The gold standard dataset was a random sample of 300 comments labelled by seven annotators familiar with the vaccination debate.

2.3.4. Extracting relevant features from World Health Organisation (WHO) Disease Outbreak News (DONs)

The objective of this use case is to extract relevant epidemiological features (i.e., affected country, date of the event, reported number of confirmed cases and deaths, and reported disease or pathogen) from WHO DONs. WHO DONs fulfill the requirement of the Article 11.4 of the International Health Regulation (2005) to make information on acute public health events available if there is a need for disseminating authoritative and independent information [15]. It is not an exhaustive list of public health events to which WHO is responding to. We selected a random sample of 300 WHO DONs between 30th of July, 1996 and 2nd of April, 2019. The used gold standard dataset was labelled by the Information Centre for International Health Protection at Robert Koch Institut [16,17].

2.5. Code and data sharing

Data and code used for the LLMs prediction prototype and evaluation and benchmarking can be found in the LLM benchmarking repository (https://github.com/digitalepidemiologylab/llm_benchmarking).

3. Results

3.1. Usability of the LLM prediction prototype and Shiny app for LLM evaluation and benchmarking

The two prototypes (LLM prediction and Shiny app for LLM evaluation and benchmarking) are open-source tools that can be used by other experts. The repositories of the prototypes include information on its use and all necessary code to execute them.

These prototypes have been developed with key features aiming towards user-friendliness and harmonisation. The LLM prediction prototype is designed to provide the same format as output (JSON file) which harmonises the LLM evaluation and benchmarking. It displays the prompts used for each experiment allowing reutilising these for other experiments or LLMs within the same experiment, and predictions can be easily downloaded after the experiment is finalised. The Shiny app prototype is versatile by allowing different types of inputs, following a predefined structure, including, but not only limited to, the JSON files from the LLMs prediction prototype, and modifiable parameters to be adapted to different experiments. Its user interface is built in R which allows for easy use without high demand of requirements.

3.2. Exploratory data analysis of the public health use cases

In this study, we have analysed the LLMs predictions for five categorical variables and two numerical variables from the four public health use cases. All categorical variables were imbalanced and numerical variables did not have a normal distribution (Fig. 2). The majority classes were “neutral” for stance towards vaccination in tweets and Facebook posts and “no” for mentions of vaccine adverse reactions in BabyCenter forum posts. There were 89 unique affected countries reported in WHO DONs and the most frequently reported were Democratic Republic of the Congo (COD) and Guinea (GIN), which each of them were reported in 6.3% of the news, and Liberia (LBR), Nigeria (NGA) and Uganda (UGA), which each of them were reported in 4% of the news. There were 42 unique diseases reported in WHO DONs and the most frequently reported were Ebola virus and cholera, which were reported in 22.0% and 19.7% of the news, respectively. The total number of cases reported in all WHO DONs included in this study had a median of 181 and an interquartile range (IQR) of 616,011. The total number of deaths reported in all WHO DONs included in this study had a median of 39 and an IQR of 3,881.

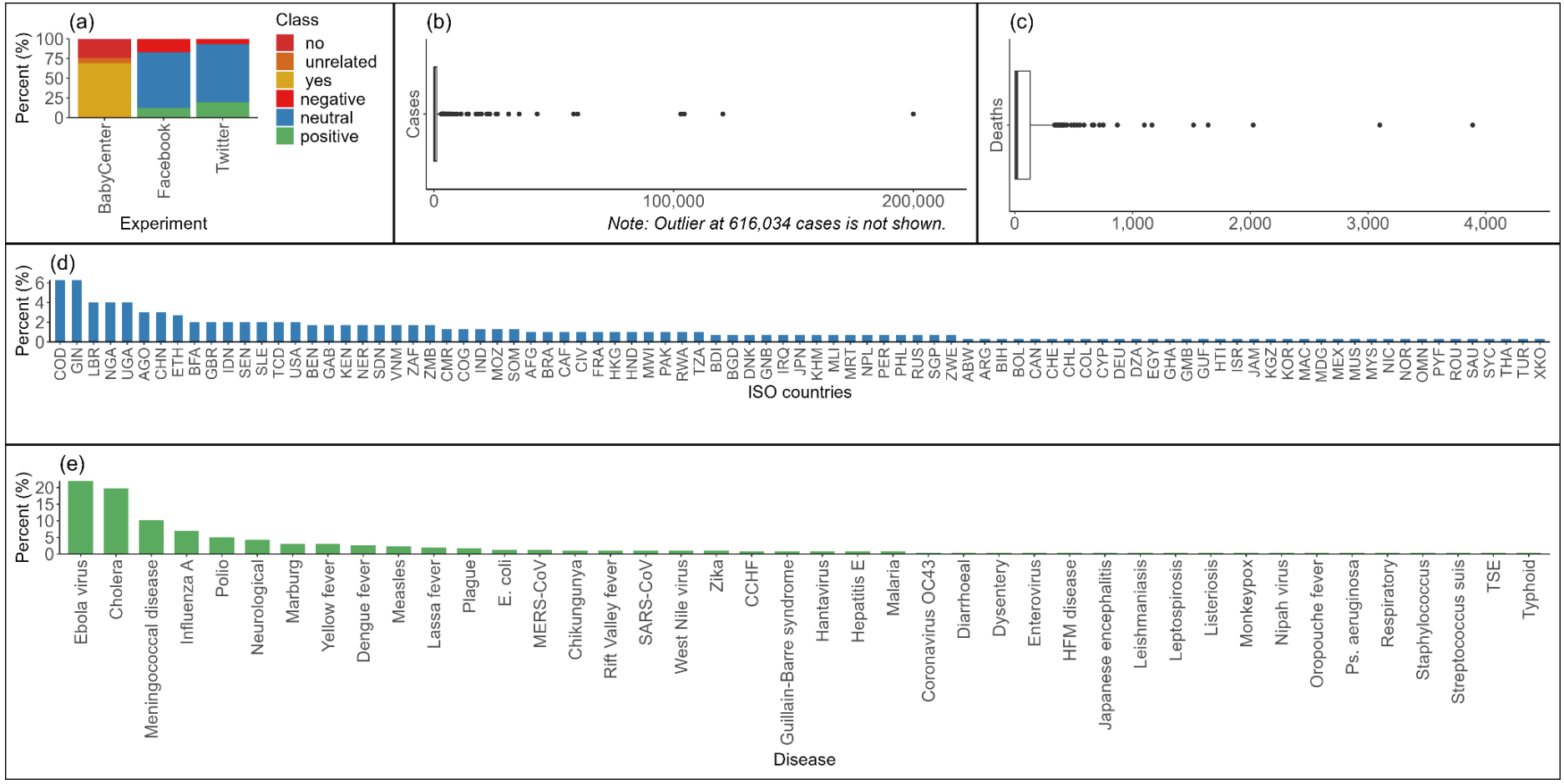


Figure 2. Distribution of classes and values of the seven variables analysed. (a) Class distribution for predictions from tweets, Facebook comments and BabyCenter posts, (b) Boxplot of cases reported in WHO DONs, (c) Boxplot of deaths reported in WHO DONs, (d) Affected countries reported in WHO DONs, and (e) Diseases or pathogens reported in WHO DONs.

3.3. LLM evaluation and benchmarking

The overall accuracy for the categorical and numerical variables differed depending on the LLM and experiment (Fig 3). The accuracy in predicting vaccine adverse reactions in BabyCenter posts was very stable among LLMs with a significant p-value (i.e., the predictions had a better accuracy than just selecting the majority class). The accuracy in predicting stance towards vaccination in Facebook posts and tweets varied according to the used LLM and, for the latter, most of the predictions had worse accuracy than just selecting the majority class. The best performing LLMs were Gemini 2.0 and GPT-4o for Facebook posts and GPT o1 mini and GPT o3 mini for tweets.

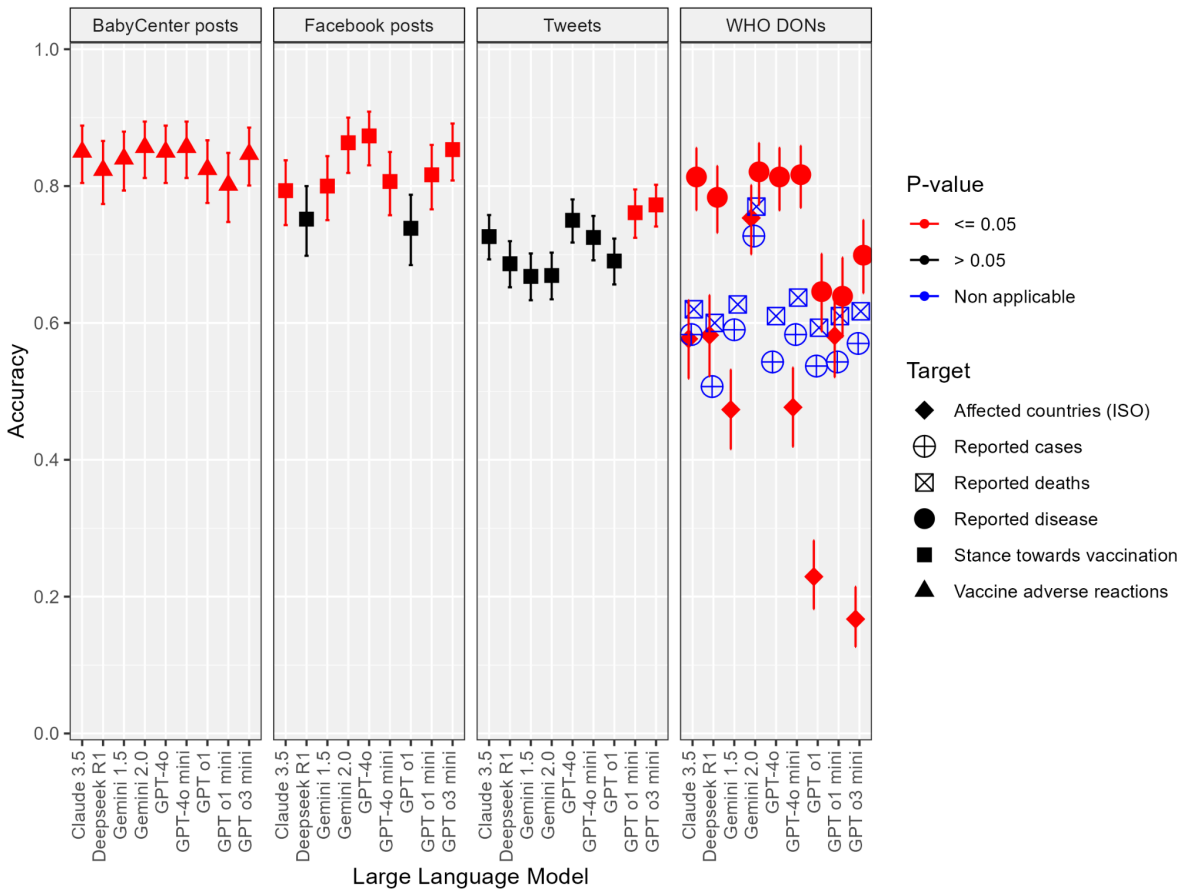


Figure 3. Accuracy for all target variables and 95% confidence interval for the numerical target variables per experiment, target and LLM

Some of the selected WHO DONs had epidemiological information on the first affected country but also included information on other countries. For example, the WHO DON on Ebola virus disease outbreak in West Africa [18] was first detected in Guinea but soon after, cases were reported in Sierra Leone and Liberia as well. GPT o1, GPT o1 mini, GPT-4o and Gemini 2.0 extracted features for all countries mentioned in the WHO DON. Furthermore, the nomenclature used by LLMs for the diseases and/or pathogen was not harmonised among diseases/pathogens and LLMs. In these cases, the LLM predictions were manually corrected by a public health expert and co-author of this study to align the nomenclature of diseases and/or virus with the labelled dataset and to select the extracted features from the initial or main country affected.

Despite requesting LLMs to provide the affected country using the three-letter ISO (International Organization for Standardization) code, there were some predictions with the full name of the country, two-letter ISO code or other abbreviations used (e.g., DRC instead of COD for the Democratic Republic of the Congo).

The overall accuracy for predicting reported disease in WHO DONs was very similar among LLM, except for GPT o1, o1 mini and o3 mini. The accuracy for predicting three-letter ISO countries differed among LLMs with Gemini 2.0 as the most accurate LLM. The accuracy for predicting reported cases and death was very similar among LLMs, except for Gemini 2.0 which had a higher accuracy than the other LLMs. Using the Hugging Face API for accessing the Deepseek model required a specific data cleaning for extracting the predictions since the outputs provided by the LLM included the reasoning from the model to reach a prediction which was not requested in the prompt or provided by other LLMs.

When evaluating the LLM prediction for each class of the categorical variables, the balanced accuracy, sensitivity, specificity and F1 score was different per each class in most of the cases, but the prevalence of each class in the predictions was very similar or equal to the prevalence of each class in the labelled dataset. In the predictions of vaccine adverse reactions in BabyCenter posts and stance towards vaccination in Facebook posts and tweets, the balanced accuracy for each class was very similar (Fig 4). Moreover, the specificity and sensitivity for each class of the Facebook posts was very similar. However, in the predictions from BabyCenter posts, the sensitivity and F1 score was higher for the majority class (yes) and the specificity was higher for the minority class (no); in the predictions from Facebook posts, the F1 score was higher for the majority class (non); and in the predictions from tweets, the sensitivity and F1 score were higher in the majority class (neutral) and stable among LLMs and differed in the other two classes depending on the LLM, and the specificity was lower in the majority class and differed depending on the LLM.

Since WHO DONs had a high number of classes for affected countries and diseases/pathogens, only those classes within the third quartile of frequency were assessed in terms of balanced accuracy, sensitivity, specificity and F1 score per each class of these categorical variables (Fig 5). Similarly to the other public health use cases, the prevalence of each class in the predictions was very similar or equal to the prevalence of each class in the labelled dataset. In the predictions of diseases/pathogens, the balanced accuracy for each class was very similar, except for Lassa fever and polio which had a 100% balanced accuracy, and GPT o1 and o1 mini which provided lower balanced accuracies for all predictions. The sensitivity, specificity and F1 score was very similar among diseases/pathogens and had lower values for GPT

o1 and o1 mini. In the predictions of affected countries, the balanced accuracy differed among countries and LLMs with a median of 75% (IQR 50% - 99.8%). GPT-4o mini, o1 and o3 mini had the lowest balanced accuracies among LLMs and countries with the lowest prevalence in the labelled dataset had the lowest balanced accuracies among countries.

4. Discussion

In this study, we have presented an open-source solution comprising two prototypes for LLM prediction, evaluation and benchmarking of public health use cases. We have shown their applicability and effectiveness in assessing LLM performance across categorical and numerical variables by testing these prototypes on four different public health use cases.

The open-source solution was designed to address the lack of public-health-specific standards for LLM benchmarking and the lack of an automated and near-real-time open-source solution to facilitate this task in the context of having recurrently new LLMs being released. The LLM prediction prototype requires standardised inputs and provides structured outputs ensuring consistency in the evaluations and includes different open-source and commercial LLMs for broad comparison. The Shiny app prototype enables automated evaluation of LLM predictions providing an accessible and transparent process for public health professionals.

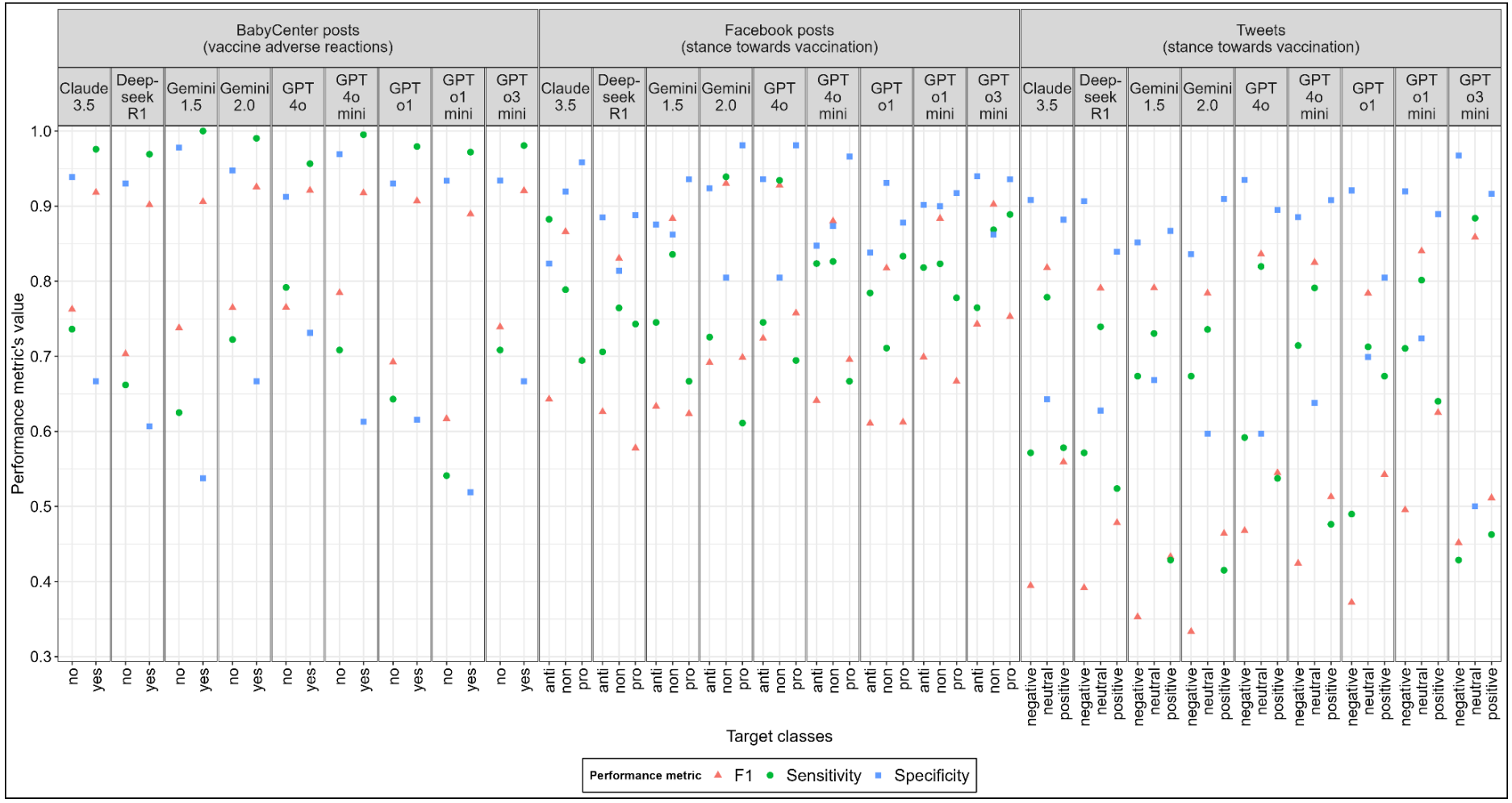


Figure 4. F1-score, sensitivity and specificity per experiment, LLM and class for BabyCenter post, Facebook posts and tweets.

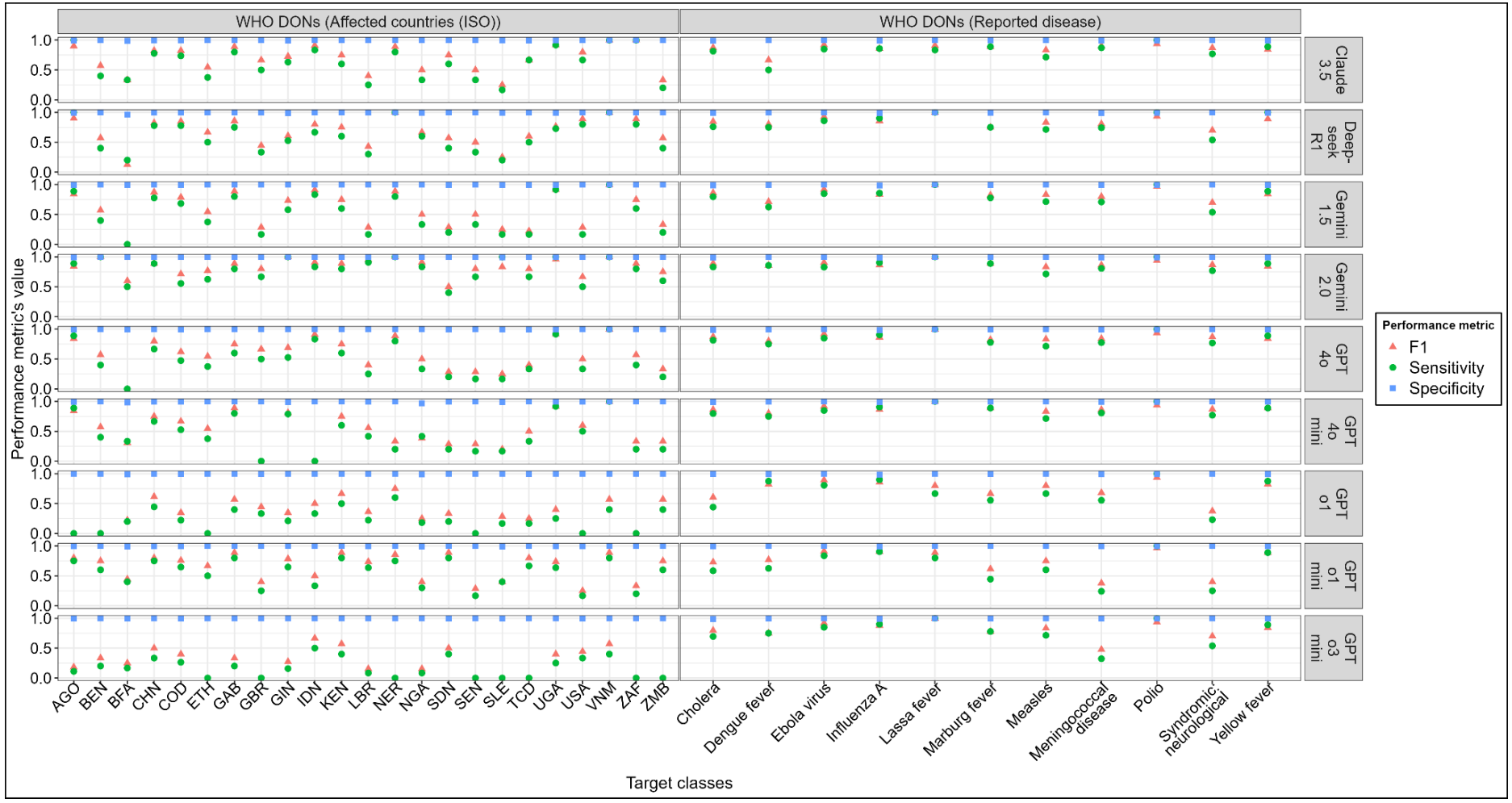


Figure 5. F1-score, sensitivity and specificity per experiment, LLM and class for WHO DONs.

Our analysis has shown variability in the LLM predictions depending on the task, dataset and LLM. For the two use cases on stance towards vaccination from social media platforms, we observed a higher proportion of negative stance posts in Facebook and was not the minority class as in tweets. This discrepancy was likely due to the dataset used from both platforms, since half of the Facebook posts originated from pages considered to be contrary to vaccination. Moreover, the lower accuracy for tweets can be explained by the random selection of tweets, not originating from pages considered favourable or contrary to vaccination, which may be inherently less objective, even for human annotators as shown in a similar study [6].

For the WHO DONs, the accuracy was overall lower than for other use cases, except for reported disease after manual harmonisation of the terminology. Some of the WHO DONs had epidemiological features for more than one country which led to some LLMs to select the firstly reported or considered more relevant and, for other LLMs, to include the epidemiological information also for the countries that were not the main focus of the WHO DON. These underlies, on the one hand, the need for integrating public health ontologies and structured terms (e.g., International Classification of Diseases -ICD-) to enable semantic interoperability for the reported diseases and a clearer and more explicit prompt specifying the most relevant information which is expected to be extracted. Despite providing such structured terminology for the affected countries, there were still inconsistencies which could be handled by explicitly pointing out to a reference with the list of countries and corresponding two-letter ISO code. These discrepancies could have also taken place by human annotators if such references and explicit instructions were not provided either.

In relation to the last public health use case in detecting vaccine adverse reactions from BabyCenter forum posts, the accuracy was higher compared to the stance towards vaccination. This can be explained by the inherent objectivity in detecting this information and by having two clearly defined classes (i.e., yes and no), whereas stance classification involves a degree of interpretation and agreement has not been fully reached even with human annotators [6]. Despite this may suggest that LLMs may have higher performance in tasks that require objective and structured information extraction, tasks requiring subjective interpretation or classification can benefit also from the use of LLM if prompt engineering and understanding of the dataset is performed.

Our findings showed that no LLM consistently outperformed others across all tasks and use cases, emphasizing the need for selecting task- and use-case specific models. Even for the same task (e.g., stance towards vaccination from social media posts), having different dataset composition impacted the performance of the same LLMs. Furthermore, smaller models outperformed larger counterparts in various cases. For example, this was the case for GPT o1 mini in predicting stance towards vaccination from Facebook posts or GPT-4o mini in predicting affected countries and reported cases or deaths in WHO DONs. This suggests that the complexity of the LLM alone does not determine effectiveness and that other aspects should also be considered, such as computational efficiency and costs, particularly for resource-constrained settings.

Furthermore, our results reinforce that non-domain-specific LLM benchmarks may not be directly applicable to the use of LLMs in public health use cases. While the models included in this study are expected to have a high performance overall, their performance was not as high in domain-specific tasks or specific use cases such as the

extraction of epidemiological data. Likewise, when comparing the performance of LLMs from the same company (e.g., OpenAI models), the latest models show higher performance for the general tasks for which these are being evaluated [19]. However, we have seen that, for example, GPT o3 mini does not always have the highest performance values when compared with the other GPTs included in this study.

While this study presented a structured approach for LLMs prediction, evaluation and benchmarking in public health, it has some limitations. The LLM prediction prototype requires deployment which may need the involvement of additional technical expertise; however, to mitigate this, the Shiny app allows other input files with LLM predictions obtained using other solutions. Furthermore, while the Shiny app was only tested in Windows, R Shiny is expected to function similarly across other operating systems. Although not all available LLMs were included in our benchmark, we ensured to have a representative mix of open-source and commercial models with reported high performance. We have not conducted a comprehensive usability assessment of the open-source solution, however, its development was guided by simplicity, user-friendliness and accessibility principles from the perspective of public health professionals.

This study contributes to the effective integration and use of LLMs in public health by providing a structured, user-friendly and scalable solution for LLM prediction, evaluation and benchmarking. Our results underline the relevance of standardised and task-specific evaluation methods and model selection, and the use of clear and structured prompts to improve LLM performance in domain-specific use cases. Future studies should evaluate its real-world usability and adoption in public health, and considering the fast-paced evolution of LLMs and other use cases requiring images,

expand the solution incorporating additional models to ensure up-to-date assessments and testing its usability for LLM predictions based on images.

Funding

This work was supported by Fondation Botnar and the European Union's Horizon H2020 grant VEO (874735). The funders had no role in the design or execution of this study and will have no role in the analyses, interpretation of the data, or decision to submit results.

Conflicts of interest

The authors declare no conflicts of interest.

Authors contributions

LE was responsible for conceptualization, data curation, formal analysis, methodology, development of the Shiny app prototype for LLM evaluation and benchmarking, visualization, and writing original draft manuscript. MS was responsible for conceptualization, funding acquisition, methodology and supervision. DK was responsible for developing the LLM prediction prototype. SC was responsible for the public health use case on extracting relevant features from WHO DONs. KK and YM were responsible for the public health use cases on stance towards vaccination in Facebook posts and detecting vaccine adverse reactions from BabyCenter US forum. All authors contributed to the editing of the final manuscript, and have seen and approved the manuscript.

References

1. Raiaan MAK, Mukta MdSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, et al. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access. 2024;12: 26839–26874. doi:10.1109/ACCESS.2024.3365742

2. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv; 2023. doi:10.48550/arXiv.1706.03762

3. Harris J, Laurence T, Loman L, Grayson F, Nonnenmacher T, Long H, et al. Evaluating Large Language Models for Public Health Classification and Extraction Tasks. arXiv; 2025. doi:10.48550/arXiv.2405.14766

4. Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, Papadopoulos N, Odavic D, Risteski P, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform. 2025;32: e101139. doi:10.1136/bmjhci-2024-101139

5. Matarazzo A, Torlone R. A Survey on Large Language Models with some Insights on their Capabilities and Limitations. arXiv; 2025. doi:10.48550/arXiv.2501.04040

6. Espinosa L, Salathé M. Use of large language models as a scalable approach to understanding public health discourse. Al-Jumeily Obe D, editor. PLOS Digit Health. 2024;3: e0000631. doi:10.1371/journal.pdig.0000631

7. Mets M, Karjus A, Ibrus I, Schich M. Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media. Grabar N, editor. PLoS ONE. 2024;19: e0302380. doi:10.1371/journal.pone.0302380

8. Guo Y, Ovadje A, Al-Garadi MA, Sarker A. Evaluating large language models for health-related text classification tasks with public social media data. Journal of the American Medical Informatics Association. 2024;31: 2181–2189. doi:10.1093/jamia/ocae210

9. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A Survey of Large Language Models. arXiv; 2023. Available: http://arxiv.org/abs/2303.18223

10. Betti L, De Francisci Morales G, Gauvin L, Kalimeri K, Mejova Y, Paolotti D, et al. Detecting adherence to the recommended childhood vaccination schedule from user-generated content in a US parenting forum. Kouyos RD, editor. PLoS Comput Biol. 2021;17: e1008919. doi:10.1371/journal.pcbi.1008919

11. Beiró MG, D'Ignazi J, Perez Bustos V, Prado MF, Kalimeri K. Moral Narratives Around the Vaccination Debate on Facebook. Proceedings of the ACM Web Conference 2023. Austin TX USA: ACM; 2023. pp. 4134–4141. doi:10.1145/3543507.3583865

12. Reddy S. Evaluating large language models for use in healthcare: A framework for translational value assessment. Informatics in Medicine Unlocked. 2023;41: 101304. doi:10.1016/j.imu.2023.101304

13. Evidently AI. 20 LLM evaluation benchmarks and how they work. 20 Feb 2025 [cited 12 Mar 2025]. Available: https://www.evidentlyai.com/llm-guide/llm-benchmarks

14. BabyCenter. About BabyCenter, L.L.C. [cited 12 Mar 2025]. Available: https://www.babycenter.com/about-babycenter/company-information/about?_gl=1*8ongr9*_up*MQ..*_ga*NDY2NzczNzMxLjE3Mzk2OTExODM.*_ga_BP5N04S2Y3*MTczOTY5MTE4Mi4xLjAuMTczOTY5MTE4Mi4wLjAuMA

15. World Health Organization. Disease Outbreak News (DONs). [cited 12 Mar 2025]. Available: https://www.who.int/emergencies/disease-outbreak-news

16. Abbood A, Ullrich A, Busche R, Ghozzi S. EventEpi—A natural language processing framework for event-based surveillance. Althouse BM, editor. PLoS Comput Biol. 2020;16: e1008277. doi:10.1371/journal.pcbi.1008277

17. Auss Abbood, Ullrich A, Busche R, Ghozzi S. Incident data base for event based surveillance. figshare; 2020. p. 47017 Bytes. doi:10.6084/M9.FIGSHARE.12575978

18. World Health Organization. Ebola virus disease, West Africa – update. In: Disease outbreak news [Internet]. [cited 12 Mar 2025]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2014_04_22_ebola-en/

19. Artificial Analysis. OpenAI: Models Intelligence, Performance & Price. [cited 12 Mar 2025]. Available: https://artificialanalysis.ai/providers/openai#intelligence