JMIR Publications
The Leading eHealth Publisher

Use of large language models and synthetic social media to optimize and validate assessment of epidemiological characteristics in social media posts about outbreaks: Infodemiology Study

Michael S. Deiner, Russell Y Deiner, Cherie Fathy, Natalie A Deiner, Vagelis Hristidis, Stephen D. McLeod, Thuy Doan, Gerami D. Seitzman, Thomas M. Lietman, Travis C. Porco

Submitted to: Journal of Medical Internet Research
on: August 09, 2024

Table of Contents

Use of large language models and synthetic social media to optimize and validate assessment of epidemiological characteristics in social media posts about outbreaks: Infodemiology Study

Michael S. Deiner[1] PhD; Russell Y Deiner[2]; Cherie Fathy[3] MD; Natalie A Deiner[4]; Vagelis Hristidis[5] PhD; Stephen D. McLeod[6] MD; Thuy Doan[1] MD, PhD; Gerami D. Seitzman[1] MD; Thomas M. Lietman[7] MD; Travis C. Porco[7] MPH, PhD

[1]Francis I. Proctor Foundation for Research in Ophthalmology and Department of Ophthalmology, University of California, San Francisco San Francisco US
[2]Abraham Lincoln High School, San Francisco, CA San Francisco US
[3]Food and Drug Administration, Center for Biologics Evaluation and Research, Document Control Center, Document Control Center, 10903 New Hampshire Ave., Bldg. 71, Rm. G112, Silver Spring US
[4]University of California, Santa Barbara, Santa Barbara, California College of Letters and Sciences Santa Barbara US
[5]University of California, Riverside Riverside US
[6]Francis I. Proctor Foundation for Research in Ophthalmology (University of California, San Francisco,, Department of Ophthalmology (University of California, San Francisco),, and American Academy of Ophthalmology San Francisco US
[7]Francis I. Proctor Foundation for Research in Ophthalmology, Department of Ophthalmology, and Department of Epidemiology and Biostatistics University of California, San Francisco San Francisco US

Corresponding Author:
Travis C. Porco MPH, PhD
Francis I. Proctor Foundation for Research in Ophthalmology, Department of Ophthalmology, and Department of Epidemiology and Biostatistics
University of California, San Francisco
University of California, San Francisco, UCSF Box 0412
San Francisco
US

Abstract

Background: Use of online search and social media can help identify epidemics, potentially earlier than clinical methods or even potentially identifying otherwise unreported outbreaks. Monitoring for eye-related epidemics can facilitate early public health intervention to reduce transmission and ocular comorbidities associated with outbreaks. However, use of social media for such monitoring is hindered by costs of laborious manual content review. To address this limitation, we have shown utility of large language models (LLMs) to assess probabilities of an outbreak from social media posts. Knowing the probability alone though may not be as informative to public health actions as also knowing more epidemiological characteristics about them, for example knowing the outbreak type, size or which ones the most severe.

Objective: We assessed if and how well LLMs can classify essential epidemiological features from individual social media posts beyond outbreak probability, including outbreak type, size, severity, etiology and location as well as other health conditions. We employed a validation framework comprising synthetic, Twitter/X and forum posts, comparing an LLMs classification to other independent LLM models and to human experts.

Methods: To develop effective prompts and test the capability of multiple LLMs, synthetic social media posts were generated. These synthetic posts were embedded with specific pre-classified epidemiological features to simulate various outbreak and control scenarios. To gauge the LLM's practical utility in real-world epidemiological surveillance, top performing LLM inter-model comparisons were made using Twitter/X and forum posts. Finally, human graders also classified a subset of posts and their classifications were compared to a leading LLM for validation. Comparisons entailed correlation, or sensitivity and specificity statistics.

Results: Seven LLMs assessed for effectively classifying epidemiological data from diverse social media posts. Notably, GPT-4 and Mixtral 8x22b exhibited high performance in predicting outbreak characteristics like probability, size, and type. Lower performing LLMs were successful for some classifications but not others. Despite strong correlations in comparative validations and known values, discrepancies were noted in a few categories of human assessments. However, overall, the models demonstrated a reliable capacity for nuanced epidemiological analysis across various data sources.

Conclusions: This investigation into the potential of LLMs for public health infoveillance suggests effectiveness in classifying key epidemiological characteristics from social media content about conjunctivitis outbreaks. Future studies may suggest that while LLMs have potential to support public health monitoring, their optimal role may be to act as a first line of documentation, assessment and classification of potential outbreaks, alerting public health organizations for follow-up of LLM-detected and classified small early outbreaks with a focus on the most severe ones. Clinical Trial: n/a

(JMIR Preprints 09/08/2024:65226)
DOI: https://doi.org/10.2196/preprints.65226

Preprint Settings

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

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

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

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Use of large language models and synthetic social media to optimize and validate assessment of epidemiological characteristics in social media posts about outbreaks: Infodemiology Study

Authors

Michael S. Deiner[1,2], Ph.D., Russell Y. Deiner[3], Cherie Fathy[4], M.D., Natalie A. Deiner[5], Vagelis Hristidis[6], Ph.D., Stephen D. McLeod[1,2,7], M.D., Thuy Doan[1,2], M.D., Ph.D., Gerami D. Seitzman[1,2], M.D,. Thomas M. Lietman[1,2,8], M.D., Travis C. Porco[1,2,8], Ph.D., M.P.H.

Affiliations:

1. Department of Ophthalmology, University of California, San Francisco, San Francisco, California
2. Francis I. Proctor Foundation for Research in Ophthalmology, University of California, San Francisco, California
3. Lincoln High School, San Francisco, CA
4. Food and Drug Administration, Center for Biologics Evaluation and Research, Document Control Center, Silver Spring, MD
5. University of California, Santa Barbara, Santa Barbara, California
6. University of California, Riverside, Riverside, California
7. American Academy of Ophthalmology, San Francisco, California
8. Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California

Corresponding Author and Address for Reprints:

Travis C. Porco, F. I. Proctor Foundation, Department of Epidemiology and Biostatistics and Department of Ophthalmology, University of California, San Francisco, UCSF Box 0412, San Francisco, California 94143-0412 USA; Email travis.porco@ucsf.edu; Phone (415) 476-4101; FAX (415) 476-0527.

Financial Support:

This work was supported in part by grant 1R01EY024608-01A1 (Lietman, PI) from the National Institutes of Health National Eye Institute (NIH-NEI), grant EY002162 (Core Grant for Vision Research, Ullian, PI) from the NIH-NEI, and an Unrestricted Grant from Research to Prevent Blindness (McLeod, PI). The sponsor or funding organization had no role in the design or conduct of this research.

Conflict of Interest:

No conflicting relationship exists for any author.

Abstract

Background: Use of online search and social media can help identify epidemics, potentially earlier than clinical methods or even potentially identifying otherwise unreported outbreaks. Monitoring for eye-related epidemics can facilitate early public health intervention to reduce transmission and ocular comorbidities associated with outbreaks. However, use of social media for such monitoring is hindered by costs of laborious manual content review. To address this limitation, we have shown utility of large language models (LLMs) to assess probabilities of an outbreak from social media posts. Knowing the probability alone though may not be as informative to public health actions as also knowing more epidemiological characteristics about them, for example knowing the outbreak type, size or which ones the most severe.

Objective:
We assessed if and how well LLMs can classify essential epidemiological features from individual social media posts beyond outbreak probability, including outbreak type, size, severity, etiology and location as well as other health conditions. We employed a validation framework comprising synthetic, Twitter/X and forum posts, comparing an LLMs classification to other independent LLM models and to human experts.

Methods:
To develop effective prompts and test the capability of multiple LLMs, synthetic social media posts were generated. These synthetic posts were embedded with specific pre-classified epidemiological features to simulate various outbreak and control scenarios. To gauge the LLM's practical utility in real-world epidemiological surveillance, top performing LLM inter-model comparisons were made using Twitter/X and forum posts. Finally, human graders also classified a subset of posts and their classifications were compared to a leading LLM for validation. Comparisons entailed correlation, or sensitivity and specificity statistics.

Results: Seven LLMs assessed for effectively classifying epidemiological data from diverse social media posts. Notably, GPT-4 and Mixtral 8x22b exhibited high performance in predicting outbreak characteristics like probability, size, and type. Lower performing LLMs were successful for some classifications but not others. Despite strong correlations in comparative validations and known values, discrepancies were noted in a few categories of human assessments. However, overall, the models demonstrated a reliable capacity for nuanced epidemiological analysis across various data sources.

Conclusions:
This investigation into the potential of LLMs for public health infoveillance suggests effectiveness in classifying key epidemiological characteristics from social media content about conjunctivitis outbreaks. Future studies may suggest that while LLMs have potential to support public health monitoring, their optimal role may be to act as a first line of documentation, assessment and classification of potential outbreaks, alerting public health organizations for follow-up of LLM-detected and classified small early outbreaks with a focus on the most severe ones.

Keywords: Conjunctivitis, large language models, epidemiology, social media, Twitter/X, Forums

Introduction

The use of online search and social media in health research to detect epidemics or outbreaks could potentially allow identification of epidemics more quickly than clinical surveillance or detection of epidemics that might have otherwise gone unreported [1-51]. Conducting such analyses using social media posts requires natural language understanding of social media content, and has been limited by the costs of human raters or alternatively by the insufficiency of computational natural language processing [41, 44, 45, 49, 51-54] . More recently large language models have been explored as an alternative tool for this purpose[50, 51, 54-56].

Monitoring ocular health to detect eye-related epidemics can reduce risks to eye health and lessen the overall societal impact[57, 58]. Surveillance can also act as an early indicator of systemic diseases, including COVID-19[59-62] or vision threatening effects of contaminated eye care products[63]. Social media content has potential to improve monitoring, but currently requires costly manual review of post content. Recently, we studied the potential of a large language model to determine if outbreaks of conjunctivitis might be detected based on the content of Twitter posts about the condition[50]. We investigated whether GPT-3.5 and GPT-4, can provide probabilistic assessments of whether or not social media posts (Twitter, now X) about conjunctivitis could indicate an outbreak. Mean probabilistic assessments from GPT-4 correlated well with human raters, with r=0.60 (95% CI: 0.47–0.70). The mean of these elicited percentages from tweets about conjunctivitis correlated with Tweet volume and with the occurrence of some known epidemics. We did not assess the ability of LLMs to assess the probability of epidemics based on content sourced from other forms of social media such as online forum discussions or blogs. In addition, we observed that quantitative analysis of clusters of posts for time series of counts over time can be insufficient as it requires enough posts for detecting differences in counts of posts over time. It is possible that in some instances, there may never be enough cases for a statistical analysis such as this.

Our prior study led us to conjecture that even one individual post about an epidemic may allow for public health identification of probable outbreaks, including very small or early outbreaks not detectable yet using analyses of timeseries counts per day, clinical data, or other means. Furthermore, we have hypothesized that there may be potential of LLMs to extract reliable information not just about the probability of epidemics, but also about the characteristics (such as severity or cause) of suspected conjunctivitis epidemics. Theoretically, this might allow us to identify epidemics, including severe ones, from single posts and without the need for a significant volume of searches or posts for time series count changes. This method could potentially yield an advantage in identifying epidemics earlier than clinical cases emerge or before data are available for analysis, or to potentially identify posts that target small but severe epidemics, earlier than observable from time series counts of clinical cases, online searches, or social media posts.

To use an LLM model to identify characteristics of outbreaks with a wide range of characteristics (e.g. one household vs. a state, mild cases vs. severe cases, infectious vs. allergic or environmentally caused, systemic vs. only ocular, bacterial vs. viral), we must first have a set of posts that represent this wide range of characteristics. The potentially more important outbreaks, such as those causing AHC, may appear rarely in social media due to their being very rare in general. As a solution to using an LLM model to properly characterize such unique and rare posts, one solution used in the field is to use a model of synthetic posts that are pre-classified for representing a full range of characteristics – including rare but important events[4, [64-67]. Preclassified values (for example a post with a known conjunctivitis health impact severity level of “mild’ vs. severe”) also provide an efficient approach to training and testing a model for sensitivity and specificity, as it bypasses the traditionally very time consuming and costly process of humans assigning classification scores to posts after obtaining the posts. Numerous other advantages for creating and using synthetic social media posts include allowing large quantities of data at very low cost which helps powering for statistical analyses, full control over variables allowing for more certainty in experiments, content modification as needed during iterative model development and refinement, and full privacy, [66, 67]. For all of these reasons, we created and used synthetic posts in this current study, allowing us to develop our model before validating findings with human subject matter experts.

LLMs are becoming highly used in all fields, including health[54-56, 68]. An emerging theme of importance is repeated, domain-specific validation of LLM outputs along with the theme of how impractical that may be for humans. Numerous solutions have been proposed, including use of multiple LLMs either as ensemble or chained LLM models to leverage the strengths and weaknesses, as well as use of one LLM to assess the validity of the output of another LLM [51, 54, 55, 69-74]. Such an approach allows a highly scalable automated assessment and correction without human validation limitations. In our study we took this such approach as a general means to validate the output of one LLMs classification of epidemic characteristics by determining the sensitivity and specificity of that LLM’s output compared to that of other leading independent LLM platforms. In this manner, there is potential to complement human validation efforts in an affordable and statistically centered approach.

The aims of this infoveillance study were therefore to ask: in addition to the probability of an outbreak, **can an LLM classify key significant epidemiological outbreak characteristics (outbreak size and type, severity and etiology of cases, other health conditions mentioned) from individual social media posts originating from a number of social media platforms?** We test the general hypothesis that the correlation, sensitivity, and specificity results would be better for some LLMs than others, and be better for some of these simpler features being assessed (e.g. probability or type of outbreak) more than harder to discern features (e.g. severity of etiology of cases). Using a diverse set of 7 LLMs to conduct several tiers of validations, we assess if LLMs can provide probabilistic assessments of the chance each post corresponds to an outbreak, and classify other characteristics. We have implemented a validation framework using synthetic data with pre-defined characteristics, allowing us to assess LLM performance against a gold standard. We also use comparative validation between LLMs, and human validation by domain experts to further strengthen reliability of our findings for posts from multiple social media platforms. If successful, leveraging these classifications could enable us to study the nature of outbreaks including those that may never be present, detected or well-characterized in clinical datasets, and it has potential for identifying higher-risk severe epidemics which in turn could aid public health authorities in taking timely actions specifically targeted to address severe outbreaks earlier.

Methods

Methods Overview

LLMs were used to extract and analyze public health information from social media posts, focusing on classifying outbreak characteristics (probability, size, type, severity, other health conditions, etiology). First, for efficient model development and to rigorously evaluate LLM capabilities, we developed a framework for generating synthetic social media posts with predefined epidemiological characteristics, using templates, controlled vocabularies, and random sampling. This allowed generation of diverse posts simulating various pre-classified outbreak scenarios, health conditions, and demographics, spanning a wide range of conditions, including rare ones for LLM assessment. From an initial pool of multiple different LLMs, we identified top performers based on their performance on these synthetic datasets after fine-tuning prompts for specific tasks. We then assessed LLM performance using a combination of correlation analysis, sensitivity and specificity calculations, and human validation by domain experts.

Data:

Twitter/X and Forum Posts: We collected social media posts from the Twitter microblogging service (subsequently rebranded as X) as well as from online Forums, and Blogs using the Brandwatch interface. To obtain posts about conjunctivitis, we used a Boolean query containing words in multiple languages representing conjunctivitis (eg, “conjuctivitis,” “conjuntivitis,” “conjuntivite,” and “pink eye”). We tailored the query to enrich for posts about outbreaks of conjunctivitis (of any size) where cases and symptoms were described and to exclude irrelevant or unrelated content, such as that related to animals, artistic or literary references, obscenities, celebrities or public figures having pink eye. Multimedia Appendix 1 provides the Boolean query details. The data cutoff window began on October 16, 2018 (October 16, 2020, for Twitter/X posts), and ended on October 16, 2023. The data were exported on October 23, 2023.

Generation of Synthetic Posts with predefined epidemiologic parameter values: We designed a modular system to allow generation of synthetic posts in which the classifications for all parameters of interest were pre-assigned (and did not need to be extracted manually) for planned comparisons to the values of these parameters extracted from LLMs. This allowed for rapid assessment of any given LLM platform’s ability to extract information of interest, comparing values elicited by the LLMs to values pre-assigned to each post for all parameters of interest in the posts. The synthetic posts were constructed by combining components from predefined categories, specifically: outbreak probability indicators, outbreak settings, disease severity descriptors, causative organisms, and associated symptoms. We selected these at random, and concatenated components from each category based on specified probabilities to generate plausible posts. Disease severity was defined as the maximum severity among the components within a post. Non-epidemic phrases and non-infectious forms, such as environmental causes, were incorporated probabilistically to ensure a diverse dataset with controls. More details of the post creation algorithm are provided in the Supplement. We generated 1152 synthetic posts using the algorithm, which was implemented in R v. 4.3 for MacIntosh (R Foundation for Statistical Computing, Vienna, Austria).

Iterative prompt development using a training set of synthetic posts

We used a training set of synthetic posts (separate from the set used in testing and analysis later) to improve our prompts, optimizing the success rate of LLM outputs for identifying the correct pre-defined values for parameters of interest in each synthetic post. We continued this iterative process until achieving consistent and adequate results (based on GPT-4-0314). Prompt optimization led us to the following prompt used in our subsequent testing studies described below.

Optimized LLM prompt

For our testing studies below, we used the following optimized user prompt for each LLM (with slight variations per LLM to accommodate any LLM-specific technical formatting requirements). For some API's that allowed them, we also included a system prompt[75] to convey a public health analyst role to the LLM. Such a system prompt may reduce the chance that the LLM will respond with helpful advice (such as telling the user to seek medical attention), respond with disclaimers, or refuse to answer the query.

> *SYSTEM PROMPT*: "You are a health analyst for a Department of Public Health. You are summarizing what individuals are saying in social media posts, helping to distinguish reports of rumors, discussions of movies, and so on from reports of actual cases of disease."
>
> *USER PROMPT: "For every snippet provide the following: how certain are you that this snippet is about a multiperson outbreak of pink eye occurring at the time the snippet was posted? A single case with no other evidence of spread or other infected people should correspond to a somewhat low probability. Suggestions of numerous people affected at one or multiple location or groups impacted (things like \"everyone at...\" or \"the entire district of...\" or \"...something is cancelled\" or \"my work is empty\" or \"something is closed today due to\" or \"school closed\" or \"daycare closed\") should have a higher probability, and the more people affected, the higher the probability should be. Any obvious conjunctivitis epidemic with more than one person should receive a high score. If it's about pink eye in non-human animals, then the probability is 0%. If it seems like it is not about a real-life occurrence (for example if it is about dreams, or about fake news, or about rumor, or about a fictional movie or tv shows, or literary fiction, etc.) then assign a probability of zero 0%. Assume all symptoms mentioned are ones that can occur in real life though, even things like \"can't taste\" or \"can't smell\" or \"lymph nodes\" are real symptoms. Assume none of these snippets are about fictional characters. Do NOT guess at location, just use the information provided for location. In addition, provide an estimate of the severity on health (mild, moderate, strong, severe, where \"mild\" is not significant, \"moderate\" has some impact on health, \"strong\" has serious health impact and \"severe\" is life-threatening). Also provide the type of outbreak: \"allergic\", \"infectious\", \"environmental\" (swimming pools, pollution, toxic spills, smoke, wildfires), or \"infectious-AHC\" (AHC is very severe and typically includes extremely red, bloody or bleeding or blistering eyes, vision loss and other severe symptoms). If there is content about drugs or drug usage (i.e. smoking weed, pot, marijuana, a joint, a bowl, a bong), getting high, sparking up, other slang terms like \"Mary Jane\", \"bud\", \"ganga\", \"reefer\", \"chronic\", \"herb\", \"spliff\", \"roach\", \"a j\") then consider the type of outbreak as \"environmental - drugs\". If you determine that the cause is infectious, please also tell us whether it is viral or bacterial. Also, tell us the health condition or disease being discussed (e.g. conjunctivitis, flu, broken leg, etc.) In your responses, label each answer. When you tell us the snippet ID, write \"snippet ID:\" followed by the snippet ID. When you tell us the location, write \"location:\" followed by the location, and so forth. Respond in the form of snippet ID, location, a numeric percentage of the probability between 0% and 100%, and then provide a numeric value for the number of people affected based in part on the estimated population of the location and how many of*

them might be affected (provide a NUMBER of people, do NOT provide a categorical term like low or high, instead provide an actual estimated number), type of outbreak and cause if infectious conjunctivitis, a list of the symptoms, the severity on health, the health condition or disease being discussed, and provide a brief explanation of your answer. Given this single snippet: \"{}\"",

Selection and interface with LLMs used in our study

In general, we chose LLM platforms that are more commonly used and published in the literature as well as on availability and accessibility. We used the following LLMs in the current study: GPT-3.5 (specifically gpt-3.5-turbo-0613), GPT-4 (specifically GPT-4-0314), GPT-4o (gpt-4o-2024-05-13) (OpenAI, San Francisco); Sonnet 3 (Anthrop\c, San Francisco); Mixtral 8x22B (Mistral AI, Paris, France); and LlaMa 3 70B (Meta AI, New York, New York). To send prompts and receive responses, we used vendor-provided APIs for OpenAI and Anthropic. We conducted inference for Mixtral 8x22B and LlaMa 3 on the Octo Labs platform (Octo Labs, Seattle, WA). A temperature of 0.0 was used for each inference.

Comparison of different LLMs' abilities to extract epidemiological characteristics from posts using a testing set of synthetic posts

After prompt optimization, we then generated a test set of 1,152 synthetic posts for conducting subsequent studies to submit the prompt above and collect the outputs for the LLMs described above. For the most part, LLMs responded with semi-structured outputs, for example: "***Snippet ID:** K1428, **Location:** household implied, **Probability:** 85%, **Number of People Affected:** 3-5 (assuming a typical number of children in a household), **Type of Outbreak:** Infectious (likely viral), **Symptoms:** Pink eye, **Severity on Health:** Moderate, **Health Condition or Disease:** Conjunctivitis (pink eye), **Explanation:** The snippet describes a household situation where multiple children are sick with pink eye. The phrase "filled with sick kids" suggests that more than one child is affected, indicating a multiperson outbreak within the household. Although the exact number of sick children is not specified, the context implies a significant number. The urgency implied by "help!" indicates concern, but there are no life-threatening symptoms mentioned, so the severity is considered moderate. The cause is likely infectious, probably viral, given the spread among children in a shared living environment.*"

After collecting outputs, using an LLM chain approach, we then used the following LLM prompt to extract these variables (among others) into structured format from the raw single-paragraph LLM replies: (1) location, (2) probability, (3) number affected, (4) the type of outbreak, (5) severity, and (6) health conditions. To assess the information provided by the LLMs regarding the health conditions, we used sciSpacy (Allen AI) for named entity recognition, based on `en_core_sci_lg` and the Unified Medical Language System (UMLS) vocabulary. This entailed some modifications (for example the word 'croup' needed to be converted to `laryngotracheobronchitis' first to avoid unwanted conversion of "croup" into "group"), and we also used the Python `TextBlob` package to further correct occasional misspellings.

DATA EXTRACTION PROMPT: "secprompt" : "Please use the information provided to fill out the XML form fields indicated: <location> </location> <probability> </probability> <number_affected> </number_affected> <type_of_outbreak> </type_of_outbreak>

<cause></cause> <severity></severity> <symptoms> </symptoms> <healthcondition> </healthcondition> <explanation> </explanation>. Include the answer between the matching tags; for instance, a probability of X% would be indicated as <probability>X%</probability>. Please do not change the format of probabilities given as percentages; leave percentages in their original form. Express number_affected in the form of a Hindu-Arabic numeral; do not use words. Given this information: \"{}\""

Analytic Methods to Compare LLM values to Synthetic Posts Values

Gold Standard Source: The data for this study were obtained from synthetic social media posts, referred to as "Synthetic Posts," which were generated to simulate real-world scenarios of outbreak reporting. The gold standard data consisted of epidemic probability score, severity categorizations, and etiological classifications pre-assigned to the components of these synthetic posts at the time of their creation. Of the posts assessed by each LLM most of the posts were about conjunctivitis in one form or another and a small number were exclusively about other conditions.

Validation Measures and Evaluation Criteria for Epidemiological Characterizations

Each LLM's assessment of the epidemiological characteristics were then validated by comparing to the predefined gold standard values of the synthetic posts as described below. To compare these with the gold standards, we conducted statistical analysis as follows.

1) Outbreak Probability: To assess the probability of an outbreak as predicted by the LLM model, we calculated the Pearson correlation coefficient between the gold standard probability scores and the model's assigned probabilities of an outbreak. Synthetic posts labeled as not about conjunctivitis at all were excluded to ensure consistency in the analysis.

It is important to note that for all additional measures below, if the outbreak probability pre-assigned to the synthetic post was less than 20%, the LLMs often refused to provide results for the prompts other than the probability of an outbreak (the LLMs often mentioned seeing no reason to continue assessing outbreak characteristics if they did not think the post was about a true outbreak). Therefore, for assessment of all variables below (i.e. except for Outbreak (OB) Probability) we only used the set of posts where the probability score pre-assigned to the synthetic was 20% or above. In addition, results from posts that were entirely not about conjunctivitis were only included in our assessments to identify other health conditions mentioned in a post.

2) Outbreak Size (# of Cases): The relationship between the gold standard outbreak size category and the LLM's estimated numeric outbreak size was evaluated using Spearman's rank correlation coefficient. 3) Outbreak Severity of Cases: The fraction of records where the gold standard severity was within one level of the model's assigned ordinal severity was determined. 4) Type of Outbreak: For each of the 4 types of outbreaks (Infectious, Allergic, Environmental and Acute Hemorrhagic Conjunctivitis (AHC), the sensitivity and specificity of the LLM model in identifying each outbreak type was compared to the gold standard. 5) Health Conditions: The sensitivity and specificity of the LLM model in identifying the following health conditions in each post were compared to the gold standard: for conjunctivitis, COVID-19,

influenza, gastrointestinal illness, croup, lice infestation, and broken leg. 6) Community Location: To compare the LLM location outputs to the gold-standard pre-defined synthetic post location values, we used mean cosine similarity[76] with Bidirectional Encoder Representations from Transformers (BERT, Google, Mountain View). We used the BERT model and tokenizer[77] (accessed from the Hugging Face `transformers` Python library, Hugging Face, Brooklyn, New York) to tokenize input texts, generate contextual embeddings, and compute cosine similarity between these embeddings to measure semantic similarity between location text pairs.

Confidence Intervals

To calculate the 95% confidence intervals of all sensitivity, specificity and percent agreement results shown, the exact (Clopper-Pearson) method was used. To calculate the 95% confidence intervals of all correlation coefficients shown, z transformation method was used.

Software and Environment

Statistical analysis was conducted using R version 4.2.0 for Macintosh. The `sciSpacy` package and all API calls were conducted through Python 3 on an Ubuntu Linux cloud server.

Validation of GPT-4 classification of epidemiological characteristics from synthetic posts by cross-comparing to other LLMs classifications of epidemiological characteristics from the same synthetic posts.

After identifying one of the more successful LLMs (GPT-4-0314, referred to as GPT-4 below) to extract classification of epidemiological characteristics from synthetic posts (above) we then sought to validate its ability to classify epidemiological characteristics compared to other LLMs. This would allow us to determine if top performing LLMs, from completely different companies and platforms, had found similar results. In this way, we used GPT-4 as the standard to validate the results obtained with other LLMs. We conducted this using the same set of 1,152 synthetic posts as in the prior section above (comparing LLM results to gold standard from synthetic posts). The methods used and components assessed (Outbreak Probability, Outbreak Size, Outbreak Severity of Cases, Type of Outbreak, Health Conditions) were identical to that described above, with the exception that the standard values in these assessments were the GPT-4 extracted values, rather than the gold-standard pre-defined values of the synthetic posts. Success of the different leading LLMs to classify epidemiological characteristics similarly to GPT-4 would suggest we could also use this approach to help validate the results of LLM classification of epidemiological characteristics of real world posts, comparing the results from GPT-4 to that of other unrelated LLMs. Leading to our next comparative assessment below.

Validation of GPT-4 classification of epidemiological characteristics from real-world Twitter/X and Forum posts by cross-comparing to other LLMs classification of epidemiological characteristics from the same set of real-world posts

We then assessed ability of LLMs to extract classification of epidemiological characteristics from a set of real world Twitter/X posts and from a set of Forum posts, using GPT-4 as a surrogate standard since there was no gold standard for classification of parameters in these posts. The components assessed and methods used were identical to that described above (Outbreak Probability, Outbreak Size, Outbreak Severity of Cases, Type of Outbreak, Health Conditions), with the exception that the standard values in these assessments were the GPT-4 extracted values for the Twitter and Forum posts. Results from other LLMs were compared to these GPT-4 values. Similar results between LLMs from different companies and platforms, could help validate result of GPT-4, but an additional approach would be to validate GPT-4 results using human graders. This approach is described below.

Human Validation: GPT-4 classification of epidemiological characteristics from posts

Having established success with GPT-4 to extract characteristics, validated by other LLMs, from real-world posts, we then conducted validation studies using human graders, classifying epidemiological characteristics of posts in a fashion similar to our prior study [50], but with more participants and more variables to be graded.

Training phase

Two groups of human graders participated in this study: two individuals with no medical training (non-expert graders) and two trained, practicing ophthalmologists specializing in surface diseases (expert ophthalmologists). All human graders underwent a training phase conducted via Zoom with one of our study team members, and using Qualtrics surveys (Qualtrics International Inc). This training phase aimed to familiarize graders with the task, ensure consistency in their assessments, and introduce them to the nuances of social media language. The training phase consisted of two distinct sets, each containing 20 posts (not all posts were necessarily used in the training). During the first training set, for each section (outbreak size, type, severity, etc.) human graders collaboratively reviewed the original prompts provided to the LLMs and then reviewed instructions provided to them on how review each post and assign classifications and/or probability scores using a Qualtrics survey. This was followed by a facilitated group discussion to address discrepancies in classifications and familiarize graders with social media communication styles, including the use of hashtags and sarcasm. For the second training set, graders independently reviewed the posts and assigned classifications and probability scores using a Qualtrics survey. A brief discussion was then conducted to ensure consensus and further refine their classification skills.

Testing Phase

Following the training phase, for each epidemiological characteristic, human graders were again provided the original LLM prompts and the human instructions for each section, and then independently assessed a testing set of posts in a separate Qualtrics survey. This testing set excluded any posts used in the training phase. Graders assigned classifications and probability scores without any knowledge of other graders' or GPT-4's classifications, ensuring an unbiased assessment.

Epidemiological Characteristics Assessed in Qualtrics Test Sets

The **non-expert human graders** classified posts across five characteristics: Outbreak Probability, Outbreak Size, Type of Outbreak, Severity of Outbreak, and Health Conditions Mentioned. The **expert trained surface disease practicing ophthalmologists** classified the same set of testing posts as the non-expert graders, focusing on Type of Outbreak and Severity of Outbreak and one additional characteristic of etiological cause, using the same questions and response options. Details are as follows and instructions from the Qualtrics for each item below are presented in the Appendix “Qualtrics Details” section. <u>Outbreak Probability:</u> For 72 posts (24 Synthetic, 24 Twitter/X, 24 Forum) graders were asked, "How certain are you that this snippet is about a multiperson outbreak of pink eye occurring at the time the snippet was posted?" and provided with response options ranging from 0% to 100% in increments of 10%. Severity of Outbreak: For 90 posts (38 Synthetic, 26 Twitter/X, 26 Forum) graders chose one the following response options: "NOT SPECIFIED", "MILD", "MODERATE", "STRONG", and "SEVERE". Outbreak Size: the same 90 posts, were assessed by asking graders to provide their best numerical estimate of affected individuals based on the post content, with the option to input "N" if the size was not defined. Type of Outbreak: for the same 90 posts, graders selected from the following options: "NOT SPECIFIED", "ALLERGIC", "INFECTIOUS", "ENVIRONMENTAL", and "AHC-INFECTIOUS". Health Conditions Mentioned: For 50 Synthetic posts, graders were asked to list any health conditions discussed in the post, excluding individual symptoms unless considered a known health condition. Etiological cause: For the set 90 posts (38 Synthetic, 26 Twitter/X, 26 Forum) used above, we also used a modified LLM prompt, with the following language specific to etiology that we used to elicit responses from GPT-4. In Qualtrics, the two trained surface disease practicing ophthalmologists were given instructions to assign an infectious or non-infectious cause for each of the 90 posts, choosing from the same 6 etiology categories as provided in the LLM prompt ("NOT INFECTIOUS", "INFECTIOUS: UNSPECIFIED", "BACTERIAL", "VIRAL: UNSPECIFIED", "VIRAL: COVID-19", "VIRAL: AHC"). See the appendix for the modified GPT-4 etiological cause prompt.

Statistical Analysis of Human Classifications

Assessment of the human classification of these components were identical to that described above (Outbreak Probability, Outbreak Size, Severity of Cases, Type of Outbreak, Other Health Conditions) with the exception that the human values were compared to that of GPT-4 and for etiological cause the MD specialist was treated as the gold standard. Sensitivity refers to the model's ability to correctly identify positive cases, while specificity refers to its ability to correctly identify negative cases. The bincmp function was used to calculate sensitivity and specificity for each cause. This function compared the binary assessments from the GPT-4 model against the MDs' assessments. For any result where the denominator was less than 5, we present results simply as a ratio rather than as a fraction. For etiological cause, outputs were calculated for all 3 sources of posts combined, as well as for each individual source (Synthetic, Twitter/X, Forums).

Results

Overall Summary of multiple LLM Responses to Synthetic posts, Twitter/X posts and Forum posts

The overall ability of the LLMs to respond in the format that we desired was high, as all of these 6 LLMs yielded usable responses to all 1152 synthetic posts, all 370 Twitter/X posts, and all 290 forum posts, except that GPT-4o failed to produce a usable result for one of the forum posts. As in our previous study with Twitter/X posts and GPT-4 [50], in this study we observed that the ability of multiple LLMs to summarize the content and characterize epidemiological characteristics of the synthetic posts, Twitter/X posts and Forum posts seemed quite reliable and surprisingly nuanced – with different abilities between LLMs. Table 1 presents a selection of GPT-4's prompt responses for synthetic posts, Twitter/X posts, and Forum posts. To ensure privacy, identifiers in posts have been redacted.

Table 1. Selected examples of GPT-4 prompt responses for synthetic posts, Twitter/X posts and Forum posts. Twitter/X and Forum Posts have been redacted where indicated with square

brackets to de-identify them. Some longer explanation content has been truncated, indicated with “…”, in order to save space. Posts were selected to show a range of epidemiological characteristics for each source of post.

Source and Original Snippet	GPT-4 Characteristics	GPT-4 explanation
Source: Synthetic **Snippet:** 'our whole household attedning Montepuez, Mozambique complinaing about diarrhea with headache and pain in limbs combined with emergency room'	**Probability:** 0% **Size:** 0 **Type:** Not applicable **Severity:** Not applicable **Etiology:** Not applicable	The snippet does not mention pink eye or conjunctivitis, so the probability of a multiperson outbreak of pink eye is 0%. The symptoms mentioned are not related to pink eye. The health condition being discussed is not specified, but the symptoms suggest a gastrointestinal issue. The severity on health is not specified in the snippet.
Source: Synthetic **Snippet:** 'in my dream in Epecuen, Argentina complinaing about pinkneye with hayfever and itchy and swollen'	**Probability:** 0% **Size:** 0 **Type:** N/A **Severity:** N/A **Etiology:** N/A	The snippet mentions a dream, which indicates that this is not a real-life occurrence. Therefore, the probability of a multiperson outbreak is 0%. No real people are affected, and no real symptoms or health conditions are being discussed.
Source: Synthetic **Snippet:** 'lotta peeps attedning The Yellowstone National Park seemt to have conjunctivitis with allergy and IntenseÂ itching with concerned'	**Probability:** 70% **Size:** hundreds **Type:** Allergic **Severity:** Mild **Etiology:** Not applicable	The snippet suggests that multiple people visiting Yellowstone National Park are experiencing symptoms of conjunctivitis, specifically intense itching, which is often associated with an allergic reaction. However, without more specific information, it\'s difficult to determine the exact number of people affected or the cause of the outbreak. The severity is considered mild as conjunctivitis, while uncomfortable, is typically not life-threatening.
Source: Synthetic **Snippet:** 'vision problems with clearly an outbreak in Bourbon Street complinaing about pink eyes with bleeding and fluid-filled, blister-like growths on top of the eye surface (chemosis) with eyelid swelling and pain and decreased vision'	**Probability:** 90% **Size:** 100 **Type:** Infectious-AHC **Severity:** Severe **Etiology:** Virus	The snippet describes an outbreak of pink eye with severe symptoms such as bleeding, blister-like growths, and decreased vision, which are indicative of Acute Hemorrhagic Conjunctivitis (AHC), a severe form of conjunctivitis. The location is specified as Bourbon Street, and the use of the term “outbreak" suggests that multiple people are affected. Therefore…
Source: Forums **Snippet:** 'I just called my ped about the same thing! He said that viral pink eye is spreading around right now. My [child’s name] has had it for a week.'	**Probability:** 70% **Size:** 2 **Type:** Infectious **Severity:** Mild **Etiology:** Viral	The snippet suggests that there is a current outbreak of viral pink eye, as confirmed by a pediatrician. The user's child has been affected, and it's implied that others in the community may also be affected. However, without more specific information about the location or the number of people affected, it's difficult to provide a more accurate estimate. Pink eye is typically a mild condition, although it can be uncomfortable and highly contagious.
Source: Forums **Snippet:** '...to the rule, but they told us blocked tear duct and it was bacterial conjunctivitis. She ended up spreading it to big sister and me!'	**Probability:** 100% **Size:** 3 **Type:** Infectious **Severity:** Moderate **Etiology:** Bacterial	The snippet clearly mentions that the person had bacterial conjunctivitis and it was spread to two other people. This indicates a small-scale outbreak of an infectious disease. The severity is considered moderate as bacterial conjunctivitis can cause discomfort and temporary vision problems but is not typically life-threatening.
Source: Forums **Snippet:** '...know Pink eye that easily infectious. Hopefully everyone recovers with no lasting	**Probability:** 90% **Size:** thousands **Type:** Infectious **Severity:** Moderate	The snippet mentions an outbreak of pink eye in multiple countries, suggesting a widespread infectious event. However, without more specific information, it\'s difficult to determine the exact

damage. I just saw that India had an outbreak in August too, it sounds like it spread to Pakistan. Damn, I saw it spread to Vietnam too Conjunctivitis (pink eye) sees a spike in cases in Vietnam, India and Pakistan Oh my God. Close all travels. And I do mean ALL. ALL...'	**Etiology:** Not specified	number of people affected or the cause of the outbreak.
Source: Twitter/X **Snippet:** '[@username] god next thing we know there's gonna be someone spreading pink eye around	**Probability:** 10% **Size:** 1 **Type:** Infectious **Severity:** Not specified **Etiology:** Not specified	The snippet suggests a hypothetical situation of someone spreading pink eye, but it does not provide concrete evidence of an actual outbreak. The \/j\" at the end of the post indicates that the user is joking
Source: Twitter/X **Snippet:** 'When it rains it pours 😭😭 Pink eye-3 [surname] family-1 Here's hoping it's done spreading and we can finally kick its butt!'"	**Probability:** 70% **Size:** 4 **Type:** Infectious **Severity:** Moderate **Etiology:** Not specified	The snippet suggests a family (the [surname] family) has been affected by pink eye, with at least three members infected. The phrase 'Here's hoping it's done spreading" suggests it's an infectious outbreak. However...
Source: Twitter/X **Snippet:** 'There's some weird virus going around that has all the cold/flu symptoms, fever, super sore throat, ear pain and pink eye. It has ran through my household and I'm feeling like death right now. 😭😭😭'"	**Probability:** 80% **Size:** 5 **Type:** Infectious **Severity:** Strong **Etiology:** Viral conjunctivitis	The snippet suggests a multi-person outbreak of pink eye, likely viral given the accompanying symptoms of cold/flu, fever, sore throat, and ear pain. The user mentions it has ran through my household

Results of Comparative LLM Analyses and Validations

Results of the top performers for many of the different comparisons made are shown in Table 2, which focuses on evaluating the ability of LLMs to accurately classify key characteristics of potential epidemics from social media content, for outbreak probability, size, severity, type and for other health conditions mentioned. Descriptions of the table for each set of comparisons in the table are provided below for: validating LLMs against known values, comparing LLM results to each other validations, and comparing human validations to LLM results. For each section below the results are discussed (and shown in Table 2) for 2 top performing LLMs. Please note that the results tables with all LLMs for each section below are provided in the Appendix.

Table 2.i Validating LLMs against known values

We first assessed the ability of LLMs to extract outbreak characteristics from synthetic social media posts where the ground truth was known. We found that when evaluated against synthetic data, GPT-4 and Mixtral8x22b demonstrated strong performance in predicting outbreak probability, severity, and size, respectively, achieving high correlations with pre-defined values (Table 2.i). For example, GPT-4 achieved a correlation of 0.73 with the true outbreak probability, indicating a strong ability to distinguish between likely and unlikely outbreaks. Similarly, Mixtral 8x22B excelled in estimating outbreak size, achieving a correlation of 0.82. While both models showed high correlations overall, Mixtral 8x22B's performance on size estimation (correlation of 0.82) was notably higher than its performance on outbreak likelihood (correlation of 0.77). We also assessed LLMs ability to determine community setting (not shown in table) and found they were fairly successful. For example, comparing the "location" values extracted by Llama2 (using mean cosine similarity scoring, BERT) to the pre-defined community settings, we found community setting was ascertained well overall for all size locations (country, state, house, etc.) combined (mean (SD): 0.91 (0.16)) and for more specific community settings as well, for example daycares (0.88 (0.16)), college/dorms (0.92 (0.13)), and homes/condos (0.77 (0.22)).

Table 2.ii Comparative Validations using real posts

Moving beyond synthetic data, we validated LLM performance against each other on real-world forum posts and found LLMs demonstrated high correlation with GPT-4's assessments across various outbreak characteristics (Table 2.ii). For instance, both Sonnet 3 and Mixtral 8x22B showed high correlation with GPT-4 in assessing outbreak likelihood, with correlations of 0.93 and 0.82 respectively. When evaluating the models on Twitter data, the correlations were slightly lower, with Sonnet 3 achieving 0.82 and Mixtral 8x22B at 0.74 for outbreak likelihood.

Table 2.iii Comparing Human Validation of LLM-Identified Insights

To understand if insights identified by LLMs could be consistently recognized by humans, non-specialist human annotators assessed the same social media posts evaluated by GPT-4. Across both synthetic and real-world data, they showed substantial agreement with GPT-4, but with some discrepancies (Table 2.iii). For outbreak type, humans achieved a mean sensitivity of 0.81 and specificity of 0.89, indicating strong agreement with GPT-4 in categorizing outbreaks as infectious, allergic, or environmental. However, there were notable exceptions, such as in identifying AHC from real-world posts, where human sensitivity was considerably lower (0.57) compared to GPT-4 (0.83). Interestingly, the two expert ophthalmologists (MD1 & MD2) also demonstrated variability in their assessments, as seen in their mean sensitivity scores of 0.53 and 0.44 respectively, which is more in line with the non-specialist human annotators.

Table 2. Characterizing Epidemics: Based on content of synthetic and real world social media posts (forums and X/Twitter) LLMs can characterize outbreak probability, severity, size, type and other health conditions mentioned. We have tested 7 LLMs, high performing LLM results are shown in i and ii. Rows 1-3: Spearman, Row 4 onward: sensitivity and specificity. Cell colors indicate values ranging from 0.0 (red) to 0.5 (yellow) to 1.0 (green). **A.i) Validating LLMs vs Gold Standard** (red columns): With known preclassifieed values of epidemiological characteristics (severity, etc.) in synthetic posts, LLM's abilities to classify these characteristics from these same post's content were assessed. **A.ii) Comparative Inter-Model Validations** (blue cols.): In absence of pre-defined known values of synthetic posts, we used one LLM (GPT-4) to help validate results from another LLM's characterization of epidemics from synthetic and real-world post content (results shown are for claasification of online Forum posts. Comparison of results from other LLMS to GPT-4 are shown **A.iii) Human Validations** (green cols.): In absence of pre-defined known values we can also have humans assign values and compare to an LLMs characterization of epidemics from posts. Comparisons of GPT-4 vs. 2 non-MD human grader's classifications are are shown, Posts = synthetic, & real Twitter/X and Forum Posts).

LLM Characterizations	i. GPT 4	i. Mixtral 8x22b	ii. Sonnet	ii. Mixtral 8x22b	iii. NON md Person 1	iii. NON md Person 2
Outbreak (OB) Probability	0.73 (0.70, 0.76)	0.77 (0.74, 0.79)	0.93 (0.90, 0.95)	0.82 (0.77, 0.86)	0.59 (0.41, 0.72)	0.62 (0.4 0.74)
OB Severity of Cases	0.82 (0.79, 0.85)	0.80 (0.77, 0.83)	1.0 (0.97, 1.0)	0.99 (0.95, 1.0)	0.73 (0.59, 0.83)	0.78 (0.6 0.86)
OB Size (# of Cases)	0.54 (0.47, 0.59)	0.37 (0.30, 0.43)	0.95 (0.92, 0.97)	0.82 (0.72, 0.89)	0.98 (0.91, 1.0)	0.98 (0.8 1.0)
OUTBREAK TYPE						
Infectious, Sensitivity	1.0 (448/450)	1.0 (450/450)	0.99 (100/101)	1.0 (101/101)	1.0 (57/57)	0.98 (47/48)
Infectious, Specificity	0.68 (191/281)	0.64 (181/281)	0.88 (7/8)	0.75 (6/8)	0.68 (21/31)	0.50 (20/40)
Allergic, Sensitivity	0.57 (87/153)	0.54 (82/153)	(0/0)	(0/0)	0.93 (13/14)	0.88 (14/16)
Allergic, Specificity	1.0 (578/578)	1.0 (578/578)	1.0 (109/109)	1.0 (109/109)	0.97 (72/74)	0.99 (71/72)
AHC, Sensitivity	0.56 (83/148)	0.52 (77/148)	(0/0)	(0/0)	0.57 (4/7)	(0/0)

Table 2. Characterizing Epidemics: Based on content of synthetic and real world social media posts (forums and X/Twitter) LLMs can characterize outbreak probability, severity, size, type and other health conditions mentioned. We have tested 7 LLMs, high performing LLM results are shown in i and ii. Rows 1-3: Spearman, Row 4 onward: sensitivity and specificity. Cell colors indicate values ranging from 0.0 (red) to 0.5 (yellow) to 1.0 (green). **A.i) Validating LLMs vs Gold Standard** (red columns): With known preclassifieed values of epidemiological characteristics (severity, etc.) in synthetic posts, LLM's abilities to classify these characteristics from these same post's content were assessed. **A.ii) Comparative Inter-Model Validations** (blue cols.): In absence of pre-defined known values of synthetic posts, we used one LLM (GPT-4) to help validate results from another LLM's characterization of epidemics from synthetic and real-world post content (results shown are for claasification of online Forum posts. Comparison of results from other LLMS to GPT-4 are shown **A.iii) Human Validations** (green cols.): In absence of pre-defined known values we can also have humans assign values and compare to an LLMs characterization of epidemics from posts. Comparisons of GPT-4 vs. 2 non-MD human grader's classifications are are shown, Posts = synthetic, & real Twitter/X and Forum Posts).

LLM Characterizations	i. GPT 4	i. Mixtral 8x22b	ii. Sonnet	ii. Mixtral 8x22b	iii. NON md Person 1	iii. NON md Person 2
AHC, Specificity	1.0 (581/583)	0.99 (580/583)	1.0 (109/109)	1.0 (109/109)	0.99 (80/81)	0.94 (83/88)
Environ., Sensitivity	1.0 (103/103)	0.95 (98/103)	0.88 (7/8)	0.75 (6/8)	0.67 (6/9)	(1/3)
Environ., Specificity	0.99 (624/628)	0.98 (618/628)	1.0 (101/101)	1.0 (101/101)	0.99 (78/79)	0.93 (79/85)
HEALTH CONDITIONS						
Conjunct., Sensitivity	0.85 (816/961)	0.95 (910/961)	0.95 (219/231)	0.91 (210/231)	0.88 (15/17)	0.88 (15/17)
Conjunct., Specificity	0.89 (170/191)	0.84 (160/191)	0.71 (42/59)	0.75 (44/59)	1.0 (33/33)	1.0 (33/33)
COVID-19, Sensitivity	1.0 (9/9)	1.0 (9/9)	0.83 (29/35)	0.66 (23/35)	0.86 (6/7)	0.86 (6/7)
COVID-19, Specificity	0.99 (1131/1143)	0.99 (1135/1143)	0.96 (246/255)	0.98 (249/255)	1.0 (43/43)	1.0 (43/43)
Influenza, Sensitivity	0.96 (46/48)	0.96 (46/48)	0.60 (6/10)	0.70 (7/10)	0.62 (5/8)	0.83 (5/6)
Influenza, Specificity	1.0 (1102/1104)	1.0 (1104/1104)	0.97 (272/280)	0.99 (277/280)	1.0 (42/42)	1.0 (44/44)
Intestinal Flu, Sensitivity	0.33 (48/145)	0.36 (52/145)	0.50 (4/8)	0.62 (5/8)	(2/2)	(2/2)
Intestinal Flu, Specificity	0.98 (990/1007)	0.99 (997/1007)	1.0 (281/282)	0.99 (280/282)	0.92 (44/48)	0.92 (44/48)
Croup, Sensitivity	0.89 (17/19)	0.74 (14/19)	(0/0)	(0/0)	(5/5)	(5/5)
Croup, Specificity	1.0 (1133/1133)	1.0 (1133/1133)	1.0 (290/290)	1.0 (290/290)	1.0 (45/45)	1.0 (45/45)
Lice, Sensitivity	0.76 (19/25)	0.56 (14/25)	(0/0)	(0/0)	0.88 (7/8)	0.88 (7/8)
Lice, Specificity	1.0 (1127/1127)	1.0 (1127/1127)	1.0 (290/290)	1.0 (290/290)	1.0 (42/42)	1.0 (42/42)
Leg Break, Sensitivity	0.88 (15/17)	0.65 (11/17)	(0/0)	(0/0)	(1/1)	(1/1)
Leg Break, Specificity	1.0 (1134/1135)	1.0 (1135/1135)	1.0 (290/290)	1.0 (290/290)	1.0 (49/49)	1.0 (49/49)

Specialist Validations of Conjunctivitis Etiological Causes

We further evaluated the ability of GPT-4 to arrive at the same conclusions as two expert ophthalmologists regarding the cause of infectious conjunctivitis from social media posts. Across all conjunctivitis cause types, GPT-4 showed high agreement with the specialist's assessments (Table 3), particularly in ruling out conditions (high specificity). For viral conjunctivitis, GPT-4 mirrored specialists' high specificity with a value of 0.96, however, GPT-4s sensitivity for identifying cases where specialists did identify viral conjunctivitis was lower at 0.38. This suggests that while GPT-4 is highly reliable in excluding viral conjunctivitis when it is not present, it might miss some actual cases. GPT-4 had very high agreement with a specialist's identification of acute hemorrhagic conjunctivitis (AHC), a rare form of conjunctivitis with high severity, achieving a specificity of 0.99 and a sensitivity of 0.71. The performance varied across different data sources: for combined data, GPT-4 had a mean sensitivity of 0.53 and specificity of 0.89. This performance dropped slightly for tweets (0.1 sensitivity, 0.81 specificity), suggesting that the limited context in tweets makes it harder for GPT-4 to identify cases. Conversely, GPT-4 performed better on synthetic data (0.63 sensitivity, 0.94 specificity), likely because the synthetic data was more standardized and potentially richer in relevant clinical information. Notably, GPT-4's performance on other forum data (0.47 sensitivity, 0.87 specificity) fell between tweets and synthetic data.

Table 3. Specialist Validations of Conjunctivitis Etiological Causes. Same data and method as “Table 2.iii” above, but ophthalmologist-assigned conjunctivitis etiological cause vs. GPT4-assigned etiology. For the set of 90 posts (38 Synthetic, 26 Twitter/X, 26 Forum) used in Table 2, etiological cause outputs from a modified LLM GPT-4 prompt are compared to that of the two trained surface disease practicing ophthalmologists and the MD specialist was treated as the gold standard All rows; sensitivity and specificity. Cell colors indicate values ranging from 0.0 (red) to 0.5 (yellow) to 1.0 (green). Results with denominator of 5 or less are shown as a fraction. Results were calculated for all 3 sources of posts combined (columns 1-2, as well as for each individual source (Synthetic, columns 3-4; Twitter/X columns 5-6; Forums columns 7-8).

Etiology	All Posts, MD1	All Posts, MD2	Synthetic, MD1	Synthetic, MD2	Twitter/X MD1	Twitter/X MD2	Forums MD1	Forums, MD2
Viral, Sensitivity	0.38 (23/61)	0.32 (23/72)	0.55 (12/22)	0.39 (11/28)	0.10 (2/20)	0.10 (2/20)	0.47 (9/19)	0.42 (10/24)
Viral, Specificity	0.96 (27/28)	0.94 (16/17)	1.0 (16/16)	0.90 (9/10)	1.0 (6/6)	1.0 (6/6)	0.83 (5/6)	(1/1)
Bacterial, Sensitivity	(3/5)	0.50 (3/6)	(3/3)	(0/1)	(0/1)	(3/4)	(0/1)	(0/1)
Bacterial, Specificity	0.89 (75/84)	0.89 (74/83)	0.91 (32/35)	0.84 (31/37)	0.80 (20/25)	0.91 (20/22)	0.96 (23/24)	0.96 (23/24)
AHC, Sensitivity	0.71 (5/7)	(3/4)	0.71 (5/7)	(3/4)	(0/0)	(0/0)	(0/0)	(0/0)
AHC, Specificity	0.99 (81/82)	0.96 (82/85)	0.97 (30/31)	0.91 (31/34)	1.0 (26/26)	1.0 (26/26)	1.0 (25/25)	1.0 (25/25)
COVID-19, Sensitivity	0.50 (3/6)	0.50 (3/6)	(0/2)	(0/1)	(0/0)	(0/1)	(3/4)	(3/4)
COVID-19, Specificity	1.0 (83/83)	1.0 (83/83)	1.0 (36/36)	1.0 (37/37)	1.0 (26/26)	1.0 (25/25)	1.0 (21/21)	1.0 (21/21)
Unspecified, Sensitivity	(0/2)	(2/2)	(0/0)	(1/1)	(0/2)	(1/1)	(0/0)	(0/0)
Unspecified, Specificity	0.60 (52/87)	0.62 (54/87)	0.84 (32/38)	0.86 (32/37)	0.25 (6/24)	0.32 (8/25)	0.56 (14/25)	0.56 (14/25)

Discussion

Principal Findings

In this study, we aimed to achieve several key objectives related to the use of large language models (LLMs) in the field of infoveillance. Specifically, we sought to determine whether an LLM could accurately classify significant epidemiological features from individual social media posts, including data from other platforms such as online forums. The epidemiological characteristics of interest included the probability of an outbreak, the number of people affected, the presence of other conditions mentioned alongside the outbreaks, and characteristics of conjunctivitis cases such as severity, type, and etiology._Our general hypothesis was that the correlation, sensitivity, and specificity results would vary among different LLMs. We anticipated that some LLMs would outperform others and that certain simpler features (e.g., probability or type of outbreak) would be classified more accurately than more complex features (e.g., severity or etiology of cases).

Overall, we found several key findings that largely supported our hypotheses: We found several LLMs from different developers performed quite well for many of the characteristics being assessed, demonstrating the capability to discern numerous epidemiological characteristics of outbreaks from both synthetic and real social media posts, including forums which we had not studied before [50]. This includes validation against known values derived from synthetic datasets or through human expert assessments. The top performing LLMs were able to extract approximate outbreak probability, outbreak size, outbreak severity, and, to a lesser extent, outbreak type and potentially etiology from social media posts in many cases.

Other Findings

Validating LLMs against known values of synthetic posts

When we evaluated the capability of LLMs to accurately classify outbreak characteristics based on synthetic social media posts with known epidemiological features, several such as GPT-4 showed a strong ability with a correlation of 0.73, indicating its effectiveness in distinguishing between potential and non-potential outbreaks. Mixtral8x22B, on the other hand, excelled in estimating the size of outbreaks, achieving a correlation of 0.82. The results underscore the potential utility in early detection scenarios where rapid assessment is crucial. This validation against known values underscores LLMs' potential as reliable tools in public health monitoring, offering a scalable solution for early outbreak detection and assessment without the need for extensive manual data analysis. **Limitations/Weaknesses:** The primary limitation in this portion of our study is the reliance on synthetic data, which might not always capture the complexity and variability of real-world data. The models' performance might differ when exposed to less controlled, more diverse datasets. Future research could focus on enhancing LLMs' training with diverse and complex synthetic datasets that mirror real-world variability in outbreak reports. As we have noted, another limitation of our approach, in the absence of sufficient gold standard pink eye epidemics is that we have not shown that our assessed outbreak probabilities (which may be interpreted as being similar to Bayesian degrees of belief) are calibrated. In other words, we cannot conclude that among tweets with assessed outbreak probability X, the relative frequency of true outbreak probability is X.

Comparative Validations using real posts from multiple sources

We evaluated and compared the ability of LLMs to classify epidemiological characteristics from real-world data, comparing the performance of LLMs like Sonnet 3 and Mixtral 8x22B against GPT-4 and found that LLM models developed by independent vendors, showed high agreement in assessing components such as outbreak likelihood, with Sonnet 3 and Mixtral 8x22B achieving correlations of 0.93 and 0.82, respectively, when evaluating forum posts. The high degree of consistency across different LLMs when analyzing real-world data underscores the reliability of using these models as supplementary tools for outbreak detection and characterization. This method of multi-model validation not only bolsters the confidence in the findings of an LLM but also suggests a potentially streamlined, automated approach to cross validate epidemiological results from LLM assessments, which traditionally relies heavily on human expertise and intervention. Limitations: Variations in performance across different social media sources like Twitter and forums indicate that LLMs might be sensitive to the format and quality of the input data, which can limit their generalizability. Further studies should explore the development of platform-agnostic LLMs that maintain high accuracy across various social media platforms, enhancing their utility in diverse public health contexts. An individual LLMs can also include inherent bias. To further improve our models and potentially help mitigate any such bias, in future studies of a larger number of LLMs we could take advantage of the unique strengths and weaknesses found per LLM, to generate an ensemble model and assess its overall ability to improve our sensitivity and specificity. We also did not include any

open-weight LLM models (which can provide flexibility of local deployment, not reliant on vendors, have controlled costs, and can be fine-tunable) in this current study. Including such models in a future ensemble model study (of LLMs with a wide range of costs) could also investigate tradeoffs between performance and cost.

Human Validation of LLM-classified outbreak characteristics

In our human validation studies, we found substantial consistency in some areas and significant discrepancies in others. For instance, while humans generally agreed with GPT-4 in categorizing outbreaks as infectious, allergic, or environmental, there were notable inconsistencies in the sensitivity and specificity values, especially for more severe or rare conditions. Human graders and GPT-4 showed a high level of agreement in identifying infectious outbreaks, with sensitivity values close to 1.0 across different LLMs and human assessments. This consistency suggests that both humans and LLMs are effective at recognizing clear signs of infectious diseases in social media posts. This consistency supports the use of LLMs as a tool to assist public health monitoring and potentially lighten the load on human analysts. **Limitations**: However, inconsistencies, especially in severe or rare conditions, highlight the challenges in using LLMs as standalone diagnostic tools. These discrepancies underscore the need for ongoing human oversight and the potential for LLMs to complement, rather than replace, human judgment in assessing urgency of detected potential outbreaks.

Specialist Validation of etiological cause

For etiology, the trained ophthalmologists demonstrated a high specificity in identifying non-infectious causes of conjunctivitis indicating a strong agreement on what clearly does not constitute a certain type of conjunctivitis based on the posts. This suggests that medical training can lead to a consistent ability to rule out certain conditions. But the sensitivity to detect specific types of infectious conjunctivitis varied, such as viral and AHC. Despite varied sensitivity for detecting AHC etiology, the sensitivities of 0.71 and (3 out of 4, too small to reliably estimate a sensitivity), in light of specificity values 0.99 and 0.96 suggest LLMs could reliably rule out irrelevant posts and find some small early AHC outbreaks (per MD1, with high sensitivity), which one could make the case has value from a public health perspective. **Limitations**: The variation seen between the specialists could stem from individual differences in clinical experience, judgment, or even the subjective interpretation of the social media content's context and descriptions. This may highlight the nuanced nature of identifying complex conditions from unstructured data, a concept that is also true in clinical settings often requiring laboratory testing in order to consistently define the type or etiology of cases [78, 79]. They also likely also reflect constraints of social media data, which may lack comprehensive clinical details, affecting the reliability of such data on its own for disease identification. Future studies could conduct targeted social media campaigns to interact with the users posting on social media and survey them on any clinical diagnoses or even collect samples for laboratory study.

Future studies of epidemics are possible using LLM extracted values

Our findings suggest that by using social media and LLMs we can leverage assessed characteristics of a large number of known outbreaks to test generalizable hypotheses about epidemics. For example, we might be able to test if a group of cases have higher (reported, "digital") severity during outbreaks vs. individual non-outbreak cases. Our approach may provide a new opportunity to leverage information about epidemics in small community settings to ask research questions at even a household epidemic level. For example, can we use information about these micro-epidemics to predict the emergence of larger more traditionally detectable epidemics, in particular, the severe ones? A future study could collect social media from the time and location of known infectious conjunctivitis outbreaks with known severity, and primary etiologies and assess the ability of LLMs to characterize them, including etiology, such a study could assess and characterize symptoms as well, if known[80-83]. Finally, with our validated findings of reliable ability to identify mentions of other health conditions, our epidemic detection and characterization approaches could be studied for their ability to reveal early significant other public health concerns from social media, such as flu outbreaks, COVID-19, foodborne illness, and contaminated products.

Conclusions

Our findings suggest that top-performing LLMs can reliably infer the probability of and classification of other conjunctivitis outbreaks as determined from multiple sources of social media data. This could lead to the ability to better understand outbreak dynamics. This approach could uncover outbreaks that might not yet have been detected, or well-characterized in traditional clinical or epidemiological reporting datasets or systems. Moreover, identifying higher-risk severe epidemics through social media monitoring could enable public health authorities to take timely, targeted actions to address severe outbreaks earlier, potentially mitigating their impact.

Acknowledgements

This work was supported in part by a grant from the National Eye Institute of the National Institutes of Health (1R01EY024608-01A1; Principal Investigator [PI]: TML), a Core Grant for Vision Research from the National Eye Institute (EY002162; PI: Erik M Ullian), and an Unrestricted Grant from Research to Prevent Blindness (PI: JLD). The funding organizations had no role in the design or conduct of this research. The authors gratefully acknowledge the contributions of the 2 human raters used in the validation of the large language model responses. The subject of the paper concerns analysis of the results of generative artificial intelligence (AI)–where we studied the use of the generative AI tools GPT-3.5 and GPT-4 (OpenAI) for LLM-based analyses of social media posts. Generative AI was not used in ideation, analysis, or manuscript preparation, other than for assistance from GPT-3.5 in helping us to refine the R code to improve the formatting of the figures and some tables and some revision editing.

Data availability

The primary social media data sets analyzed during this study are not publicly available owing to our terms of use agreement with the Brandwatch platform but are available from the corresponding author using a data sharing agreement on reasonable request. We have placed the posts in the Qualitative Data Repository via the University of California San Francisco.

Conflicts of interest

None declared in relation to the topic of this manuscript

Abbreviations

API: application programming interface
GPT: Generative Pre-trained Transformers
LLM: large language model
BERT: Bidirectional Encoder Representations from Transformers

Multimedia appendix 1

Boolean query, additional code, and additional tables

Appendix Item 1: Boolean Query

The details of the query are as shown below (not shown: additional exclusion of terms to remove posts about animals, obscenities, artistic or literary references, celebrities and politicians):
“(("aankh aana" OR "azoumounou" OR "Bindehautentzündung" OR "bindhinneinflammation" OR "bindvliesontsteking" OR "congiuntivite" OR "conjonctivite" OR "conjunctivită" OR "conjunctivitis" OR "conjuntivite" OR "conjuntivitis" OR "je woz" OR "konjonktivit" OR "konjuktiviti" OR "konjuktivitis" OR "konjunktivitis" OR "konjunktivīts" OR "konjunktivitt" OR "konjunktyvitas" OR "kötőhártya-gyulladás" OR "œil rose" OR "pinkeye" OR "pink eye" OR "sidekalvontulehdus" OR "sidekestapõletik" OR "Viêm kết mạc" OR "viem ket mac" OR "zánětspojivek" OR "zapaleniespojówek" OR "zápalspojiviek" OR " конјуктивит " OR " конъюнктивит " OR " конюнктивит " OR " التهابالملتحمة" OR " آنكهكالالهوجانا" OR " □□□□ □□ □□□ □□ □" OR " □□□□□ □□ □□□□□ □□" OR " □□□□□ " OR "□□□ " OR "□□□ " OR "□□□ " OR "□□□"))
AND
(our OR my OR hers OR his OR her OR him OR their OR he OR she OR he OR we)
AND
(
(((outbreak OR epidemic OR "going around" OR "spreading" OR cancelled OR closed OR postponed OR "the entire" OR "everybody" OR "everyone" OR "my entire" OR "the whole") NEAR/5 (has OR got OR gotten OR "coming down with" OR getting OR sick OR spreading)) NEAR/15 (pollen OR smoke OR swimming OR pool OR pollution OR fires OR pus OR "stuck shut" OR green OR yellow OR swollen OR edema OR itch* OR burn*))
OR
(((outbreak OR epidemic OR "going around" OR "spreading" OR cancelled OR closed OR postponed OR "the entire" OR "everybody" OR "everyone" OR "my entire" OR "the whole") NEAR/5 (has OR got OR gotten OR "coming down with" OR getting OR sick OR spreading)) NEAR/5 (bacteri* OR virus OR viral OR allergic))
)

Appendix Item 2: Algorithm for generating synthetic posts.

```
```{r, eval=FALSE}
function generateSyntheticTweets(numTweets, probDB, settDB, conjDB, orgDB, sevDB, altProb)
   for i = 1 to numTweets do
      probComponent = randomly select a component from probDB
      settComponent = randomly select a component from settDB
      sevComponent = randomly select a component from sevDB
```
```

```
        if random number < altProb then
            mainPhrase = randomly select a phrase from alternativePhrases
            isAlt = true
        else
            mainPhrase = randomly select a phrase from conjunctivitisRelatedPhrases
            isAlt = false
        end if

        j1, j2, j3, j4, j4a = randomly select joining phrases from joiningPhraseLists

        if isAlt then
            probScore = 5
        else
            probScore = get probability score from probComponent
        end if

        sizeCat = get size category from settComponent
        severity = get severity from sevComponent

        if probScore <= 25 and sizeCat <= 50 and severity is "mild" then
            conjType = randomly select from all conjunctivitis types in conjDB
        else
            conjType = randomly select from non-environmental types in conjDB
        end if

        if conjType is "blank" then
            conjComponent = ""
            j4 = ""
            conjSeverity = "mild"
        else
            j5 = randomly select a joining phrase
            conjComponent = concatenate 2 random components from conjDB where type=conjType, joined by j5
            conjSeverity = get severity from first selected conjunctivitis component
        end if

        if conjType is "allergic" or "environmental" or "blank" then
            if random number < 1.0 then
                tweet = concatenate(probComponent, j1, settComponent, j2, mainPhrase, j3, conjComponent, j4, sevComponent)
            else
                tweet = concatenate(sevComponent, j4, probComponent, j1, settComponent, j2, mainPhrase, j3, conjComponent)
            end if
            orgComponent = ""
            orgType = null
            orgSeverity = ""
```

```
    else if isAlt then
      j5a = randomly select a joining phrase
      tweet = concatenate(probComponent, j1, settComponent, j2, mainPhrase, j3, conjComponent, j4, sevComponent)
      orgComponent = concatenate 2 random non-eye-related components from orgDB, joined by j5a
      orgType = null
      orgSeverity = get severity from selected organism components
    else
      orgType = randomly select an organism type from orgDB
      if orgType is not "blank" then
        j5a = randomly select a joining phrase
        orgComponent = concatenate 2 random components from orgDB where type=orgType, joined by j5a
        if random number < 0.5 then
          tweet = concatenate(probComponent, j1, settComponent, j2, mainPhrase, j3, conjComponent, j4, orgComponent, j4a, sevComponent)
        else
          tweet = concatenate(sevComponent, j4a, probComponent, j1, settComponent, j2, mainPhrase, j3, conjComponent, j4, orgComponent)
        end if
        orgSeverity = get severity from selected organism components
      else
        if random number < 0.5 then
          tweet = concatenate(probComponent, j1, settComponent, j2, mainPhrase, j3, conjComponent, j4, sevComponent)
        else
          tweet = concatenate(sevComponent, j4, probComponent, j1, settComponent, j2, mainPhrase, j3, conjComponent)
        end if
        orgComponent = ""
        orgSeverity = ""
      end if
    end if

    maxSeverity = max(severity, conjSeverity, orgSeverity)

    store generated tweet and metadata
  end for

  return generated tweets and metadata
end function
```

Appendix Item 3: Qualtrics Instructions that Human Graders Were Provided, and Additional Etiological Cause GPT-4 Prompt

1) Outbreak Probability

The instructions in the Qualtrics survey were: *“Here is a text portion of our request "prompt" that we had provided to the LLM to get its responses: [LLM PROMPT WAS HERE]. Below, is a modified version of the same request for you: Please read each snippet in blue and then based upon that content, indicate* ***how certain are you that this snippet is about a multiperson outbreak of pink eye occurring at the time the snippet was posted****? For each row,* ***you can hover your mouse over the "hover here for full instructions" to review the entire LLM prompt instructions again,*** *including when to assign a 0% or other very low likelihood (for example if it you are certain it is fictional or a rumor proven incorrect etc). Please try the hover function to be sure you can see the pop-up of the full portion of the prompt instructions to the LLMs.”*

2) Outbreak Severity, Size, Type

The instructions in the Qualtrics survey were: “*Here is a text portion of our request "prompt" that we had provided to the LLM to get its responses: [LLM PROMPT WAS HERE]. Below, in each row we will provide a modified version of the above prompt as requests for you. For each row, we will ask you to read the blue snippet and then estimate the info about the mentioned conjunctivitis cases: Severity, Number of Cases, Type of Cases. Here are guidelines and definitions to please read before proceeding (these will be briefly described in each row for your convenience): <u>1) Severity of these cases on health:</u> "****NOT SPECIFIED****" (too hard to tell severity), "****MILD****" (not significant), "****MODERATE"**** (has some impact on health),"****STRONG****" (has serious health impact), "****SEVERE****" (life-threatening). <u>2) Number of people affected:</u> Best guess based in part on the estimated population of the location and how many of them might be affected. Type the <u>NUMBER(integer)</u>--or--<u>type "N" for "not defined"</u>. Examples: 2 | 5,500 | N . <u>3) Type of conjunctivitis cases:</u> "****NOT SPECIFIED****" (select this if it's too hard to feel very confident guessing the type of conjunctivitis). "****ALLERGIC****" (select this if you think allergic conjunctivitis for example if it's about allergy season, or symptoms of allergy or pollen). "****INFECTIOUS****" (select this if you think it may be viral or bacterial conjunctivitis, but NOT AHC). "****ENVIRONMENTAL****" (select this if you think the conjunctivitis is from swimming pools, pollution, toxic spills, smoke, wildfires, drug-usage). "****AHC-INFECTIOUS****" (select this if you suspect is may be AHC -- acute hemorrhagic conjunctivitis, also known epidemic keratoconjunctivitis, hemorrhagic conjunctivitis -- is very severe and typically includes extremely red, bloody or bleeding eyes, vision loss and other severe symptoms).”*

3) Health Conditions:

The instructions in the Qualtrics survey were: *“Here is a text portion of our request "prompt" that we had provided to the LLM to get its responses: [LLM PROMPT WAS HERE]. For every snippet provide the following: tell us the health condition(s) or disease(s) being discussed (e.g. conjunctivitis, flu, broken leg, etc.)" Below is our request for you, a modified version of the prompt above.* ***For each row, read the blue snippet and*** *based upon that content, please provide the following: Tell us* ***what health condition(s) or disease(s) are being discussed (e.g. conjunctivitis, flu, broken leg, etc.)****? (please <u>DO</u> provide health condition/disease names, <u>even if you suspect this is not about a real case of that condition</u>). Please list health conditions (e.g. conjunctivitis, flu, broken leg, etc) but do <u>not</u> list individual symptoms (e.g. itchy eyes, fever, pain) unless you feel that that symptom is a known health condition.”.*

4) Specialist Validation of Conjunctivitis Etiological cause

The additional GPT-4 prompt for the conjunctivitis etiological cause included: “…*Cause: Please assign an infectious or non-infectious cause, choosing one of the following etiology categories: 1. \"NOT INFECTIOUS\" (select this if you think it seems not infectious), 2. \"INFECTIOUS:*

UNSPECIFIED\" (select this if it's too hard to feel very confident guessing the type of organism but it seems infectious), 3. \"BACTERIAL\" (select this if you think it may be bacterial conjunctivitis), 4. \"VIRAL: UNSPECIFIED\" (select this if you think it may be viral conjunctivitis but not COVID-19 or AHC), 5. \"VIRAL: COVID-19\" (select this if you think it may be viral COVID-19 conjunctivitis more than other forms of viral conjunctivitis), 6. \"VIRAL: AHC\" (select this if you think it may be viral AHC conjunctivitis more than other forms of viral conjunctivitis)...."

The instructions in the Qualtrics survey for the two trained surface disease practicing ophthalmologists to assess the conjunctivitis etiological cause included: ***"...For each row, we will ask you to read the blue snippet and then estimate the conjunctivitis cases: Severity, Type of Cases and Etiology of Cases.*** *Here are guidelines and definitions to please read before proceeding (these will be briefly described in each row for your convenience)...**Etiology of cases: "NOT INFECTIOUS"** (select this if you think it seems not infectious); **"INFECTIOUS: UNSPECIFIED"** (select this if it's too hard to feel very confident guessing the type of organism but it seems infectious); **"BACTERIAL"** (select this if you think it may be bacterial conjunctivitis); **"VIRAL: UNSPECIFIED"** (select this if you think it may be viral conjunctivitis but not COVID-19 or AHC); **"VIRAL: COVID-19"** (select this if you think it may be viral COVID-19 conjunctivitis more than other forms of viral conjunctivitis); **"VIRAL: AHC"** (select this if you think it may be viral AHC conjunctivitis more than other forms of viral conjunctivitis). Once you become familiar with the definitions above, please now proceed with each row below"*

Appendix Item 4: Full Set of LLMs, Post Sources and Graders (Expansion of Table 2)

Supplement Table S2.i Validating LLMs against known values – all 7 LLMs shown

LLM Characterizations	GPT 4	Sonnet 3	Mixtral 8x22B	Opus	LlaMa 3 70B	GPT 4o	GPT 3.5
Outbreak (OB) Probability	0.73 (0.70, 0.76)	0.73 (0.70, 0.76)	0.77 (0.74, 0.79)	0.72 (0.69, 0.75)	0.71 (0.67, 0.74)	0.71 (0.68, 0.74)	0.71 (0.68, 0.74)
OB Severity of Cases	0.82 (0.79, 0.85)	0.81 (0.78, 0.84)	0.80 (0.77, 0.83)	0.91 (0.89, 0.93)	0.86 (0.83, 0.88)	0.86 (0.83, 0.88)	0.78 (0.75, 0.81)
OB Size (# of Cases)	0.54 (0.47, 0.59)	0.45 (0.39, 0.52)	0.37 (0.30, 0.43)	0.67 (0.63, 0.71	0.45 (0.37, 0.53)	0.56 (0.51, 0.61)	0.63 (0.52, 0.72)
OUTBREAK TYPE							
Infectious, Sensitivity	1.0 (448/450)	0.99 (445/450)	1.0 (450/450)	0.99 (444/450)	0.98 (443/450)	0.99 (444/450)	1.0 (449/450)
Infectious, Specificity	0.68 (191/281)	0.62 (174/281)	0.64 (181/281)	0.52 (146/281)	0.42 (117/281)	0.66 (185/281)	0.33 (93/281)
Allergic, Sensitivity	0.57 (87/153)	0.41 (63/153)	0.54 (82/153)	0.37 (57/153)	0.12 (18/153)	0.52 (79/153)	0.34 (52/153)
Allergic, Specificity	1.0 (578/578)	1.0 (578/578)	1.0 (578/578)	1.0 (578/578)	1.0 (578/578)	1.0 (578/578)	1.0 (578/578)
AHC, Sensitivity	0.56 (83/148)	0.28 (42/148)	0.52 (77/148)	0.52 (77/148)	0.56 (83/148)	0.66 (97/148)	0.034 (5/148)
AHC, Specificity	1.0 (581/583)	1.0 (581/583)	0.99 (580/583)	1.0 (583/583)	0.99 (580/583)	0.98 (571/583)	1.0 (583/583)
Environ., Sensitivity	1.0 (103/103)	1.0 (103/103)	0.95 (98/103)	0.96 (99/103)	0.85 (88/103)	1.0 (103/103)	0.35 (36/103)
Environ., Specificity	0.99 (624/628)	0.97 (612/628)	0.98 (618/628)	1.0 (628/628)	0.99 (620/628)	1.0 (625/628)	1.0 (626/628)
HEALTH CONDITIONS							
Conjunct., Sensitivity	0.85 (816/961)	0.87 (839/961)	0.95 (910/961)	0.85 (819/961)	0.96 (922/961)	0.83 (801/961)	0.75 (718/961)
Conjunct., Specificity	0.89 (170/191)	0.89 (170/191)	0.84 (160/191)	0.84 (160/191)	0.77 (148/191)	0.91 (173/191)	0.83 (158/191)
COVID-19, Sensitivity	1.0 (9/9)	1.0 (9/9)	1.0 (9/9)	1.0 (9/9)	1.0 (9/9)	1.0 (9/9)	0.78 (7/9)
COVID-19, Specificity	0.99 (1131/1143)	0.99 (1133/1143)	0.99 (1135/1143)	0.99 (1129/1143)	0.99 (1133/1143)	0.99 (1136/1143)	1.0 (1139/1143)
Influenza, Sensitivity	0.96 (46/48)	0.94 (45/48)	0.96 (46/48)	1.0 (48/48)	0.96 (46/48)	0.88 (42/48)	0.79 (38/48)
Influenza, Specificity	1.0 (1102/1104)	0.99 (1098/1104)	1.0 (1104/1104)	1.0 (1102/1104)	1.0 (1101/1104)	1.0 (1104/1104)	1.0 (1104/1104)
Int. Flu, Sensitivity	0.33 (48/145)	0.27 (39/145)	0.36 (52/145)	0.31 (45/145)	0.26 (38/145)	0.27 (39/145)	0.15 (22/145)
Int. Flu, Specificity	0.98 (990/1007)	0.99 (1001/1007)	0.99 (997/1007)	0.98 (990/1007)	0.99 (999/1007)	0.99 (1001/1007)	1.0 (1005/1007)
Croup, Sensitivity	0.89 (17/19)	0.95 (18/19)	0.74 (14/19)	0.95 (18/19)	0.79 (15/19)	0.74 (14/19)	0.42 (8/19)
Croup, Specificity	1.0 (1133/1133)	1.0 (1133/1133)	1.0 (1133/1133)	1.0 (1133/1133)	1.0 (1133/1133)	1.0 (1133/1133)	1.0 (1133/1133)
Lice, Sensitivity	0.76 (19/25)	0.84 (21/25)	0.56 (14/25)	0.80 (20/25)	0.68 (17/25)	0.60 (15/25)	0.12 (3/25)
Lice, Specificity	1.0 (1127/1127)	1.0 (1127/1127)	1.0 (1127/1127)	1.0 (1127/1127)	1.0 (1127/1127)	1.0 (1127/1127)	1.0 (1127/1127)
Leg Break, Sensitivity	0.88 (15/17)	0.65 (11/17)	0.65 (11/17)	0.24 (4/17)	0.53 (9/17)	0.53 (9/17)	0.0 (0/17)
Leg Break, Specificity	1.0 (1134/1135)	1.0 (1134/1135)	1.0 (1135/1135)	1.0 (1135/1135)	1.0 (1135/1135)	1.0 (1134/1135)	1.0 (1135/1135)

Supplement Table S2.ii - Synthetic posts, Comparative Validations of 6 LLMs vs. GPT-4

LLM Characterizations	Sonnet	Mixtral 8x22b	Opus	LlaMa 3	GPT 4o	GPT 3.5
Outbreak (OB) Probability	0.92 (0.91, 0.93)	0.81 (0.79, 0.83)	0.83 (0.80, 0.86)	0.82 (0.80, 0.84)	0.91 (0.90, 0.92)	0.67 (0.63, 0.71)
OB Severity of Cases	1.0 (0.99, 1.0)	1.0 (0.99, 1.0)	0.99 (0.98, 1.0)	0.99 (0.98, 1.0)	0.99 (0.98, 1.0)	0.98 (0.96, 0.99)
OB Size (# of Cases)	0.81 (0.78, 0.84)	0.75 (0.71, 0.78)	0.79 (0.75, 0.82)	0.80 (0.76, 0.84)	0.75 (0.72, 0.79)	0.71 (0.61, 0.79)

LLM Characterizations	Sonnet	Mixtral 8x22b	Opus	LlaMa 3	GPT 4o	GPT 3.5
OUTBREAK TYPE						
Infectious, Sensitivity	0.97 (549/565)	0.99 (557/565)	0.98 (554/565)	0.97 (549/565)	0.96 (541/565)	0.99 (562/565)
Infectious, Specificity	0.90 (164/182)	0.92 (168/182)	0.77 (140/182)	0.60 (110/182)	0.95 (173/182)	0.47 (86/182)
Allergic, Sensitivity	0.70 (57/81)	0.90 (73/81)	0.65 (53/81)	0.21 (17/81)	0.86 (70/81)	0.59 (48/81)
Allergic, Specificity	1.0 (666/666)	1.0 (663/666)	1.0 (666/666)	1.0 (666/666)	1.0 (663/666)	1.0 (666/666)
AHC, Sensitivity	0.47 (40/85)	0.85 (72/85)	0.84 (71/85)	0.91 (77/85)	1.0 (85/85)	0.047 (4/85)
AHC, Specificity	1.0 (659/662)	0.99 (654/662)	0.99 (658/662)	0.99 (654/662)	0.96 (637/662)	1.0 (661/662)
Environ., Sensitivity	0.98 (103/105)	0.93 (98/105)	0.92 (97/105)	0.82 (86/105)	0.97 (102/105)	0.33 (35/105)
Environ., Specificity	0.98 (628/642)	0.99 (634/642)	1.0 (642/642)	0.99 (633/642)	1.0 (639/642)	1.0 (641/642)
HEALTH CONDITIONS						
Conjunct., Sensitivity	0.88 (739/837)	0.94 (790/837)	0.98 (822/837)	0.96 (805/837)	0.84 (705/837)	0.87 (726/837)
Conjunct., Specificity	0.62 (194/315)	0.52 (164/315)	0.91 (287/315)	0.49 (155/315)	0.64 (201/315)	0.92 (290/315)
COVID-19, Sensitivity	0.81 (17/21)	0.71 (15/21)	0.86 (18/21)	0.81 (17/21)	0.76 (16/21)	0.52 (11/21)
COVID-19, Specificity	1.0 (1129/1131)	1.0 (1129/1131)	1.0 (1126/1131)	1.0 (1129/1131)	1.0 (1131/1131)	1.0 (1131/1131)
Influenza, Sensitivity	0.90 (43/48)	0.92 (44/48)	0.96 (46/48)	0.92 (44/48)	0.85 (41/48)	0.79 (38/48)
Influenza, Specificity	0.99 (1096/1104)	1.0 (1102/1104)	1.0 (1100/1104)	1.0 (1099/1104)	1.0 (1103/1104)	1.0 (1104/1104)
Int. Flu, Sensitivity	0.52 (34/65)	0.63 (41/65)	0.60 (39/65)	0.55 (36/65)	0.52 (34/65)	0.34 (22/65)
Int. Flu, Specificity	0.99 (1076/1087)	0.98 (1066/1087)	0.98 (1064/1087)	0.99 (1077/1087)	0.99 (1076/1087)	1.0 (1085/1087)
Croup, Sensitivity	0.94 (16/17)	0.76 (13/17)	0.94 (16/17)	0.76 (13/17)	0.76 (13/17)	0.47 (8/17)
Croup, Specificity	1.0 (1133/1135)	1.0 (1134/1135)	1.0 (1133/1135)	1.0 (1133/1135)	1.0 (1134/1135)	1.0 (1135/1135)
Lice, Sensitivity	0.84 (16/19)	0.63 (12/19)	0.84 (16/19)	0.74 (14/19)	0.68 (13/19)	0.16 (3/19)
Lice, Specificity	1.0 (1128/1133)	1.0 (1131/1133)	1.0 (1129/1133)	1.0 (1130/1133)	1.0 (1131/1133)	1.0 (1133/1133)
Leg Break, Sensitivity	0.69 (11/16)	0.62 (10/16)	0.25 (4/16)	0.56 (9/16)	0.56 (9/16)	0.0 (0/16)
Leg Break, Specificity	1.0 (1135/1136)	1.0 (1135/1136)	1.0 (1136/1136)	1.0 (1136/1136)	1.0 (1135/1136)	1.0 (1136/1136)

Supplement Table S2.ii - Twitter/X posts, Comparative Validations of 6 LLMs vs. GPT-4

LLM Characterizations	Sonnet	Mixtral 8x22b	Opus	LlaMa 3	GPT 4o	GPT 3.5
Outbreak (OB) Probability	0.82 (0.76, 0.87)	0.74 (0.67, 0.79)	0.73 (0.63, 0.80)	0.38 (0.11, 0.60)	0.75 (0.65, 0.83)	0.57 (0.48, 0.66)
OB Severity of Cases	0.99 (0.96, 1.0)	1.0 (0.98, 1.0)	1.0 (0.98, 1.0)	1.0 (0.98, 1.0)	1.0 (0.98, 1.0)	0.98 (0.95, 1.0)
OB Size (# of Cases)	0.85 (0.79, 0.90)	0.84 (0.78, 0.88)	0.74 (0.65, 0.81)	0.71 (0.60, 0.79)	0.73 (0.66, 0.80)	0.13 (-0.19, 0.43)
OUTBREAK TYPE						
Infectious, Sensitivity	0.98 (235/241)	0.96 (232/241)	0.95 (228/241)	0.88 (213/241)	0.93 (224/241)	0.71 (171/241)
Infectious, Specificity	0.57 (4/7)	0.29 (2/7)	0.43 (3/7)	0.71 (5/7)	0.57 (4/7)	0.57 (4/7)
Allergic, Sensitivity	(1/1)	(1/1)	(1/1)	(1/1)	(1/1)	(0/1)
Allergic, Specificity	1.0 (246/247)	1.0 (247/247)	1.0 (247/247)	1.0 (247/247)	1.0 (247/247)	1.0 (247/247)
AHC, Sensitivity	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)
AHC, Specificity	1.0 (248/248)	1.0 (248/248)	1.0 (248/248)	1.0 (248/248)	1.0 (248/248)	1.0 (248/248)
Environ., Sensitivity	(3/5)	(1/5)	(1/5)	(1/5)	(1/5)	(0/5)
Environ., Specificity	0.98 (239/243)	1.0 (243/243)	1.0 (243/243)	0.99 (241/243)	1.0 (242/243)	1.0 (243/243)
HEALTH CONDITIONS						
Conjunct., Sensitivity	0.94 (333/354)	0.95 (335/354)	0.94 (332/354)	0.97 (343/354)	0.86 (306/354)	0.57 (203/354)
Conjunct., Specificity	0.44 (7/16)	0.50 (8/16)	0.62 (10/16)	0.44 (7/16)	0.88 (14/16)	0.94 (15/16)
COVID-19, Sensitivity	0.81 (21/26)	0.50 (13/26)	0.77 (20/26)	0.69 (18/26)	0.69 (18/26)	0.50 (13/26)
COVID-19, Specificity	0.98 (338/344)	0.99 (340/344)	0.95 (328/344)	0.99 (339/344)	0.98 (336/344)	0.99 (339/344)
Influenza, Sensitivity	0.68 (13/19)	0.63 (12/19)	0.53 (10/19)	0.58 (11/19)	0.89 (17/19)	0.53 (10/19)
Influenza, Specificity	0.98 (344/351)	0.98 (345/351)	0.98 (345/351)	0.99 (346/351)	0.98 (343/351)	0.99 (346/351)
Int. Flu, Sensitivity	0.89 (16/18)	0.72 (13/18)	0.61 (11/18)	0.61 (11/18)	0.89 (16/18)	0.28 (5/18)
Int. Flu, Specificity	0.99 (348/352)	0.99 (348/352)	0.99 (347/352)	0.99 (348/352)	0.99 (349/352)	0.99 (348/352)
Croup, Sensitivity	(2/2)	(2/2)	(2/2)	(1/2)	(2/2)	(0/2)
Croup, Specificity	1.0 (368/368)	1.0 (368/368)	1.0 (368/368)	1.0 (368/368)	1.0 (368/368)	1.0 (368/368)
Lice, Sensitivity	(1/1)	(0/1)	(0/1)	(0/1)	(0/1)	(0/1)
Lice, Specificity	0.99 (367/369)	0.99 (365/369)	0.99 (366/369)	1.0 (368/369)	1.0 (368/369)	1.0 (369/369)
Leg Break, Sensitivity	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)
Leg Break, Specificity	1.0 (370/370)	1.0 (370/370)	1.0 (370/370)	1.0 (370/370)	1.0 (370/370)	1.0 (370/370)

Supplement Table S2.ii - Forum posts, Comparative Validations of 6 LLMs vs. GPT-4

LLM Characterizations	Sonnet	Mixtral 8x22b	Opus	LlaMa 3	GPT 4o	GPT 3.5
Outbreak (OB) Probability	0.93 (0.90, 0.95)	0.82 (0.77, 0.86)	0.79 (0.71, 0.85)	0.73 (0.59, 0.82)	0.56 (0.44, 0.67)	0.50 (0.39, 0.60)
OB Severity of Cases	1.0 (0.97, 1.0)	0.99 (0.95, 1.0)	1.0 (0.96, 1.0)	1.0 (0.97, 1.0)	0.98 (0.92, 1.0)	0.98 (0.93, 1.0)
OB Size (# of Cases)	0.95 (0.92, 0.97)	0.82 (0.72, 0.89)	0.64 (0.46, 0.77)	0.84 (0.74, 0.90)	0.36 (0.14, 0.55)	0.69 (0.44, 0.84)
OUTBREAK TYPE						
Infectious, Sensitivity	0.99 (100/101)	1.0 (101/101)	0.92 (93/101)	0.92 (93/101)	0.71 (72/101)	0.89 (90/101)

LLM Characterizations	Sonnet	Mixtral 8x22b	Opus	LlaMa 3	GPT 4o	GPT 3.5
Infectious, Specificity	0.88 (7/8)	0.75 (6/8)	0.88 (7/8)	0.62 (5/8)	0.88 (7/8)	0.50 (4/8)
Allergic, Sensitivity	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)
Allergic, Specificity	1.0 (109/109)	1.0 (109/109)	0.99 (108/109)	1.0 (109/109)	0.99 (108/109)	1.0 (109/109)
AHC, Sensitivity	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)
AHC, Specificity	1.0 (109/109)	1.0 (109/109)	1.0 (109/109)	0.99 (108/109)	1.0 (109/109)	1.0 (109/109)
Environ., Sensitivity	0.88 (7/8)	0.75 (6/8)	0.75 (6/8)	0.62 (5/8)	0.88 (7/8)	0.38 (3/8)
Environ., Specificity	1.0 (101/101)	1.0 (101/101)	1.0 (101/101)	0.99 (100/101)	0.97 (98/101)	1.0 (101/101)
HEALTH CONDITIONS						
Conjunct., Sensitivity	0.95 (219/231)	0.91 (210/231)	0.82 (189/231)	0.95 (220/231)	0.80 (184/231)	0.65 (151/231)
Conjunct., Specificity	0.71 (42/59)	0.75 (44/59)	0.90 (53/59)	0.68 (40/59)	0.58 (34/59)	0.86 (51/59)
COVID-19, Sensitivity	0.83 (29/35)	0.66 (23/35)	0.80 (28/35)	0.80 (28/35)	0.43 (15/35)	0.54 (19/35)
COVID-19, Specificity	0.96 (246/255)	0.98 (249/255)	0.95 (243/255)	0.98 (249/255)	0.91 (233/255)	0.99 (253/255)
Influenza, Sensitivity	0.60 (6/10)	0.70 (7/10)	0.60 (6/10)	0.50 (5/10)	0.50 (5/10)	0.30 (3/10)
Influenza, Specificity	0.97 (272/280)	0.99 (277/280)	0.99 (276/280)	0.99 (276/280)	0.99 (277/280)	1.0 (280/280)
Int. Flu, Sensitivity	0.50 (4/8)	0.62 (5/8)	0.75 (6/8)	0.38 (3/8)	0.50 (4/8)	0.12 (1/8)
Int. Flu, Specificity	1.0 (281/282)	0.99 (280/282)	0.98 (276/282)	1.0 (282/282)	0.99 (279/282)	0.99 (279/282)
Croup, Sensitivity	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)
Croup, Specificity	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)
Lice, Sensitivity	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)
Lice, Specificity	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)
Leg Break, Sensitivity	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)	(0/0)
Leg Break, Specificity	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)	1.0 (290/290)

Supplement Table S2.iii Comparing Human Validation to GPT-4 Insights, all 4 Human Graders

LLM Characterizations	NON md Person 1	NON md Person 2	MD1	MD2
Outbreak (OB) Probability	0.59 (0.41, 0.72)	0.62 (0.45, 0.74)	n/a	n/a
OB Severity of Cases	0.73 (0.59, 0.83)	0.78 (0.65, 0.86)	n/a	n/a
OB Size (# of Cases)	0.98 (0.91, 1.0)	0.98 (0.87, 1.0)	0.99 (0.93, 1.0)	1.0 (0.96, 1.0)
OUTBREAK TYPE				
Infectious, Sensitivity	1.0 (57/57)	0.98 (47/48)	0.91 (50/55)	0.84 (66/79)
Infectious, Specificity	0.68 (21/31)	0.50 (20/40)	0.48 (16/33)	0.89 (8/9)
Allergic, Sensitivity	0.93 (13/14)	0.88 (14/16)	0.92 (12/13)	0.89 (8/9)
Allergic, Specificity	0.97 (72/74)	0.99 (71/72)	0.96 (72/75)	0.91 (72/79)
AHC, Sensitivity	0.57 (4/7)	(0/0)	0.57 (4/7)	(0/0)
AHC, Specificity	0.99 (80/81)	0.94 (83/88)	0.99 (80/81)	0.94 (83/88)
Environ., Sensitivity	0.67 (6/9)	(1/3)	(1/1)	(0/0)
Environ., Specificity	0.99 (78/79)	0.93 (79/85)	0.93 (81/87)	0.92 (81/88)
HEALTH CONDITIONS				
Conjunct., Sensitivity	0.88 (15/17)	0.88 (15/17)	n/a	n/a
Conjunct., Specificity	1.0 (33/33)	1.0 (33/33)	n/a	n/a
COVID-19, Sensitivity	0.86 (6/7)	0.86 (6/7)	n/a	n/a
COVID-19, Specificity	1.0 (43/43)	1.0 (43/43)	n/a	n/a
Influenza, Sensitivity	0.62 (5/8)	0.83 (5/6)	n/a	n/a
Influenza, Specificity	1.0 (42/42)	1.0 (44/44)	n/a	n/a
Int. Flu, Sensitivity	(2/2)	(2/2)	n/a	n/a
Int. Flu, Specificity	0.92 (44/48)	0.92 (44/48)	n/a	n/a
Croup, Sensitivity	(5/5)	(5/5)	n/a	n/a
Croup, Specificity	1.0 (45/45)	1.0 (45/45)	n/a	n/a
Lice, Sensitivity	0.88 (7/8)	0.88 (7/8)	n/a	n/a
Lice, Specificity	1.0 (42/42)	1.0 (42/42)	n/a	n/a
Leg Break, Sensitivity	(1/1)	(1/1)	n/a	n/a
Leg Break, Specificity	1.0 (49/49)	1.0 (49/49)	n/a	n/a

Literature Cited

1. Brownstein JS, Mandl KD. Reengineering real time outbreak detection systems for influenza epidemic monitoring. AMIA Annu Symp Proc. 2006:866. PubMed PMID: 17238486; PubMed Central PMCID: PMC1839573.
2. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection--harnessing the Web for public health surveillance. N Engl J Med. 2009;360(21):2153-5, 7. doi: 10.1056/NEJMp0900702. PubMed PMID: 19423867; PubMed Central PMCID: PMC2917042.
3. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 2009;11(1):e11. Epub 20090327. doi: 10.2196/jmir.1157. PubMed PMID: 19329408; PubMed Central PMCID: PMCPMC2762766.
4. Tall AA, Cliff%AWang, Jun%Anull Ed.%D%I. Generating Connected Synthetic Electronic Health Records and Social Media Data for Modeling and Simulation. Medium: X p.
5. Aramaki E, Maskawa S, Morita M. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In: Barzilay R, Johnson M, editors.; July 2011: Association for Computational Linguistics; 2011. p. 1568-76.
6. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med. 2011;40(5 Suppl 2):S154-8. doi: 10.1016/j.amepre.2011.02.006. PubMed PMID: 21521589.
7. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. Am J Trop Med Hyg. 2012;86(1):39-45. doi: 10.4269/ajtmh.2012.11-0597. PubMed PMID: 22232449; PubMed Central PMCID: PMC3247107.
8. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. PLoS Med. 2013;10(4):e1001413. doi: 10.1371/journal.pmed.1001413. PubMed PMID: 23565065; PubMed Central PMCID: PMC3614504.
9. Salathe M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza A (H7N9) and the importance of digital epidemiology. N Engl J Med. 2013;369(5):401-4. doi: 10.1056/NEJMp1307752. PubMed PMID: 23822655.
10. Aleo CL, Hark L, Leiby B, Dai Y, Murchison AP, Martinez P, Haller JA. Ophthalmic patients' utilization of technology and social media: an assessment to improve quality of care. Telemed J E Health. 2014;20(10):887-92. doi: 10.1089/tmj.2013.0365. PubMed PMID: 25188532.
11. Aslam AA, Tsou MH, Spitzberg BH, An L, Gawron JM, Gupta DK, Peddecord KM, Nagel AC, Allen C, Yang JA, Lindsay S. The reliability of tweets as a supplementary method of seasonal influenza surveillance. J Med Internet Res. 2014;16(11):e250. Epub 20141114. doi: 10.2196/jmir.3532. PubMed PMID: 25406040; PubMed Central PMCID: PMCPMC4260066.
12. Child RJ, Mentes JC, Pavlish C, Phillips LR. Using Facebook and participant information clips to recruit emergency nurses for research. Nurse Res. 2014;21(6):16-21. Epub 2014/07/26. doi: 10.7748/nr.21.6.16.e1246. PubMed PMID: 25059083.
13. Finfgeld-Connett D. Twitter and Health Science Research. West J Nurs Res. 2014. doi: 10.1177/0193945914565056. PubMed PMID: 25542190.
14. Grajales FJ, 3rd, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res. 2014;16(2):e13. doi: 10.2196/jmir.2912. PubMed PMID: 24518354; PubMed Central PMCID: PMC3936280.
15. McGregor F, Somner JE, Bourne RR, Munn-Giddings C, Shah P, Cross V. Social media use by patients with glaucoma: what can we learn? Ophthalmic Physiol Opt. 2014;34(1):46-52. doi: 10.1111/opo.12093. PubMed PMID: 24325434.
16. Paul MJ, Dredze M. Discovering health topics in social media using topic models. PLoS One. 2014;9(8):e103408. Epub 2014/08/02. doi: 10.1371/journal.pone.0103408. PubMed PMID: 25084530; PubMed Central PMCID: PMCPMC4118877.

17. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr. 2014;6. doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117. PubMed PMID: 25642377; PubMed Central PMCID: PMC4234396.
18. Pawelek KA, Oeldorf-Hirsch A, Rong L. Modeling the impact of twitter on influenza epidemics. Math Biosci Eng. 2014;11(6):1337-56. doi: 10.3934/mbe.2014.11.1337. PubMed PMID: 25365604.
19. Santos JC, Matos S. Analysing Twitter and web queries for flu trend prediction. Theor Biol Med Model. 2014;11 Suppl 1:S6. doi: 10.1186/1742-4682-11-S1-S6. PubMed PMID: 25077431; PubMed Central PMCID: PMC4108891.
20. Velardi P, Stilo G, Tozzi AE, Gesualdo F. Twitter mining for fine-grained syndromic surveillance. Artif Intell Med. 2014;61(3):153-63. doi: 10.1016/j.artmed.2014.01.002. PubMed PMID: 24613716.
21. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q. 2014;92(1):7-33. doi: 10.1111/1468-0009.12038. PubMed PMID: 24597553; PubMed Central PMCID: PMC3955375.
22. Micieli JA, Tsui E. Ophthalmology on social networking sites: an observational study of Facebook, Twitter, and LinkedIn. Clin Ophthalmol. 2015;9:285-90. doi: 10.2147/OPTH.S79032. PubMed PMID: 25709390; PubMed Central PMCID: PMC4334327.
23. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance Tools Emerging From Search Engines and Social Media Data for Determining Eye Disease Patterns. JAMA Ophthalmol. 2016;134(9):1024-30. Epub 2016/07/15. doi: 10.1001/jamaophthalmol.2016.2267. PubMed PMID: 27416554; PubMed Central PMCID: PMCPMC5227006.
24. Rodriguez-Martinez M. Experiences with the Twitter Health Surveillance (THS) System. Proc IEEE Int Congr Big Data. 2017;2017:376-83. Epub 20170911. doi: 10.1109/BigDataCongress.2017.55. PubMed PMID: 29607412; PubMed Central PMCID: PMCPMC5872152.
25. Chen X, Faviez C, Schuck S, Lillo-Le-Louet A, Texier N, Dahamna B, Huot C, Foulquie P, Pereira S, Leroux V, Karapetiantz P, Guenegou-Arnoux A, Katsahian S, Bousquet C, Burgun A. Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate. Front Pharmacol. 2018;9:541. Epub 20180524. doi: 10.3389/fphar.2018.00541. PubMed PMID: 29881351; PubMed Central PMCID: PMCPMC5978246.
26. Deiner MS, McLeod SD, Chodosh J, Oldenburg CE, Fathy CA, Lietman TM, Porco TC. Clinical Age-Specific Seasonal Conjunctivitis Patterns and Their Online Detection in Twitter, Blog, Forum, and Comment Social Media Posts. Invest Ophthalmol Vis Sci. 2018;59(2):910-20. Epub 2018/02/17. doi: 10.1167/iovs.17-22818. PubMed PMID: 29450538; PubMed Central PMCID: PMCPMC5815847.
27. Hammond AS, Paul MJ, Hobelmann J, Koratana AR, Dredze M, Chisolm MS. Perceived Attitudes About Substance Use in Anonymous Social Media Posts Near College Campuses: Observational Study. JMIR Ment Health. 2018;5(3):e52. Epub 2018/08/04. doi: 10.2196/mental.9903. PubMed PMID: 30072359; PubMed Central PMCID: PMCPMC6096169.
28. Reuter K, Angyan P, Le N, MacLennan A, Cole S, Bluthenthal RN, Lane CJ, El-Khoueiry AB, Buchanan TA. Monitoring Twitter Conversations for Targeted Recruitment in Cancer Trials in Los Angeles County: Protocol for a Mixed-Methods Pilot Study. JMIR Res Protoc. 2018;7(9):e177. Epub 2018/10/03. doi: 10.2196/resprot.9762. PubMed PMID: 30274964; PubMed Central PMCID: PMCPMC6231794.
29. Sarkar U, Le GM, Lyles CR, Ramo D, Linos E, Bibbins-Domingo K. Using Social Media to Target Cancer Prevention in Young Adults: Viewpoint. J Med Internet Res. 2018;20(6):e203. Epub 2018/06/07. doi: 10.2196/jmir.8882. PubMed PMID: 29871850; PubMed Central PMCID: PMCPMC6008512.

30. Deiner MS, McLeod SD, Wong J, Chodosh J, Lietman TM, Porco TC. Google Searches and Detection of Conjunctivitis Epidemics Worldwide. Ophthalmology. 2019;126(9):1219-29. Epub 2019/04/15. doi: 10.1016/j.ophtha.2019.04.008. PubMed PMID: 30981915.
31. Le GM, Radcliffe K, Lyles C, Lyson HC, Wallace B, Sawaya G, Pasick R, Centola D, Sarkar U. Perceptions of cervical cancer prevention on Twitter uncovered by different sampling strategies. PLoS One. 2019;14(2):e0211931. Epub 20190211. doi: 10.1371/journal.pone.0211931. PubMed PMID: 30742683; PubMed Central PMCID: PMCPMC6370210.
32. Lyson HC, Le GM, Zhang J, Rivadeneira N, Lyles C, Radcliffe K, Pasick RJ, Sawaya G, Sarkar U, Centola D. Social Media as a Tool to Promote Health Awareness: Results from an Online Cervical Cancer Prevention Study. J Cancer Educ. 2019;34(4):819-22. doi: 10.1007/s13187-018-1379-8. PubMed PMID: 29948924; PubMed Central PMCID: PMCPMC6289895.
33. Tsui E, Rao RC. Navigating Social Media in #Ophthalmology. Ophthalmology. 2019;126(6):779-82. Epub 2019/05/28. doi: 10.1016/j.ophtha.2019.02.015. PubMed PMID: 31122357.
34. Aiello AE, Renson A, Zivich PN. Social Media- and Internet-Based Disease Surveillance for Public Health. Annu Rev Public Health. 2020;41:101-18. Epub 20200106. doi: 10.1146/annurev-publhealth-040119-094402. PubMed PMID: 31905322; PubMed Central PMCID: PMCPMC7959655.
35. Allen JD, Hollander J, Gualtieri L, Alarcon Falconi TM, Savir S, Agenor M. Feasibility of a twitter campaign to promote HPV vaccine uptake among racially/ethnically diverse young adult women living in public housing. BMC Public Health. 2020;20(1):830. Epub 2020/06/04. doi: 10.1186/s12889-020-08824-0. PubMed PMID: 32487045; PubMed Central PMCID: PMCPMC7268465.
36. Joshi A, Sparks R, Karimi S, Yan SJ, Chughtai AA, Paris C, MacIntyre CR. Automated monitoring of tweets for early detection of the 2014 Ebola epidemic. PLoS One. 2020;15(3):e0230322. Epub 20200317. doi: 10.1371/journal.pone.0230322. PubMed PMID: 32182277; PubMed Central PMCID: PMCPMC7077840.
37. Joshi A, Sparks R, McHugh J, Karimi S, Paris C, MacIntyre CR. Harnessing Tweets for Early Detection of an Acute Disease Event. Epidemiology. 2020;31(1):90-7. doi: 10.1097/ede.0000000000001133. PubMed PMID: 31651659; PubMed Central PMCID: PMCPMC6889901.
38. Ola O, Sedig K. Understanding Discussions of Health Issues on Twitter: A Visual Analytic Study. Online J Public Health Inform. 2020;12(1):e2. Epub 20200516. doi: 10.5210/ojphi.v12i1.10321. PubMed PMID: 32577151; PubMed Central PMCID: PMCPMC7295584.
39. Wolffsohn JS, Leteneux-Pantais C, Chiva-Razavi S, Bentley S, Johnson C, Findley A, Tolley C, Arbuckle R, Kommineni J, Tyagi N. Social Media Listening to Understand the Lived Experience of Presbyopia: Systematic Search and Content Analysis Study. J Med Internet Res. 2020;22(9):e18306. Epub 20200921. doi: 10.2196/18306. PubMed PMID: 32955443; PubMed Central PMCID: PMCPMC7536603.
40. Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PLoS One. 2020;15(9):e0239441. Epub 20200925. doi: 10.1371/journal.pone.0239441. PubMed PMID: 32976519; PubMed Central PMCID: PMCPMC7518625.
41. Karafillakis E, Martin S, Simas C, Olsson K, Takacs J, Dada S, Larson HJ. Methods for Social Media Monitoring Related to Vaccination: Systematic Scoping Review. JMIR Public Health Surveill. 2021;7(2):e17149. Epub 20210208. doi: 10.2196/17149. PubMed PMID: 33555267; PubMed Central PMCID: PMCPMC7899807.
42. Lossio-Ventura JA, Gonzales S, Morzan J, Alatrista-Salas H, Hernandez-Boussard T, Bian J. Evaluation of clustering and topic modeling methods over health-related tweets and emails. Artif

Intell Med. 2021;117:102096. Epub 20210507. doi: 10.1016/j.artmed.2021.102096. PubMed PMID: 34127235; PubMed Central PMCID: PMCPMC9040385.
43. Syntosi A, Felizzi F, Bouchet C. A Social Media Listening Study to Understand the Unmet Needs and Quality of Life in Adult and Pediatric Amblyopia Patients. Ophthalmol Ther. 2022;11(6):2183-96. Epub 20220929. doi: 10.1007/s40123-022-00571-2. PubMed PMID: 36175822; PubMed Central PMCID: PMCPMC9587203.
44. Zhang Y, Chen K, Weng Y, Chen Z, Zhang J, Hubbard R. An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US. Expert Syst Appl. 2022;198:116882. Epub 20220314. doi: 10.1016/j.eswa.2022.116882. PubMed PMID: 35308584; PubMed Central PMCID: PMCPMC8920081.
45. Gao W, Li L, Tao X, Zhou J, Tao J. Identifying informative tweets during a pandemic via a topic-aware neural language model. World Wide Web. 2023;26(1):55-70. Epub 20220316. doi: 10.1007/s11280-022-01034-1. PubMed PMID: 35308294; PubMed Central PMCID: PMCPMC8924578.
46. McMann T, Wenzel C, Le N, Li Z, Xu Q, Cuomo RE, Mackey T. Detection and Characterization of Web-Based Pediatric COVID-19 Vaccine Discussions and Racial and Ethnic Minority Topics: Retrospective Analysis of Twitter Data. JMIR Pediatr Parent. 2023;6:e48004. Epub 20231130. doi: 10.2196/48004. PubMed PMID: 38038663; PubMed Central PMCID: PMCPMC10706352.
47. Pradeep T, Ravipati A, Melachuri S, Fu R. More than just a stye: identifying seasonal patterns using google trends, and a review of infodemiological literature in ophthalmology. Orbit. 2023;42(2):130-7. Epub 20220303. doi: 10.1080/01676830.2022.2040542. PubMed PMID: 35240907.
48. Wang Y, Chen Y. Characterizing discourses about COVID-19 vaccines on Twitter: a topic modeling and sentiment analysis approach. J Commun Healthc. 2023;16(1):103-12. Epub 20220324. doi: 10.1080/17538068.2022.2054196. PubMed PMID: 36919802.
49. Ashraf AR, Mackey TK, Fittler A. Search Engines and Generative Artificial Intelligence Integration: Public Health Risks and Recommendations to Safeguard Consumers Online. JMIR Public Health Surveill. 2024;10:e53086. Epub 20240321. doi: 10.2196/53086. PubMed PMID: 38512343; PubMed Central PMCID: PMCPMC10995787.
50. Deiner MS, Deiner NA, Hristidis V, McLeod SD, Doan T, Lietman TM, Porco TC. Use of Large Language Models to Assess the Likelihood of Epidemics From the Content of Tweets: Infodemiology Study. J Med Internet Res. 2024;26:e49139. Epub 20240301. doi: 10.2196/49139. PubMed PMID: 38427404; PubMed Central PMCID: PMCPMC10943433.
51. Deiner MS, Honcharov V, Li J, Mackey TK, Porco TC, Sarkar U. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study JMIR Infodemiology. 2024;01/07/2024:59641 ((forthcoming/in press)). doi: 10.2196/59641.
52. Elyashar A, Plochotnikov I, Cohen IC, Puzis R, Cohen O. The State of Mind of Health Care Professionals in Light of the COVID-19 Pandemic: Text Analysis Study of Twitter Discourses. J Med Internet Res. 2021;23(10):e30217. Epub 20211022. doi: 10.2196/30217. PubMed PMID: 34550899; PubMed Central PMCID: PMCPMC8544741.
53. Deiner MS, Fathy C, Kim J, Niemeyer K, Ramirez D, Ackley SF, Liu F, Lietman TM, Porco TC. Facebook and Twitter vaccine sentiment in response to measles outbreaks. Health Informatics J. 2019;25(3):1116-32. Epub 2017/11/18. doi: 10.1177/1460458217740723. PubMed PMID: 29148313; PubMed Central PMCID: PMCPMC5930144.
54. Elyoseph Z, Levkovich I. Comparing the Perspectives of Generative AI, Mental Health Experts, and the General Public on Schizophrenia Recovery: Case Vignette Study. JMIR Ment Health. 2024;11:e53043. Epub 20240318. doi: 10.2196/53043. PubMed PMID: 38533615; PubMed Central PMCID: PMCPMC11004608.
55. de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KG, van Smeden M.

From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Health. 2024;6(7):e441-e3. doi: 10.1016/S2589-7500(24)00111-0. PubMed PMID: 38906607.
56. Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, Linos E, Pereira F. A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data. JMIR Ment Health. 2024;11:e50150. Epub 20240125. doi: 10.2196/50150. PubMed PMID: 38271138; PubMed Central PMCID: PMCPMC10813836.
57. Filleul L, Pages F, Wan GC, Brottet E, Vilain P. Costs of Conjunctivitis Outbreak, Reunion Island, France. Emerg Infect Dis. 2018;24(1):168-70. Epub 2017/12/21. doi: 10.3201/eid2401.170916. PubMed PMID: 29260662; PubMed Central PMCID: PMCPMC5749441.
58. Yen MY, Wu TS, Chiu AW, Wong WW, Wang PE, Chan TC, King CC. Taipei's use of a multi-channel mass risk communication program to rapidly reverse an epidemic of highly communicable disease. PLoS One. 2009;4(11):e7962. Epub 2009/12/04. doi: 10.1371/journal.pone.0007962. PubMed PMID: 19956722; PubMed Central PMCID: PMCPMC2776508.
59. Prajna NV, Lalitha P, Teja GV, Gunasekaran R, Sharma SS, Hinterwirth A, Ruder K, Zhong L, Chen C, Deiner M, Huang C, Pinsky BA, Lietman TM, Seitzman GD, Doan T, Group SS. Outpatient human coronavirus associated conjunctivitis in India. J Clin Virol. 2022;157:105300. Epub 20220927. doi: 10.1016/j.jcv.2022.105300. PubMed PMID: 36209621; PubMed Central PMCID: PMCPMC9512524.
60. Lalitha P, Prajna NV, Gunasekaran R, Teja GV, Sharma SS, Hinterwirth A, Ruder K, Zhong L, Chen C, Deiner M, Huang C, Pinsky BA, Lietman TM, Doan T, Seitzman GD, Group SS. Deep sequencing analysis of clinical samples from patients with acute infectious conjunctivitis during the COVID-19 delta surge in Madurai, India. J Clin Virol. 2022;157:105318. Epub 20221005. doi: 10.1016/j.jcv.2022.105318. PubMed PMID: 36242841; PubMed Central PMCID: PMCPMC9534536.
61. County PHdoLA. With New COVID-19 Strain Confirmed in Los Angeles County, Residents Advised to Be Aware of Symptoms, Take Precautions http://www.publichealth.lacounty.gov/phcommon/public/media/mediapubhpdetail.cfm?prid=4372: Public Health department of Los Angeles County 2023.
62. Vashishtha VM, Kumar P. Preliminary clinical characteristics of Pediatric Covid-19 cases during the ongoing Omicron XBB.1.16 driven surge in a north Indian city. medRxiv. 2023:2023.04.18.23288715. doi: 10.1101/2023.04.18.23288715.
63. Centers for Disease Control and Prevention H-AN. Outbreak of Extensively Drug-resistant Pseudomonas aeruginosa Associated with Artificial Tears. https://www.cdc.gov/hai/outbreaks/crpa-artificial-tears.html: Centers for Disease Control and Prevemntion, Healthcare-Associated Notifications Notifications H-A; 2023 05/15/2023. Report No.
64. Hampton A, Bhatt S, Smith A, Brunn J, Purohit H, Shalin VL, Flach JM, Sheth AP. On Using Synthetic Social Media Stimuli in an Emergency Preparedness Functional Exercise2015 March 01, 2015:[arXiv:1503.00760 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2015arXiv150300760H.
65. E. Sagduyu Y, Grushin A, Shi Y. Synthetic Social Media Data Generation. IEEE Transactions on Computational Social Systems. 2018;5:605-20.
66. Li Z, Zhu H, Lu Z, Yin M. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations2023 October 01, 2023:[arXiv:2310.07849 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2023arXiv231007849L.
67. Liu R, Wei J, Liu F, Si C, Zhang Y, Rao J, Zheng S, Peng D, Yang D, Zhou D, Dai AM. Best Practices and Lessons Learned on Synthetic Data for Language Models2024 April 01, 2024:[arXiv:2404.07503 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv240407503L.
68. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer K, Succi MD.

Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J Med Internet Res. 2023;25:e48659. Epub 20230822. doi: 10.2196/48659. PubMed PMID: 37606976; PubMed Central PMCID: PMCPMC10481210.
69. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models2022 January 01, 2022:[arXiv:2201.11903 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2022arXiv220111903W.
70. Yang H, Li M, Zhou H, Xiao Y, Fang Q, Zhang R. One LLM is not Enough: Harnessing the Power of Ensemble Learning for Medical Question Answering. medRxiv. 2023. Epub 20231224. doi: 10.1101/2023.12.21.23300380. PubMed PMID: 38196648; PubMed Central PMCID: PMCPMC10775333.
71. Jiang D, Ren X, Yuchen Lin B. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion2023 June 01, 2023:[arXiv:2306.02561 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2023arXiv230602561J.
72. Sher DV. A Gentle Intro to Chaining LLMs, Agents, and utils via LangChain: Towards Data Science; 2023. Available from: https://towardsdatascience.com/a-gentle-intro-to-chaining-llms-agents-and-utils-via-langchain-16cd385fca81.
73. Popper B, Donovan R, Geden M. How do you evaluate an LLM? Try an LLM. Stack Overflow; 2023.
74. Babenko K, Pankova M, Prashchur H. Validating LLM using LLM2024. Available from: https://www.processica.com/articles/validating-llm-using-llm/.
75. Zheng M, Pei J, Jurgens D. Is “A Helpful Assistant” the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts. arXiv, 2023.
76. Kotu V, Deshpande B. Chapter 4 - Classification. In: Vijay K, Bala D, editors. Data Science (Second Edition): Morgan Kaufmann; 2019. p. 65-163.
77. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding2018 October 01, 2018:[arXiv:1810.04805 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D.
78. Azari AA, Barney NP. Conjunctivitis: a systematic review of diagnosis and treatment. JAMA. 2013;310(16):1721-9. doi: 10.1001/jama.2013.280318. PubMed PMID: 24150468; PubMed Central PMCID: PMC4049531.
79. Lalitha P, Seitzman GD, Kotecha R, Hinterwirth A, Chen C, Zhong L, Cummings S, Lebas E, Sahoo MK, Pinsky BA, Lietman TM, Doan T. Unbiased Pathogen Detection and Host Gene Profiling for Conjunctivitis. Ophthalmology. 2019;126(8):1090-4. Epub 20190404. doi: 10.1016/j.ophtha.2019.03.039. PubMed PMID: 30953744; PubMed Central PMCID: PMCPMC6646074.
80. Bountogo M, Sie A, Coulibaly B, Ruder K, Chen C, Zhong L, Colby E, Lebas E, Deiner M, Hinterwirth A, Lietman TM, Seitzman GD, Doan T. Deep sequencing analysis of acute conjunctivitis in Burkina Faso, Africa. Int Health. 2023;15(1):101-3. doi: 10.1093/inthealth/ihac001. PubMed PMID: 35076074; PubMed Central PMCID: PMCPMC9808514.
81. Johnson K, Taleo F, Willie K, Amel E, Nanu M, Alguet M, Wass J, Rymill P, Solomon A, Ruder K, Chen C, Zhong L, Hinterwirth A, Liu D, Abraham T, Seitzman G, Lietman T, Doan T, Group SS. Viruses Associated with Acute Conjunctivitis in Vanuatu. Am J Trop Med Hyg. 2023;108(3):595-8. Epub 20230116. doi: 10.4269/ajtmh.22-0600. PubMed PMID: 36646076; PubMed Central PMCID: PMCPMC9978563.
82. Tran H, Hoang L, Tran Y, Ha T, Ruder K, Zhong L, Chen C, Yu D, Liu Y, Abraham T, Hinterwirth A, Deiner M, Porco T, Lietman TM, Seitzman GD, Doan T. Pathogen Profiles of Infectious Conjunctivitis in Ho Chi Minh City, Vietnam. Cornea Open. 2023;2(4). Epub 20231207. doi: 10.1097/coa.0000000000000025. PubMed PMID: 38855500; PubMed Central PMCID: PMCPMC11160933.
83. Tran H, Ha T, Hoang L, Tran Y, Ruder K, Zhong L, Chen C, Liu Y, Yu D, Abraham T,

Hinterwirth A, Deiner M, Porco TC, Lietman TM, Doan T, Seitzman GD, Group SS. Coxsackievirus A24 Causing Acute Conjunctivitis in a 2023 outbreak in Vietnam. Int J Infect Dis. 2024:107133. Epub 20240612. doi: 10.1016/j.ijid.2024.107133. PubMed PMID: 38876162.