Large Language Models at Population Scale: A Survey and Taxonomy of Public Health Applications

Zhengxu Tang[1*], **Bohan Wang**[1*], **Yiming Lu**[1*], **Zewen Liu**[1], **Max S. Y. Lau**[2], **and Wei Jin**[1]

[1]Department of Computer Science, Emory University

[2]Department of Biostatistics and Bioinformatics, Emory University

{yiming.lu, bohan.wang2, wei.jin}@emory.edu

Abstract

Large language models (LLMs) have been widely studied for individual level clinical tasks, yet their role in public health remains far less systematically understood. Unlike clinical AI, public health operates primarily at the population level, introducing distinct challenges in data modality, reasoning scope, and societal-scale risk. We present the first comprehensive survey of LLM in public health through a two-dimensional taxonomy, which maps six core public health tasks against five functional LLM roles. Our analysis reveals a critical gap: while current LLMs excel at extracting health signals from text, they struggle with the complex reasoning required for population-level challenges. Ultimately, applying LLMs to public health is not simply about scaling up clinical models; it introduces fundamentally new risks, from the mass generation of misinformation to the amplification of systemic inequities. To address these challenges, we argue that LLMs must evolve from individual-level clinical assistants into societal-scale reasoning engines capable of modeling disease dynamics, simulating policy interventions, and ensuring health equity on a global scale.

1 Introduction

Public health is concerned with protecting and improving the health of entire populations, and it shapes decisions with far-reaching societal consequences (Hattab et al., 2025; Acosta, 2025). Authorities must decide when to declare an outbreak, how to allocate scarce vaccines, and whether a new intervention will mitigate or exacerbate health disparities. These decisions are often made in real time, under substantial uncertainty, and across heterogeneous populations with different risks, resources, and needs. As global health threats become more frequent and complex, from pandemic preparedness to climate-driven disease emergence, the need for scalable, data-driven tools that can support population-level decision-making has become increasingly urgent (Dixit, 2025; Rizzo et al., 2024; Belova et al., 2025; Liu et al., 2024).

Recent advances position large language models (LLMs) as promising public health tools (De Angelis et al., 2023; Grattafiori et al., 2024; Guo et al., 2025; Xu et al., 2026). These foundation models efficiently process unstructured text, extract signals from noisy data, and generate actionable summaries (Bommasani et al., 2021). These capabilities naturally support core functions, including social media-based outbreak detection, cross-lingual evidence synthesis, tailored health communication, and equity-focused policy auditing.

However, the current evidence base for LLMs in health is centered primarily on clinical medicine rather than public health. A growing ecosystem of specialized foundation models, clinical benchmarks, and systematic reviews has documented LLM applications in diagnosis, treatment recommendation, medical documentation, and biomedical text mining. More

Task / Section	N	Dominant LLM Role(s)	Typical Data Modality
T1 Surveillance	53	Sensor (n=37), Predictor (n=7)	News / outbreak reports; social media / reviews; EHR / clinical text; structured surveillance records
T2 Forecasting	18	Predictor (n=10), Simulator (n=5)	Epidemic time series; simulated epidemic trajectories; policy / genomic covariates
T3 Infodemiology	52	Sensor (n=28), Communicator (n=10), Auditor (n=9)	Social media; news / fact-check content; surveys; synthetic / multimodal misinformation data
T4 Health Equity	22	Auditor (n=8), Sensor (n=6)	EHR / clinical notes; surveys / benchmarks; simulated prompts; literature / policy corpora
T5 Intervention	18	Communicator (n=8), Sensor (n=4), Simulator (n=3)	Surveys / chatbot interactions; policy / guideline documents; social media; EHR / mobile-health data
T6 Governance	25	Simulator (n=8), Auditor (n=7)	Simulated epidemic environments; policy documents; benchmark / literature corpora; public health QA / evaluation data

Table 1: Section-level summary of the surveyed literature, organized by public health task, dominant LLM role, and typical data modality. The number of papers (n) assigned to the leading roles is shown in parentheses.

recently, specialized reviews have examined narrower health sub-domains, including social determinants of health extraction (Keloth et al., 2025), mental health screening (Xu et al., 2024b), and pharmacovigilance (Hakim et al., 2025). Yet these efforts remain isolated: each addresses a single task at the individual level without connecting it to the broader public health pipeline or examining how LLM roles shift across surveillance, forecasting, communication, and governance. No existing work provides a comprehensive, cross-functional synthesis of LLMs across the entire public health spectrum.

Unlike clinical medicine, which focuses primarily on individual patients, public health addresses questions such as how diseases spread through communities, how to counteract vaccine misinformation at scale, and whether interventions exacerbate systemic inequities. These differences create three structural challenges for LLMs. **(1) Data heterogeneity:** public health requires reasoning over heterogeneous, population-scale signals, ranging from social media streams and epidemiological time series to verbal autopsy narratives and wastewater metagenomics (Deiner et al., 2024; Liu et al., 2025b), with no clear parallel in clinical AI. **(2) Cascading failure modes:** whereas a clinical AI hallucination may affect a single patient, a public health AI error, such as a false epidemic alert, culturally insensitive messaging, or algorithmically amplified misinformation, can cascade across entire populations. **(3) Equity as a first-class objective:** public health models must perform equitably across languages, cultures, and healthcare systems, rather than treating bias mitigation as an afterthought.

To bridge this gap, we present the first comprehensive survey of LLMs across diverse public health applications, synthesizing 192 papers published from 2023 to early 2026. To ensure a strict population-health focus, we employed a three-step boundary test (institutional attribution, core function, and substitution; Appendix A.2) to exclude individual clinical diagnostics and drug discovery. The resulting corpus is then organized through a two-dimensional taxonomy(Table 1). The **task dimension** follows the operational pipeline of public health systems, *Sense* $\rightarrow$ *Predict* $\rightarrow$ *Communicate* $\rightarrow$ *Protect* $\rightarrow$ *Intervene* $\rightarrow$ *Govern*, yielding six categories (T1–T6). The **role dimension** captures how an LLM functions within each task: as a Sensor, Predictor, Communicator, Simulator, or Auditor. Our main contributions are:

- We conduct the first cross-functional survey of LLMs in public health, covering the full pipeline from surveillance to governance with a population-health boundary test.
- We introduce a task $\times$ role taxonomy that reveals systematic gaps: the field is heavily concentrated in extraction-oriented sensing (Sensor dominates T1 and T3, which together account for over half the corpus), while anticipatory roles (Predictor, Simulator) and equity-focused auditing remain underexplored.
- We quantify structural imbalances, including that 80.7% of papers use English-only data and most depend on proprietary models, and distill four gap-driven research directions grounded in the taxonomy.

Feature	Public Health	Clinical Medicine / Healthcare
Target	Population, community, region, or jurisdiction (Chen et al., 2025b; Burstein et al., 2025)	Individual patient
Primary goal	Prevention, surveillance, preparedness, equity (Quigley et al., 2025; Singh et al., 2023)	Diagnosis, treatment, recovery, continuity of care
Typical data	Surveillance counts, news reports, policies, environmental and social signals, mobility (Deiner et al., 2024; Wang et al., 2025b)	EHRs, labs, imaging, medications, clinical notes
Typical actions	Vaccination, screening, outbreak response, risk communication, resource allocation (Song et al., 2025; Verma et al., 2025)	Prescription, procedure, triage, monitoring, care planning
Key stakeholders	Health departments, CDC/WHO, policymakers, communities (van Hoek et al., 2024; Harris et al., 2025)	Clinicians, hospitals, care teams, patients

Table 2: Public health versus clinical medicine. Public health targets population-level data and systemic goals, while clinical medicine focuses on individual patient care.

2 Background & preliminaries

Public health as population-level AI. Unlike clinical medicine's focus on individual patients, public health targets populations, communities, and jurisdictions. Its core actions are therefore collective rather than individual: surveillance, prevention, preparedness, risk communication, screening, vaccination, and resource allocation. This population-level scope expands both the data regime and the evaluation target. Public health systems must reason over social media, policy documents, news, surveillance counts, mobility, environmental signals, scientific evidence, and clinical records, often under reporting delays and distribution shift. Success is not accuracy alone: models must be timely, robust, useful for intervention planning, and sensitive to disparities across populations. These demands shape the surveyed literature across epidemic forecasting (Du et al., 2025; Gong et al., 2025; Liu et al., 2025f), misinformation (Song et al., 2025), pharmacovigilance (Sidorov et al., 2025; Carpenter & Altman, 2023), health equity (Pierson et al., 2025; Pfohl et al., 2024), and frontline support (Al Ghadban et al., 2023; Ramjee et al., 2025).

From public health functions to survey tasks. The task dimension in this survey follows the operational pipeline of public health systems. Surveillance and early warning motivate *T1 Population Health Surveillance*; predictive preparedness motivates *T2 Epidemic Forecasting & Modeling*; health promotion and information management motivate *T3 Infodemiology & Health Communication*; population assessment and disparities motivate *T4 Health Equity, SDOH & Global Health*; health protection and prevention motivate *T5 Population Intervention & Practice*; and preparedness, coordination, and responsible deployment motivate *T6 Governance, Ethics & Policy*. This framing explains why LLMs in public health are used not only for medical question answering, but also for event extraction, epidemic intelligence, misinformation analysis, evidence review, intervention design, and policy support. Table 2 summarizes the contrast with clinical medicine.

LLM adaptation strategies in public health. Large language models (LLMs) are foundation models trained on large text corpora to predict and generate language (Vaswani et al., 2017; Bommasani et al., 2021). Rather than review architecture in detail, we focus on the adaptation patterns that recur across public health tasks; additional technical background is provided in Appendix A.1. *Prompting* enables rapid classification, extraction, and communication without parameter updates, and is common in misinformation detection and public health discourse analysis (Brown et al., 2020; Wei et al., 2022). *Fine-tuning*, including parameter-efficient methods such as LoRA (Hu et al., 2022), is used when tasks require domain-specific labels or numerical adaptation, such as influenza forecasting, SDOH extraction, or adverse-event mining. *Retrieval-augmented generation* grounds outputs in external evidence and is especially important for fact-checking, guideline interpretation, and policy support (Lewis et al., 2020). Finally, *agentic systems* connect LLMs to tools, simulators, databases, or multi-agent workflows, enabling applications such as epidemic simulator construction and emergency response planning (Datta et al., 2026; Aoki & Ghaffarzadegan, 2026; Shi et al., 2026).

Why public health poses unique challenges for LLMs. Public health introduces challenges that go beyond individual-level clinical AI. First, many tasks require spatiotemporal and mechanistic reasoning: disease spread depends on contact networks, geography, seasonality, reporting delays, interventions, and changing pathogen dynamics, whereas current LLMs have no native representation of transmission parameters or compartmental models (Du et al., 2025; Gong et al., 2025; Liu et al., 2025f). Second, public health operates inside an information ecosystem. The same generative capability that supports misinformation detection, fact-checking, and counterspeech can also generate persuasive health misinformation at scale, creating a dual-role paradox with no direct clinical analog (De Angelis et al., 2023; Hussain et al., 2025; Modi et al., 2025). Third, public health treats equity as a first-class objective rather than a secondary bias metric: models must work across languages, cultures, infrastructures, and marginalized communities, while current evidence remains heavily English-centered and high-resource (Zhou et al., 2025b; Chen et al., 2025b; Pfohl et al., 2024; Ji et al., 2025). These challenges motivate the task-specific taxonomy and recur throughout the survey.

3 LLM role dimension

We define LLM roles by the model's primary functional output in a public health workflow rather than by architecture, training strategy, or data modality. This distinction matters because the same decoder-only model, prompting method, or social-media dataset can support very different public health functions: extracting a signal, forecasting a trend, generating a message, auditing a system, or simulating an intervention. The role dimension therefore asks what the model produces, who consumes the output, and what downstream public health action it supports.

The five roles correspond to recurring questions in population health workflows: **Sensor** asks what is happening, **Predictor** asks what will happen next, **Simulator** asks what could happen under alternative conditions, **Auditor** asks whether a system is reliable, fair, and safe, and **Communicator** asks what should be conveyed to people. These roles are not strictly sequential, but together they span the core stages of public health AI from evidence extraction to anticipation, planning, evaluation, and response. This role-based view also previews a structural imbalance in the literature: current work remains strongest in retrospective sensing, while forecasting, simulation, and deployment-oriented evaluation are still emerging.

Sensor. Sensor models convert heterogeneous public health data into structured labels, entities, or signals for downstream analysis. The output is usually consumed by analysts, surveillance systems, or statistical models rather than directly by the public. Examples include event extraction from outbreak reports (Consoli et al., 2024; Parekh et al., 2024; Bhadelia et al., 2026), adverse-event extraction in pharmacovigilance (Li et al., 2024), verbal-autopsy and mortality coding (Wen et al., 2026; Coutinho et al., 2026), and SDOH extraction from clinical text (Keloth et al., 2025; Gu et al., 2025). Across these settings, LLMs function as an evidence-extraction layer that transforms noisy population-level inputs into structured public health intelligence.

Predictor. Predictor models produce forecasts of population-health trajectories, risks, or future states. This role is concentrated in epidemic forecasting and selected surveillance settings. In T2, LLM-based and foundation-model forecasters integrate epidemic time series with policy, mobility, genomic, or spatial context for infectious disease prediction (Du et al., 2025; Gong et al., 2025; Moon et al., 2025; Li et al., 2025a; Kalahasti et al., 2025; Panja et al., 2025; Liu et al., 2025f; Dudley et al., 2025). Outside classical forecasting, related predictive uses appear in genomic surveillance, such as SARS-CoV-2 variant fitness prediction (Ito et al., 2025). The key distinction from Sensor is temporal: Predictor systems estimate what is likely to happen next rather than structuring evidence about what has already happened.

Simulator. Simulator models generate or control synthetic epidemic dynamics, policy scenarios, or agent behavior in closed-loop environments. This role appears in work on generative epidemic agents (Williams et al., 2023), simulator construction from natural-language specifications (Datta et al., 2026), policymaking agents in structured epidemic settings (Aoki

& Ghaffarzadegan, 2026), and coordinated intervention planning across interacting regions (Shi et al., 2026). The key distinction from Predictor is counterfactual: Simulator systems explore possible trajectories and decision sequences under alternative assumptions, making them especially relevant for stress testing and planning under uncertainty.

Auditor. Auditor models evaluate whether AI systems, public health decisions, or generated outputs are reliable, fair, and deployment-ready. Unlike the other roles, the auditor's object is often the model or system itself rather than a public health phenomenon. Applications include equity toolkits and bias mitigation (Pfohl et al., 2024; Ji et al., 2025), anti-LGBTQIA+ and mental-health bias audits (Chang et al., 2025; Haider et al., 2025), adversarial disinformation-risk analyses (Hussain et al., 2025; Modi et al., 2025), and epidemiological QA or policy benchmarking (Wei et al., 2026; Harris et al., 2025; Espinosa et al., 2025). In T6, auditing encompasses institutional risks, role creep, and geographic bias (Zhou et al., 2025a; Chu et al., 2025; Wilson et al., 2024). Auditors therefore serve as a primary mechanism for measuring system bias, robustness, and safety.

Communicator. Communicator models translate public health knowledge into human-facing dialogue, guidance, or decision support for communities, frontline workers, patients, and policymakers. The boundary with Sensor is the output audience: Communicator outputs are intended to be read and acted on by people, not merely processed by downstream analytic systems. Central to T3 and T5, they generate evidence-grounded counterspeech, corrective notes, and vaccine messaging (Song et al., 2025; Anik et al., 2025; Wu et al., 2025a; Hou et al., 2025). In intervention settings, they power chatbots and assistants for HIV prevention, smoking cessation, and frontline support (Narayan et al., 2026; Humphries et al., 2026; Govathson et al., 2026; Ramjee et al., 2025; Abroms et al., 2025a). While LLMs increasingly serve as public- and worker-facing interfaces, fluency alone is insufficient; communicators must remain rigorously evidence-grounded, audience-aware, and operationally safe (Liu et al., 2025d; Abroms et al., 2025b).

4 A taxonomy of LLMs in public health

Public health tasks progress from sensing threats to governing systems. Accordingly, the **Task dimension** (Table 1) categorizes the literature into six areas: T1 Population Health Surveillance (53), T2 Epidemic Forecasting & Modeling (18), T3 Infodemiology & Health Communication (52), T4 Health Equity, SDOH & Global Health (22), T5 Population Intervention & Practice (18), and T6 Governance, Ethics & Policy (25). **T1** and **T3** comprise over half the corpus, skewing research toward signal extraction and discourse monitoring, whereas **T4** and **T6** tackle systemic fairness and deployment infrastructure. Across T1–T6, we identify shared methodologies, evaluation gaps, and a shift in LLM roles from early-stage sensing to later-stage evaluation, communication, and simulation.

4.1 Population health surveillance (T1)

This section reviews LLMs for population health surveillance, that is, the detection and monitoring of health threats from heterogeneous population-level data. Compared with standard clinical diagnostics, surveillance must operate over much broader signals, ranging from social media and informal news to verbal autopsy narratives and wastewater metagenomics, and must do so under conditions of incomplete reporting, multilingual coverage, and rapidly evolving threats. The defining computational challenge is therefore not a single prediction problem, but the conversion of noisy, unstructured, and cross-modal inputs into reliable, timely, and structured signals that can drive public health response.

Event and syndromic surveillance. The first and largest strand treats surveillance as an information-extraction problem over public-facing text, where LLMs are used to convert news articles, official outbreak bulletins, and social-media posts into structured event records. For early outbreak detection, few-shot prompting with frontier models and domain-adapted ensembles outperform rule-based baselines on WHO Disease Outbreak News and global news streams, with several systems framing extraction as multilingual or schema-driven event extraction (Consoli et al., 2024; 2025; Parekh et al., 2024; Bhadelia et al., 2026;

Hong et al., 2023). Social media platforms serve as complementary real-time biosensors: fine-tuned systems have predicted COVID-19 surges 7.63 days ahead of official reports (Xie et al., 2025) and generated influenza alerts in Wales weeks before primary care consultations (Sheridan et al., 2025), while PH-LLM scales infoveillance across dozens of multilingual classification and extraction tasks (Zhou et al., 2025b) and multilingual symptom detection extends surveillance coverage into low-resource languages (Jannah et al., 2025). Beyond infectious diseases, the same extraction-centric pattern has been adapted to injury and behavioral surveillance, from agricultural injury narratives (Muller et al., 2025) to child maltreatment case reports analyzed with local, privacy-preserving models (Stoll et al., 2025) and large-scale Reddit analysis of opioid-use-disorder discussions (Testagrose et al., 2025). A notable recent trend is the explicit push toward deployment-ready architectures, including Health Sentinel for national-scale monitoring (Pant et al., 2025) and multi-agent swarms designed for autonomous, continuous intelligence gathering (Wattamwar & Akwafuo, 2026), signaling a shift from isolated extraction studies toward operational early-warning pipelines.

Clinical and mortality surveillance. A second strand moves inside the health system itself, extracting infection and mortality signals from electronic health records, emergency department notes, and civil registration narratives. On the infection side, zero-shot GPT-4 achieves strong symptom identification across multi-site emergency department notes (McMurry et al., 2025), and combining LLM-based case extraction with spatiotemporal scan statistics detects localized outbreaks earlier than case-count-only baselines (Goncalves et al., 2025); complementary studies show that integrating LLMs with clinical expertise can support detection of catheter-associated and broader hospital-acquired infections (Perret & Schmid, 2024; Pan et al., 2025), while in-hospital deployments are beginning to move from pilot to routine infection surveillance (Wu et al., 2025b). On the mortality side, automating verbal autopsy is a flagship application in settings where civil registration is weak: recent work shows that LLMs substantially outperform traditional probabilistic algorithms for cause-of-death assignment in sub-Saharan Africa, both in general adult mortality (Wen et al., 2026; Carshon-Marsh et al., 2026) and in dedicated systems such as LAVA for language-model-assisted verbal autopsy (Chen et al., 2025c). Other work targets structured mortality coding and high-stakes injury data, including constrained decoding for ICD-10 assignment (Coutinho et al., 2026), supervised classification of violent-death narratives with compact LLMs (Parker, 2025), and improved extraction of overdose-related drugs from death investigation reports (Funnell et al., 2026; Omaki et al., 2025). Across this strand, LLMs are increasingly positioned as a way to convert legacy free-text mortality and clinical archives into machine-readable surveillance data, especially for diseases and populations that current structured systems fail to capture well.

Biological and safety surveillance. A third strand extends surveillance from text and records into biological and safety data streams. Sequence language models track pathogen evolution at population scale: CoVFit predicts SARS-CoV-2 variant fitness (Ito et al., 2025), ARNLE links autoregressive next-token prediction to variant prevalence (Liu et al., 2025e), and METAGENE-1 demonstrates that a metagenomic foundation model can detect pathogens from wastewater without species-specific references (Liu et al., 2025b). Beyond viruses, related models have been applied to predicting antibiotic resistance from genomic sequences (Kim & Kim, 2025) and to broader biological-sequence analysis for infectious-disease research (Kaur & Butt, 2025). For post-market drug and vaccine safety, a parallel literature uses fine-tuned LLMs to extract adverse events and vaccine-reaction relationships from clinical text and triage notes, outperforming zero-shot prompting and prior NLP baselines (Li et al., 2024; 2025c; Khademi et al., 2024; 2025; Daluwatte et al., 2024), while LLMs are also being evaluated for upstream pharmacovigilance tasks such as literature screening and structured-query generation (Laurence et al., 2025; Painter et al., 2025). More complex reasoning tasks, however, remain challenging: off-the-shelf LLMs struggle on AEFI causality assessment and case-level adjudication (Abate et al., 2025). This strand also places stronger emphasis on safeguards, including semantic guardrails for multilingual pharmacovigilance (Hakim et al., 2025) and audits of sociodemographic bias and prompt-induced suggestibility in drug-safety decisions (Liu et al., 2025c).

Across these three strands, T1 is dominated by the *Sensor* role, with LLMs primarily used to convert heterogeneous, unstructured inputs into structured surveillance signals. Adaptation strategies cluster along a clear axis: prompting and few-shot use of frontier models are typically sufficient for broad extraction over public text and outbreak news, whereas fine-tuning of open or compact models becomes critical for highly structured, safety-critical, or privacy-sensitive tasks such as pharmacovigilance, mortality coding, and child-protection classification. Relative to traditional rule-based or supervised pipelines, LLM-based surveillance systems consistently improve flexibility, cross-source and cross-lingual coverage, and timeliness, with some studies reporting detection leads of days to weeks over official statistics. At the same time, several recurring weaknesses limit operational uptake: causal and protocol-based reasoning (e.g., AEFI adjudication) remains brittle, multilingual and LMIC coverage is still uneven, and most evaluations are retrospective rather than prospective. The main open problem is therefore not whether LLMs can extract surveillance signals at all, but how to build surveillance systems that are simultaneously accurate, timely, calibrated, multilingual, privacy-aware, and integrable into the workflows of real public health agencies.

4.2 Epidemic forecasting & modeling (T2)

This section reviews LLMs, time-series foundation models, and agentic systems for epidemic forecasting, mechanistic modeling, and public health decision support. Compared with standard clinical prediction, epidemic modeling must operate under sparse, revised, and spatially coupled surveillance data, together with shifting interventions and behavior. These challenges make T2 distinct from conventional time-series forecasting: beyond point accuracy, useful systems must remain robust to reporting delays and distribution shift, represent uncertainty, and support decisions whose consequences propagate across populations and regions.

Time-series epidemic prediction. Most T2 work extends standard time-series forecasting with either multimodal context or epidemiological structure. A first line of work augments epidemic trajectories with exogenous public health signals. PandemicLLM combines epidemic time series with policy, genomic, and spatial signals (Du et al., 2025); EpiLLM and MIFlu integrate mobility or contextual text into autoregressive forecasting (Gong et al., 2025; Moon et al., 2025); and fine-tuned LLMs improve longer horizon influenza prediction (Li et al., 2025a). A second line introduces more explicit epidemiological inductive bias. Structurally informed approaches include compartment pretraining for epidemics(CAPE) (Liu et al., 2025f), simulation-trained foundation models (Mantis) (Dudley et al., 2025), and zero-shot and general-purpose forecasters (Kalahasti et al., 2025; Panja et al., 2025). A third, lighter-weight use treats LLMs less as standalone epidemic models and more as forecasting assistants or off-the-shelf baselines (Huang et al., 2025b; Liang et al., 2025). Taken together, these studies suggest that current progress comes less from replacing epidemic forecasting with generic language modeling than from enriching forecasts with broader context, transfer, or mechanistic structure. At the same time, much of the literature still emphasizes forecast performance on benchmark datasets, while leaving more operational questions, such as robustness to revisions, calibration under uncertainty, and generalization across outbreaks, regions, and diseases, less fully examined.

Epidemic simulation & agent-based modeling. Fewer papers use LLMs for simulation, and most do not replace classical epidemic simulators directly. Instead, one line of work uses LLMs to support simulator construction, specification, or iterative model design. ChatGPT supports iterative design of transmission models (Kwok et al., 2024), while EpiAgent formulates simulator construction as program synthesis with epidemiological flow-graph constraints (Datta et al., 2026). A complementary line embeds LLM agents directly in epidemic environments for policy reasoning (Aoki & Ghaffarzadegan, 2026). These two directions serve different purposes: the former treats LLMs as interfaces for building or modifying epidemiological models, whereas the latter uses them as decision-making agents inside structured epidemic worlds. This distinction is important because simulation in public health is not only a prediction problem, but also a counterfactual and planning problem. The main challenge is therefore not simply generating plausible trajectories, but

ensuring that natural-language flexibility remains aligned with epidemiological validity, intervention logic, and interpretable assumptions about transmission dynamics.

Emergency response & decision support. A third strand targets preparedness and operational decision-making. EPIWATCH provides open-source epidemic intelligence for early warning under incomplete surveillance (Quigley et al., 2025), while foundation model forecasting is being extended to counterfactual policy analysis (Kalahasti et al., 2025). LLM agents have been applied to coordinated intervention design across mobility-coupled regions (Shi et al., 2026) and adaptive policymaking in structured epidemic environments (Aoki & Ghaffarzadegan, 2026). Perspective papers highlight both the promise and deployment risks of these systems (Dixit, 2025; Rizzo et al., 2024; Lau et al., 2026). Relative to retrospective forecasting, this strand places stronger emphasis on actionability: systems are expected not only to estimate future epidemic states, but also to support scenario comparison, intervention timing, and policy coordination under uncertainty. This raises a higher bar for evaluation, since strong predictive performance alone does not guarantee reliable decision support when recommendations may alter behavior, resource allocation, or downstream disease dynamics.

Across these three strands, T2 is driven mainly by the Predictor and Simulator roles rather than the extraction-oriented Sensor role that dominates earlier parts of the public health pipeline. Current work develops along four broad directions: multimodal epidemic forecasting, mechanistically informed foundation models, LLM-assisted simulator construction, and agentic decision-making in structured epidemic settings. Compared with traditional forecasting pipelines, these systems broaden the range of inputs and planning functions that can be incorporated, but evaluation remains relatively narrow: many studies still prioritize forecast accuracy or proof-of-concept simulation over revision robustness, calibrated uncertainty, counterfactual validity, and real operational usefulness. The main open problem is therefore not merely forecasting epidemic curves more accurately, but building systems that couple flexible multimodal reasoning with epidemiological realism, reliable uncertainty quantification, and decision-ready support for public health practice.

4.3 Infodemiology & health communication (T3)

This section reviews how LLMs and related generative AI systems are used to monitor, interpret, and shape public health information ecosystems. T3 is broader and more socially embedded than classical epidemic forecasting: online discourse can function as an early population signal, the content of that discourse can itself become a target for intervention, and generative models can both repair and pollute the information environment. The literature therefore follows a maturity gradient from monitoring and classification, to evidence-grounded verification, to audience-specific intervention, and finally to governance of model-generated risk. This dual-use structure makes T3 one of the clearest settings in which LLMs are simultaneously public health tools and public health threats (De Angelis et al., 2023; Morita et al., 2024; Joseph et al., 2025; Saeidnia et al., 2026).

Misinformation detection & fact-checking. The largest T3 strand treats LLMs as sensors and auditors for health misinformation. Early work uses prompting or lightweight adaptation to classify claims, explain decisions, and combine machine judgments with human or crowd signals (Rostami & Hawamdeh, 2025; Zong et al., 2025). A second wave moves beyond classification toward evidence-grounded verification: retrieval-augmented systems check COVID-19 and vaccine claims against scientific literature, while long-form fact-checking decomposes complex narratives into claims that can be individually verified (Li et al., 2025b; Huang et al., 2025a). A third wave turns the lens back on the models themselves. New benchmarks include multimodal and LLM-generated vaccine misinformation (Zhang et al., 2025; Ahmad et al., 2025), and adversarial audits show that health chatbots can be induced to generate or endorse disinformation under jailbreak, role-playing, or system-instruction attacks (Hussain et al., 2025; Modi et al., 2025). Thus, progress in T3.1 is not only a story of better misinformation detection, but also a warning that detection systems must be evaluated against the generative threats they help create.

Vaccine attitudes & hesitancy monitoring. A second strand focuses on vaccine confidence, where the technical trajectory runs from stance detection to behaviorally relevant communication. LLMs classify sentiment, hesitancy, and fine-grained anti-vaccine concerns across COVID-19, HPV, and cross-national mRNA vaccine discourse (Annan et al., 2025; Kim et al., 2024; Mu et al., 2024; Boatman et al., 2024; Liu et al., 2025a; Xu et al., 2024a). This line is beginning to connect digital traces to real-world outcomes: online vaccination behavior signals have been linked to population vaccination patterns (Oh et al., 2025), suggesting that LLM-coded discourse may complement official uptake surveillance. The field is also moving from measurement to intervention. Tailored pro-vaccine messages, counter-arguments, common-ground-based framing, and chatbot systems explore whether LLMs can translate population attitudes into audience-aware communication (Xia et al., 2025; Stureborg et al., 2024; Dhanuka et al., 2025; Hou et al., 2025). These studies make vaccine communication the most mature T3 area for testing whether offline language-model performance can produce measurable public health behavior change.

Public health discourse analysis. Beyond misinformation and vaccination, LLMs are increasingly used as qualitative-analysis engines for large public health corpora. Rather than predicting disease trajectories, these systems summarize themes, identify concerns, and structure discourse at a scale that would be infeasible for manual coding. Human-validation studies suggest that LLMs can support inductive thematic analysis and broad discourse classification at substantially lower cost than expert-only workflows (Deiner et al., 2024; Espinosa & Salathé, 2024). Applications include climate-related public health discourse, Alzheimer's disease infodemiology, cytomegalovirus awareness, and campaign evaluation (Belova et al., 2025; Dai et al., 2025; Rosebrock et al., 2026). The main methodological value of this strand is not simply automation: it turns otherwise diffuse public conversation into analyzable public health intelligence, while raising unresolved questions about representativeness, platform bias, and whether discourse changes translate into policy or health outcomes.

Substance use and tobacco surveillance. Substance-use and tobacco studies form a distinct infodemiology strand because they combine high-stigma topics, rapidly evolving language, and strong links to population harm. LLMs have been used to monitor opioid-related chatter, identify self-disclosures of use and recovery, expand slang lexicons, classify tobacco and e-cigarette sentiment, and audit adherence to cessation guidance (Sidorov et al., 2025; Mittal et al., 2025; Han et al., 2025; Elmitwalli et al., 2024; Carpenter & Altman, 2023; Yang et al., 2024). This area provides unusually direct evidence that social-media signals can track real-world burden, with opioid discourse trends compared against external mortality patterns (Mittal et al., 2025). It also shows how T3 can move from passive monitoring to language intervention: stigma-detection and rewriting systems attempt to reduce harmful framing around substance use rather than merely describe it (Bouzoubaa et al., 2024). The limitations are equally clear: most studies remain U.S.-centric, text-only, retrospective, and weakly connected to downstream prevention or treatment programs.

Risk communication & health messaging. The most intervention-oriented part of T3 studies whether LLMs can generate useful, credible, and audience-appropriate responses. Evidence-grounded counterspeech uses retrieval-augmented and multi-agent frameworks to rebut misinformation while maintaining factuality and tone (Song et al., 2025; Anik et al., 2025). Broader corrective mechanisms include community notes, inoculation, myth correction, pro-vaccine message generation, common-ground tailoring, and chatbot-based information support (Wu et al., 2025a; Malek et al., 2026; Gullison & Fu, 2025; Stureborg et al., 2024; Reis et al., 2026; Hou et al., 2025; Laily et al., 2026). More generally, work on health-news quality evaluation, belief change in human–AI interaction, and adherence to public health guidance emphasizes that fluent communication is insufficient: public health communicators must be evidence-grounded, culturally and literacy appropriate, and operationally safe (Liu et al., 2025d; Lu et al., 2026; Abroms et al., 2025b).

Across these strands, prompting is the dominant adaptation strategy, fine-tuning is most useful for stance classification and domain-specific behavior labels, and RAG is emerging as the default architecture for fact-checking and counterspeech. The central gap is the transition from classification to intervention. Many systems report accuracy or F1, yet few assess downstream outcomes such as belief change, vaccination uptake, reduced stigma, or

improved adherence to guidance. Evaluation is also fragmented: misinformation detection often relies on automated metrics, while substance-use and communication studies more often require human or expert judgment. Finally, the field remains narrow in language, geography, platform, and modality. Moving from strong offline classification to robust, multilingual, multimodal, and prospectively evaluated communication systems remains the defining open challenge for T3.

4.4 Health equity, SDOH & global health (T4)

This section reviews how LLMs are used to identify social determinants of health (SDOH), audit bias and fairness, and extend health tools to low-resource settings. Health equity is a foundational public health principle, yet the social and structural factors driving disparities are among the hardest signals to measure at population scale: they are unevenly documented, embedded in free text, often invisible in structured fields, and shaped by language, geography, and the design of health systems themselves. LLMs are increasingly framed as a possible response to this measurement problem, both because they can read unstructured records at scale and because they can extend clinical and public health tools into linguistic and infrastructural environments that conventional NLP pipelines have largely ignored. At the same time, the same flexibility that makes LLMs attractive for equity research also makes them a new source of equity risk, motivating a parallel literature on bias auditing and equitable deployment.

SDOH information extraction. A first line of work uses LLMs as scalable extractors of social and structural risk factors that are otherwise locked inside free-text clinical notes (Ong et al., 2024). Instruction-tuned open models such as LLaMA have been used to extract a broad set of SDOH factors across multiple institutional EHR datasets, generalizing across documentation styles better than encoder-based clinical baselines, especially for class-imbalanced or under-represented factors (Keloth et al., 2025), while SBDH-Reader shows that prompt-engineered general-purpose LLMs can match this performance under independent cross-institutional validation without fine-tuning (Gu et al., 2025). To handle infrequent or sensitive categories, recent work proposes staged zero-shot pipelines that decompose extraction into context retrieval, relevance verification, and factor extraction, with the explicit goal of improving recall on rare SDOH factors that supervised baselines tend to miss (Wang et al., 2025a), and synthetic note generation, in which LLM-generated clinical notes are combined with authentic data to train downstream classifiers (Gabriel et al., 2024). A complementary line analyzes the equity orientation of the broader research landscape itself, using LLMs to extract health-equity covariates from registered research projects and to characterize how often demographic and structural factors are studied at all (Nananukul & Kejriwal, 2024). Together, these studies suggest that SDOH extraction has moved from feasibility studies toward multi-institution and increasingly deployment-oriented use, but most evidence is still drawn from US English-language EHRs, leaving generalization across health systems, languages, and documentation cultures as the central open question.

Bias detection and fairness auditing. A second strand explicitly audits whether LLMs reproduce or amplify existing health disparities, treating LLMs themselves as objects of equity evaluation. EquityMedQA introduces a 4,619-example adversarial benchmark and a multifactorial human assessment framework covering six dimensions of bias. It shows that adversarial prompts and diverse rater pools reveal substantially more bias than standard medical QA evaluations (Pfohl et al., 2024). EquityGuard takes a mitigation-oriented approach, using contrastive learning to reduce inequities that arise when clinically irrelevant socio-demographic factors are added to prompts for clinical trial matching and medical question answering (Ji et al., 2025). Domain-specific audits further identify concrete failure modes, including persistent anti-LGBTQIA+ bias across four mainstream LLMs, with inappropriate response rates of 43–65% (Chang et al., 2025), and intersectional bias in mental health reasoning that is amplified during multi-hop question answering (Haider et al., 2025). A complementary line of work uses LLMs to audit human rather than algorithmic bias: counterfactual analyses of emergency department triage notes show that female patients receive systematically lower severity ratings than male patients with otherwise identical clinical content (Adames et al., 2025). Perspective papers broaden the discussion from

model development to deployment-stage risks and proactive uses, arguing that clinical documentation, equity datasets, and equitable access must be redesigned as LLMs become routine intermediaries in health care (Singh et al., 2023; Pierson et al., 2025). Overall, this strand marks a shift from generic capability benchmarks toward adversarial and explicitly equity-focused evaluation. However, most existing audits remain English-only, focus on a narrow set of clinical scenarios, and rarely connect observed biases to downstream clinical or population health outcomes.

Global health and low-resource settings. A third strand asks whether LLMs can meaningfully serve populations and health systems beyond high-income, English-speaking contexts. Reviews and roadmaps document persistent inequities at the research level, with the bulk of LLM-in-health work originating in HICs and disproportionately optimized for English-language clinical contexts, and propose multidimensional frameworks for equitable adoption in low- and middle-income countries (LMICs) (Ong et al., 2026; Chen et al., 2025b). Benchmark efforts make this gap concrete: AfriMed-QA spans 60+ medical schools across 16 African countries and 32 specialties and shows that even strong general-purpose LLMs lag on USMLE-style questions in this setting, with biomedical-specialized models often underperforming general ones (Olatunji et al., 2024), while simulated-patient evaluations of ChatGPT for diseases disproportionately affecting LMICs report high diagnostic accuracy alongside frequent inclusion of unnecessary or harmful medications (Si et al., 2024). Linguistic adaptation is a second major theme: Vietnamese-specific fine-tuning with LoRA/QLoRA on 337K health prompt-response pairs improves communication quality while enabling on-premise deployment (Bui et al., 2025), and modular architectures such as L2M3 combine LLMs with machine translation to support multilingual medical knowledge for community health workers (Gangavarapu, 2024). Applied evaluations move closer to real LMIC workflows: silent trials and prospective protocols evaluate LLM assistance to community health workers in Rwanda, including ambient listening of CHW–patient encounters in Kinyarwanda, where strong frontier models can approach CHW-level referral accuracy while smaller models fall sharply behind (Shimelash et al., 2026; Menon et al., 2025), and LLMs have been used to analyze open-ended survey responses about why children in the DRC fail to access vaccination services, achieving up to 96% accuracy with as few as 20–100 labeled examples (Burstein et al., 2025). Taken together, this work shifts the equity question from whether LLMs are biased on standard benchmarks toward whether they can be deployed in linguistically, infrastructurally, and clinically realistic LMIC settings without amplifying existing global health inequities.

Across these three strands, T4 is shaped by two distinct LLM roles: a *Sensor* role used to extract SDOH and equity-relevant signals from unstructured records, and an *Auditor* role used to evaluate, mitigate, or surface bias in models, datasets, and human decisions. Adaptation strategies follow this division: instruction tuning and prompt engineering dominate SDOH extraction, while bias auditing relies more heavily on adversarial benchmarks, counterfactual prompts, and contrastive learning, and global-health work increasingly couples lightweight fine-tuning with translation and CHW-facing deployment. Compared to standard NLP equity research, this literature is distinctive in its move toward intersectional, deployment-stage, and multilingual evaluation rather than purely demographic-attribute accuracy gaps. Yet several gaps persist: most extraction work remains tied to US EHRs and English text, most bias audits do not connect to measurable health outcomes, and most LMIC studies are still pilot- or protocol-stage with limited prospective evidence. The central open problem for T4 is therefore to develop equity-aware LLM systems whose accuracy, bias, and downstream impact are jointly evaluated under realistic linguistic, geographic, and clinical conditions, rather than at the level of individual benchmarks or single-site studies.

4.5 Population intervention & practice (T5)

This section reviews how LLMs are used for direct public health action: deploying interventions, supporting frontline workers, allocating scarce resources, screening for mental and behavioral health conditions, and shaping everyday health practice. Compared to surveillance and forecasting, T5 is the most user-facing part of the public health pipeline, and the people on the other side of the system are rarely clinicians: they are community

health workers, smokers trying to quit, individuals navigating HIV-prevention services, socially isolated older adults, or recovery seekers turning to online forums for support. This shifts the technical problem from "can the model produce a correct answer?" to "can the model produce a credible, safe, and contextually appropriate response in an ongoing interaction with a non-expert user?", and it makes deployment design, human oversight, and evaluation under real-world use central rather than peripheral. Most systems in T5 therefore take a *Communicator*-oriented form and pair generative capabilities with domain knowledge, clinical guidelines, retrieval augmentation, or expert-in-the-loop validation.

Digital health interventions. A first strand studies LLM-powered chatbots, conversational agents, and decision-support tools that deliver counseling, screening, or intervention content directly to end users. HIV prevention is one of the most active testbeds: the AHF retrospective cohort of more than 155,000 adults shows that AI-augmented communication is associated with significantly higher PrEP initiation, follow-up attendance, and appointment adherence, with the largest gains among younger and historically underserved patients (Narayan et al., 2026); "Your Choice" achieves a System Usability Scale of 92/100 in stigma-free PrEP eligibility assessment among South African adults (Govathson et al., 2026); CHIA, a GPT-4o-based bilingual PrEP chatbot grounded in motivational interviewing, scores high on accuracy and trustworthiness in internal testing while exposing significant Spanish-language performance gaps (Tao et al., 2026); and qualitative work with men who have sex with men and transgender women in KwaZulu-Natal shows that multi-agent HIV/mental-health chatbots are valued for privacy, anonymity, and human-like interaction, but penalized for slow responses, repetitive replies, and limited emotional rapport (Humphries et al., 2026). A second cluster of systems supports community health workers (CHWs) and frontline staff in LMIC settings: ASHABot, a WhatsApp-based RAG chatbot deployed with India's ASHA workers, is treated as both a private channel for sensitive questions and a near-authoritative reference, with supervisors expanding the knowledge base in response to usage patterns (Ramjee et al., 2025); the SMART health GPT pilot uses RAG to provide ASHAs with locally relevant maternal-care information and reports that retrieval-grounded designs are preferred over fine-tuning for this use case (Al Ghadban et al., 2023); and PRIORITY2REWARD takes the same population further by translating health-worker preferences into reward functions for restless multi-armed-bandit allocation in maternal mHealth programs, allowing CHWs to specify equity priorities and immediately inspect the resulting policies (Verma et al., 2025). Closely related work uses LLM agents to simulate maternal-health beneficiary listening behavior with explicit uncertainty estimation and decision-focused evaluation, demonstrating how LLM predictions can shape allocation decisions rather than just describe them (Martinson et al., 2025). Behavioral health and substance use form a third cluster: BeFreeBot, a ChatGPT-based SMS smoking-cessation chatbot, achieves 70% engagement, 100% USPSTF guideline alignment, and 30% self-reported 7-day abstinence in a 23-person pilot (Abroms et al., 2025a), while LLM fine-tuning with data downsampling raises message-intent F1 from 0.41 off-the-shelf to 0.91 in a deployed smoking-cessation mHealth program (Rahman et al., 2026); complementary studies use LLMs to mine Reddit for evolving barriers to opioid recovery, surfacing pandemic-era shifts from stigma-related obstacles to systemic barriers like treatment discontinuity (Ekanayake et al., 2025); and CHI 2023's CareCall study of a deployed LLM chatbot for socially isolated individuals provides one of the most explicit accounts of the deployment trade-offs at this layer, including gains in holistic understanding and emotional support against challenges in unpredictability and reliability for vulnerable users (Jo et al., 2023).

Population mental health screening. A second, narrower but important strand uses LLMs as scalable screeners for mental and behavioral health signals in non-clinical text. Mental-LLM is representative of the technical pattern: by instruction-fine-tuning on mental-health data, Mental-Alpaca and Mental-FLAN-T5 outperform GPT-3.5 by 10.9% and GPT-4 by 4.8% in balanced accuracy on population mental-health prediction from online text, performing on par with task-specific state-of-the-art models while remaining substantially smaller (Xu et al., 2024b). This style of work is appealing for public health precisely because clinical screening cannot reach the populations whose distress is most visible online, but it raises distinct concerns that do not arise in clinical NLP: known racial and gender biases in mental-health classification, the ethical status of repurposing non-clinical social-media content, and the absence of meaningful follow-up infrastructure when screening runs at scale. As a

result, T5.2 currently sits at a tension point: technical performance is approaching a level where deployment is conceivable, but the governance and equity questions surrounding population-level mental-health screening from social media remain effectively unresolved.

Health promotion and literacy. A third strand examines broader health-promotion and public health practice applications, where LLMs are typically embedded in routine workflows rather than user-facing chat. This includes LLM-based identification of workplace violence and communication failures from healthcare safety reports, where physical-violence and verbal-abuse classification reaches F1 above 0.80 and 0.94 respectively and outperforms prior NLP techniques (Becker et al., 2025; Sridi & Brigui, 2023); evaluation of LLMs as natural-language-to-code translators for aggregated public health data, in which 11 systems are benchmarked on Czech public-health analyses and shown to vary widely in reliability, with executable but incorrect code as a recurring failure mode (Klempir et al., 2025); and domain-specific reviews and questionnaires assessing LLM use in public health dentistry and disease prevention, where the literature still emphasizes opportunity-and-threat framings and curriculum readiness rather than empirical effectiveness (Tiwari et al., 2023; Angyal et al., 2025). A growing thread within this strand explicitly studies occupational and population-level practice settings, asking how LLMs reshape the role of public health practitioners themselves rather than only the content they produce (Vos et al., 2025).

Across these three strands, T5 is the part of the survey where "does this system work?" depends most directly on real users, real workflows, and real time. A few patterns are consistent. First, evaluation has moved beyond accuracy: the most informative T5 studies report engagement, behavioral outcomes (e.g., PrEP initiation, abstinence, referral accuracy), or qualitative acceptance from end users, and they increasingly draw participants from the populations targeted by the intervention rather than from convenience samples. Second, retrieval grounding, expert-in-the-loop validation, and modest fine-tuning on program-specific data (rather than larger generic models) tend to be the design choices that move pilots toward deployment, especially for CHWs and underserved populations. Third, equity and safety concerns are not a separate workstream: they appear inside almost every deployment study, from Spanish-language gaps in HIV chatbots to bias in population mental-health prediction. The main open problem for T5 is therefore to move from short, single-site pilots toward sustained, multi-site, outcome-evaluated deployments in which intervention effectiveness, equity, and operational safety are measured jointly under everyday conditions of use.

4.6 Governance, ethics & policy (T6)

This section reviews how LLM-based public health systems are governed, evaluated, and deployed responsibly. As these tools move from research prototypes toward operational use in surveillance, decision support, and policy work, governance questions become unavoidable: who should benchmark these systems, what classes of risk should be measured, how should hallucinations and biases be surfaced before they shape resource allocation, and what institutional arrangements are needed before deployment is responsible? Compared with preceding task categories, T6 focuses less on what LLMs can do in principle and more on how they should be built, benchmarked, and integrated in practice. We organize this literature into three strands: governance-oriented system design, evaluation and policy-readiness benchmarking, and deployment and implementation science.

Ethics, safety, and risk frameworks. A first strand studies governance-oriented system design, treating the system architecture itself as the locus of accountability. One sub-direction democratizes mechanistic and modeling tools that previously required expert programming: an LLM-based assistant grounded in the open-source CMS framework lets users generate, simulate, and analyze compartmental disease models from natural-language descriptions, with few-shot EMODL examples dramatically reducing zero-shot syntactic errors (Proctor & Chabot-Couture, 2024). A closely related sub-direction builds multi-agent research and response systems with explicit roles and validation steps. EpidemIQs uses a dual-agent (scientist + task-expert) architecture to autonomously conduct epidemic research from literature review through simulation to manuscript generation, achieving 79% task success across diverse scenarios at roughly $1.57 per study (Samaei et al., 2026), while EpiPlanAgent

integrates task decomposition, knowledge grounding, and plan simulation/validation to automate digital emergency response planning with expert-in-the-loop checks on guideline alignment (Mao et al., 2025). Generative agent-based modeling extends this paradigm into population-level behavioral simulation: agents in epidemic environments spontaneously quarantine when sick or self-isolate when cases rise, producing emergent multi-wave dynamics that begin to resemble real pandemics (Williams et al., 2023), and "Infected Smallville" uses the same machinery to test behavioral immune system hypotheses by manipulating disease threat in agent settings (Choi et al., 2025). A fourth sub-direction targets disaster and outbreak response: DisasterResponseGPT generates plans of action grounded in domain guidelines that are comparable to human-authored ones while reducing turnaround from hours to seconds (Goecks & Waytowich, 2023), and broader review and commentary work argues that the field underinvests in integrating LLM-based agents into outbreak analytics pipelines and disaster-management workflows more generally (Xu et al., 2025a; van Hoek et al., 2024). Finally, several proposals reframe public health infrastructure itself around retrieval and architectural grounding: MEGA-RAG combines multi-source evidence retrieval, cross-encoder reranking, and discrepancy-aware refinement to cut hallucination rates by more than 40% on public health QA (Xu et al., 2025b), and a position paper argues that public health systems should embrace a multi-layered epidemic early-warning architecture coupling distributed LLM agents, centralized analytics, and regionally enriched knowledge bases (DENG & Jin, 2025). Across these systems, evidence grounding, expert checkpoints, and modular validation are recurring design commitments, even though most evaluations remain proof-of-concept and real-world emergency deployment is still rare.

Policy analysis and evaluation. A second strand asks the more direct question of whether current LLMs are fit to participate in public health reasoning and policy support, and answers it by building benchmarks that probe progressively harder tasks. EpiQAL, the first diagnostic benchmark for epidemiological question answering, partitions tasks into factual recall, multi-step inference, and conclusion reconstruction, and shows that current model rankings shift across subsets, that scale alone does not predict success, and that strong general LLMs can still fail systematically on evidence-grounded epidemiological reasoning (Wei et al., 2026). PubHealthBench, built from 8,000+ questions derived from 687 UK government public-health guidance documents, finds that frontier proprietary models exceed 90% accuracy on multiple-choice items, outperforming humans with cursory search-engine use, but drop below 75% on free-form responses where deployment risk is highest (Harris et al., 2025), and an open-source benchmarking platform combining a Ruby/Rails prediction prototype with an R Shiny evaluation app shows high task- and dataset-level variability across LLMs even when overall accuracy looks competitive (Espinosa et al., 2025). Application-oriented evaluations make these limitations concrete in policy-relevant settings: LLMs evaluated on childhood lead-testing resource allocation average only 0.46 accuracy and frequently overlook the highest-risk neighborhoods, citing outdated data and stereotypical narratives in place of current epidemiological evidence (Afane et al., 2025); bespoke RAG pipelines for public-health evidence reviews achieve 90%+ acceptable accuracy on objective study-design fields but fall to 54% or below on the outcome data that matters most for synthesis (Simmons et al., 2025); retrieval-augmented text-to-SQL pipelines materially improve epidemiological querying over EHR/claims data but remain insufficient for unsupervised use (Ziletti & DAmbrosi, 2024); and SatHealth demonstrates that satellite-derived environmental factors, when fused with claims data and SDOH indices, substantially improve regional public-health modeling, illustrating the more general point that benchmarks confined to text leave critical multimodal signals out of view (Wang et al., 2025b). Taken together, this literature pushes back against unsupervised deployment of LLMs in high-stakes policy settings, but it also helps locate the boundary: structured factual recall is increasingly tractable, while free-form reasoning, resource allocation, and outcome-level evidence synthesis still require careful human supervision and grounded retrieval.

Deployment and implementation science. A third strand turns from system design and benchmarking to the institutional and regulatory conditions under which LLMs can be safely operationalized in public health. A risk taxonomy and reflection tool for LLM adoption, derived from focus groups with public health professionals and individuals with lived experience across vaccines, opioid use disorder, and intimate partner violence, identifies four

distinct risk dimensions, individual behaviors, human-centered care, information ecosystems, and technology accountability, and translates them into reflection questions that public health programs can use before adopting LLMs (Zhou et al., 2025a). A complementary cross-sectional audit of the OpenAI GPT Store finds 1,055 health-related custom GPTs serving over 360,000 cumulative conversations, with 624 (59%) using healthcare-professional titles in their names despite none being approved by FDA, EU MDR, or TGA: an empirical case study of unregulated "role creep" in consumer-facing health AI (Chu et al., 2025). Geographic and global-health blind spots are also documented systematically: an evaluation of generative AI for health-policy identification across all UN Member States reports 78% concordance with expert datasets for vaccination policy but markedly worse performance in African, Southeast Asian, and Eastern Mediterranean regions, raising the prospect that uncritical deployment could entrench global inequities (Wilson et al., 2024). Against this backdrop, AMR-Policy GPT provides one of the clearest worked examples of a governance-aware, retrieval-grounded deployment: built on a multilingual database of antimicrobial-resistance national action plans from 146 countries, it answers AMR policy queries with traceable references and explicit One Health framing (Chen et al., 2025a). Position pieces tie these threads together by arguing for validation-board frameworks for public health AI (Hattab et al., 2025), systematic AI literacy in public health curricula (Acosta, 2025), and clearer institutional accountability for generative AI in scientific communication and policy work (Yoga Ratnam, 2025), alongside more skeptical perspectives on whether epidemiology can in fact be automated given current hallucination, dataset-access, and reliability constraints (Bann et al., 2026).

Across these three strands, T6 is dominated by the *Auditor* and *Simulator* roles: LLMs are evaluated as candidate components of public health infrastructure, audited for hallucination, bias, accountability, and privacy risks, and embedded in agentic systems whose architectures are themselves the unit of governance. Several patterns are consistent. First, retrieval grounding, evidence traceability, and multi-agent validation are emerging as default design choices for any system intended to support real public health reasoning. Second, frontier models perform well on structured benchmarks but degrade sharply on free-form reasoning, resource allocation, and LMIC settings, exposing a gap between benchmark accuracy and decision-readiness. Third, governance-oriented work increasingly couples technical evaluation with institutional and regulatory analysis, including risk taxonomies, role-creep audits, validation boards, and AI literacy in public health training. The main open problem is therefore not whether LLMs can produce useful public health artifacts, but how to build the evaluation, oversight, and deployment infrastructure that allows their use in high-stakes settings to be responsibly accountable rather than only technically impressive.

5 Discussion and Conclusion

Our analysis highlights a central tension: current LLMs are designed mainly for document-level language processing, whereas public health reasoning is population-level, spatiotemporal, and socially embedded. This mismatch is not merely technical; it shapes what current systems do well, what they still miss, and where the field is structurally imbalanced. Across our taxonomy, the literature is strongest in extraction-oriented sensing and communication-heavy tasks, while forecasting, simulation, and deployment-oriented evaluation remain comparatively underdeveloped. More broadly, public health tasks form an operational pipeline rather than a collection of isolated benchmarks: surveillance signals feed forecasts, forecasts influence communication and intervention, and all of these ultimately affect governance and resource allocation. As a result, weaknesses at one stage can propagate downstream. We summarize the resulting research agenda in four directions.

Cross-modal integration of epidemiological data. Key non-textual signals—epidemic time series, genomics, mobility, and wastewater—remain largely siloed. Initial alignment efforts combine time series with policy (Du et al., 2025), integrate compartmental priors (Liu et al., 2025f), or map infection dynamics to language (Gong et al., 2025). Future progress requires multimodal epidemiological foundation models with mechanistic grounding to handle distribution shifts and emerging outbreaks (Kwok et al., 2024). More fundamentally, public health forecasting and preparedness cannot be reduced to text understanding with

auxiliary covariates appended at the end. Epidemic trajectories evolve under feedback from interventions, behavior, seasonality, and spatial coupling, while genomic and environmental signals often operate on different temporal and biological scales. A central challenge is therefore representation: future systems must jointly encode textual guidance, temporal dynamics, spatial interaction, and mechanistic constraints, rather than treating these modalities as loosely coupled inputs. This also changes what counts as success. It is not enough to improve retrospective benchmark performance; models must remain robust under reporting delays, revision, regime shifts, and emerging pathogens. In this sense, the next generation of public health LLM systems will likely look less like standalone chat models and more like multimodal reasoning systems grounded in epidemiological structure.

The dual-role paradox: LLMs as both tool and threat. The same models used to detect misinformation can also generate it at scale (De Angelis et al., 2023; Hussain et al., 2025). This risk is already visible in system-instruction attacks on health chatbots (Modi et al., 2025) and in the difficulty of distinguishing AI-generated from human-authored health misinformation (Zhang et al., 2025). In public health, such failures can propagate through information networks and undermine trust. This makes safety a population-level problem, motivating retrieval-grounded systems tied to evolving guidance (Xu et al., 2025b) and literacy-aware counterspeech generation (Song et al., 2025). The key implication is that safety in public health cannot be framed only as single-response harmlessness. A fluent but subtly misleading model may distort risk perception, reduce adherence to guidance, or amplify uncertainty during outbreaks, even when each individual interaction appears plausible in isolation. This is especially important because T3 and T5 systems are often designed to operate at scale, where seemingly small per-instance errors can accumulate into large population-level effects. The field therefore needs a broader safety lens that includes provenance, susceptibility to adversarial prompting, consistency with evolving guidance, and downstream effects on trust and behavior. More generally, the same generative flexibility that makes LLMs attractive for surveillance summarization, public communication, and triage support also creates new attack surfaces. Public health applications should therefore be evaluated not only for utility, but for their potential to accelerate the very harms they are intended to mitigate.

Population-scale alignment and global health equity. Public health demands equitable population-level performance. However, current systems often reproduce disparities, showing higher adverse-event risks for disadvantaged patients (Liu et al., 2025c) and racial gaps in mortality classification (Parker, 2025). Globally, an English-centric reliance on proprietary models (Chen et al., 2025b) creates a "multilingual penalty," degrading performance across Southeast Asian languages (Zhou et al., 2025b), Rwanda (Shimelash et al., 2026), and other African settings (Wilson et al., 2024). Mitigating these challenges requires equity-aware alignment data, subgroup-sensitive rewards, improved cross-lingual transfer, and smaller on-premise models tailored for low-resource deployment (Pfohl et al., 2024; Ji et al., 2025). Yet equity in public health should be understood as more than a post hoc fairness audit. In this domain, inequity can emerge at multiple levels: in what data are collected, which populations are represented in benchmarks, what languages are supported, which institutions can afford deployment, and how model outputs are translated into policy. The current literature still leans heavily toward English-centric and resource-rich settings, even though many of the most consequential public health use cases involve multilingual communication, weak surveillance infrastructure, and low-resource operational environments. This means that alignment for public health must be population-aware from the outset. Models must be evaluated across regions, subgroups, and health systems, and they must support local adaptation rather than assuming that a single frontier model can generalize universally. In practice, this also elevates questions of reproducibility, accessibility, and deployment sovereignty: open and smaller models may be less capable in some settings, but they can be essential for privacy-preserving, locally governed, and infrastructure-constrained public health use.

Evaluation beyond static benchmarks: synthetic societies. Static QA benchmarks are insufficient for tasks involving sequential decisions, delayed feedback, and collective behavior. Generative agent-based modeling offers a promising alternative: LLM agents can quarantine in simulated epidemics (Williams et al., 2023), reduce social contact under dis-

ease threat (Choi et al., 2025), and improve coordinated intervention outcomes (Shi et al., 2026). A natural next step is synthetic societies that can serve as population-level stress tests for scenarios that are difficult to evaluate in the real world. More broadly, public health evaluation must move beyond single-step correctness toward decision relevance. A system that performs well on extraction, classification, or QA may still fail when deployed in settings with delayed surveillance, conflicting guidance, heterogeneous populations, or intervention feedback. This suggests a need for evaluation protocols that are longitudinal, subgroup-aware, and operationally grounded. Useful benchmarks should test not only whether a model can answer correctly, but whether it remains calibrated under uncertainty, robust to temporal drift, consistent across linked tasks, and safe under iterative human–AI interaction. Synthetic societies are promising precisely because they provide a bridge between offline benchmarking and real-world deployment: they can expose emergent failure modes, coordination breakdowns, and policy side effects that are invisible in static test sets. At the same time, they should be viewed as complements rather than substitutes for real deployment studies, prospective trials, and implementation science. The long-term goal is not merely more realistic simulation, but evaluation frameworks that reflect how public health decisions unfold over time and across populations.

Conclusion. This survey organizes 192 papers through a two-dimensional taxonomy spanning six public health tasks and five LLM roles. Overall, the field is strongest in extraction-oriented sensing but remains early in forecasting, simulation, alignment, and multilingual deployment. Our synthesis suggests that the main bottleneck is not simply the lack of larger or more capable language models, but the mismatch between document-centered architectures and the requirements of population health reasoning. Public health systems must integrate heterogeneous modalities, support action under uncertainty, remain equitable across languages and settings, and be evaluated in ways that reflect real downstream consequences. Public health should therefore be viewed not only as an application domain, but also as a demanding testbed for multimodal learning, equitable alignment, and realistic evaluation. The accompanying repository at `https://github.com/publichealthllm/llms-in-public-health-survey` is maintained to track this rapidly evolving literature.

References

Andrea Abate, Elisa Poncato, Maria Antonietta Barbieri, Greg Powell, Andrea Rossi, Simay Peker, Anders Hviid, Andrew Bate, and Maurizio Sessa. Off-the-shelf large language models for causality assessment of individual case safety reports: A proof-of-concept with covid-19 vaccines: A. abate et al. *Drug Safety*, 48(7):805–820, 2025.

Lorien C Abroms, Christina N Wysota, Artin Yousefi, Tien-Chin Wu, and David A Broniatowski. Chatgpt-based chatbot for help quitting smoking via text messaging: An interventional study. *JMIR formative research*, 9:e79402, 2025a.

Lorien C Abroms, Artin Yousefi, Christina N Wysota, Tien-Chin Wu, and David A Broniatowski. Assessing the adherence of chatgpt chatbots to public health guidelines for smoking cessation: content analysis. *Journal of medical Internet research*, 27:e66896, 2025b.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jose A Acosta. Perspective: advancing public health education by embedding ai literacy. *Frontiers in digital health*, 7:1584883, 2025.

Ariel Guerra Adames, Marta Avalos, Océane Doremus, Cédric Gil-Jardiné, and Emmanuel Lagarde. Uncovering judgment biases in emergency triage: A public health approach based on large language models. *Proceedings of Machine Learning Research*, 259:420–439, 2025.

Mohamed Afane, Ying Wang, and Juntao Chen. Can llms help allocate public health resources? a case study on childhood lead testing. *arXiv preprint arXiv:2511.18239*, 2025.

Syed Talal Ahmad, Haohui Lu, Sidong Liu, Annie Lau, Amin Beheshti, Mark Dras, and Usman Naseem. Vaxguard: A multi-generator, multi-type, and multi-role dataset for detecting llm-generated vaccine misinformation. *arXiv preprint arXiv:2503.09103*, 2025.

Yasmina Al Ghadban, Huiqi Lu, Uday Adavi, Ankita Sharma, Sridevi Gara, Neelanjana Das, Bhaskar Kumar, Renu John, Praveen Devarsetty, and Jane E Hirst. Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. *medRxiv*, pp. 2023–12, 2023.

Viola Angyal, Ádám Bertalan, Péter Domján, Helga Judit Feith, and Elek Dinya. Development of a questionnaire for assessing the use of chatgpt in primary and secondary disease prevention. *Frontiers in Public Health*, 13:1709611, 2025.

Anirban Saha Anik, Xiaoying Song, Elliott Wang, Bryan Wang, Bengisu Yarimbas, and Lingzi Hong. Multi-agent retrieval-augmented framework for evidence-based counterspeech against health misinformation. *arXiv preprint arXiv:2507.07307*, 2025.

Augustine Annan, Amanda L Eiden, Dong Wang, Jingcheng Du, Majid Rastegar-Mojarad, Varun Kumar Nomula, and Xiaoyan Wang. Evaluating large language models for sentiment analysis and hesitancy analysis on vaccine posts from social media: Qualitative study. *JMIR Formative Research*, 9:e64723, 2025.

Goshi Aoki and Navid Ghaffarzadegan. Ai agents as policymakers in simulated epidemics. *arXiv preprint arXiv:2601.04245*, 2026.

David Bann, Ed Lowther, Liam Wright, and Yevgeniya Kovalchuk. Why can't epidemiology be automated (yet)? *International Journal of Epidemiology*, 55(1):dyaf210, 2026.

Mike Becker, Sy Hwang, Emily Schriver, Caryn Douma, Caoimhe Duffy, Joshua Atkins, Caitlyn McShane, Jason Lubken, Asaf Hanish, John D McGreevey, et al. Automatically identifying event reports of workplace violence and communication failures using large language models. *AMIA Summits on Translational Science Proceedings*, 2025:74, 2025.

Anna Belova, Raquel A Silva, Dylan M Vorndran, and Natalie R Sampson. Using large language models to learn from recent climate change discourse in public health. *PLoS One*, 20(4):e0321309, 2025.

Nahid Bhadelia, Ioannis Ch Paschalidis, John S Brownstein, and Britta Lassman. A beacon for novel disease threats: Leveraging artificial intelligence for informal event-based outbreak surveillance, 2026.

Dannell Boatman, Abby Starkey, Lori Acciavatti, Zachary Jarrett, Amy Allen, and Stephenie Kennedy-Rea. Using social listening for digital public health surveillance of human papillomavirus vaccine misinformation online: Exploratory study. *JMIR infodemiology*, 4: e54000, 2024.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arber, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Layla Bouzoubaa, Elham Aghakhani, and Rezvaneh Rezapour. Words matter: Reducing stigma in online conversations about substance use with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9139–9156, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nhat Bui, Giang Nguyen, Nguyen Nguyen, Bao Vo, Luan Vo, Tom Huynh, Arthur Tang, Van Nhiem Tran, Tuyen Huynh, Huy Quang Nguyen, et al. Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine*, 263:108655, 2025.

Roy Burstein, Eric Mafuta, and Joshua L Proctor. Large language models for analyzing open text in global health surveys: why children are not accessing vaccine services in the democratic republic of the congo. *International health*, 17(5):843–852, 2025.

Kristy A Carpenter and Russ B Altman. Using gpt-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. *Biomolecules*, 13(2):387, 2023.

Ronald Carshon-Marsh, Richard Wen, Thomas Kai Sze Ng, Rajeev Kamadod, Isaac Bogoch, Susan J Bondy, Theodore J Witek, and Prabhat Jha. Comparison of verbal autopsy using a large language model to biologically confirmed causes of death for malaria and other communicable diseases among children in six sub-saharan african countries. *Malaria Journal*, 25(1):77, 2026.

Crystal T Chang, Neha Srivathsa, Charbel Bou-Khalil, Akshay Swaminathan, Mitchell R Lunn, Kavita Mishra, Sanmi Koyejo, and Roxana Daneshjou. Evaluating anti-lgbtqia+ medical bias in large language models. *PLOS Digital Health*, 4(9):e0001001, 2025.

Cai Chen, Shu-Le Li, Anthony D So, Yao-Yang Xu, Zhao-Feng Guo, Xinbing Wang, David W Graham, and Yong-Guan Zhu. Using large language models to assist antimicrobial resistance policy development: Integrating the environment into health protection planning. *Environmental Science & Technology*, 59(2):1243–1252, 2025a.

Haichao Chen, Dian Zeng, Yiming Qin, Zeyue Fan, Faye Ng Yu Ci, David C Klonoff, John S Ji, Shuyang Zhang, Kwesi Nyan Amissah-Arthur, Michelle María Jiménez de Tavárez, et al. Large language models and global health equity: a roadmap for equitable adoption in lmics. *The Lancet Regional Health–Western Pacific*, 63, 2025b.

Yiqun T Chen, Tyler H McCormick, Li Liu, and Abhirup Datta. Lava: Language model assisted verbal autopsy for cause-of-death determination. *arXiv preprint arXiv:2509.09602*, 2025c.

Soyeon Choi, Kangwook Lee, Oliver Sng, and Joshua M Ackerman. Infected smallville: How disease threat shapes sociality in llm agents. *arXiv preprint arXiv:2506.13783*, 2025.

Bianca Chu, Natansh D Modi, Bradley D Menz, Stephen Bacchi, Ganessan Kichenadasse, Catherine Paterson, Joshua G Kovoor, Imogen Ramsey, Jessica M Logan, Michael D Wiese, et al. Generative ai's healthcare professional role creep: a cross-sectional evaluation of publicly accessible, customised health-related gpts. *Frontiers in Public Health*, 13:1584348, 2025.

Sergio Consoli, Peter Markov, Nikolaos I Stilianakis, Lorenzo Bertolini, Antonio Puertas Gallardo, and Mario Ceresa. Epidemic information extraction for event-based surveillance using large language models. In *International Congress on Information and Communication Technology*, pp. 241–252. Springer, 2024.

Sergio Consoli, Pietro Coletti, Peter V Markov, Lia Orfei, Indaco Biazzo, Lea Schuh, Nicolas Stefanovitch, Lorenzo Bertolini, Mario Ceresa, and Nikolaos I Stilianakis. An epidemiological knowledge graph extracted from the world health organization's disease outbreak news. *Scientific Data*, 12(1):970, 2025.

Isabel Coutinho, Gonçalo M Correia, Bruno Martins, Afonso Moreira, and André Peralta-Santos. Icd coding of death certificates with generative language models. *PLOS Digital Health*, 5(2):e0001245, 2026.

Haixing Dai, Yiwei Li, Zhengliang Liu, Lin Zhao, Zihao Wu, Suhang Song, Shen Ye, Dajiang Zhu, Xiang Li, Sheng Li, et al. Ad-autogpt: An autonomous gpt for alzheimer's disease infodemiology. *PLOS Global Public Health*, 5(5):e0004383, 2025.

Chathuri Daluwatte, Alena Khromava, Yuning Chen, Laurence Serradell, Anne-Laure Chabanon, Anthony Chan-Ou-Teung, Cliona Molony, and Juhaeri Juhaeri. Application of a language model tool for covid-19 vaccine adverse event monitoring using web and social media content: Algorithm development and validation study. *JMIR infodemiology*, 4:e53424, 2024.

Rituparna Datta, Zihan Guan, Baltazar Espinoza, Yiqi Su, Priya Pitre, Srini Venkatramanan, Naren Ramakrishnan, and Anil Vullikanti. Agentic framework for epidemiological modeling. *arXiv preprint arXiv:2602.00299*, 2026.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in public health*, 11:1166120, 2023.

Michael S Deiner, Vlad Honcharov, Jiawei Li, Tim K Mackey, Travis C Porco, and Urmimala Sarkar. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: human validation study. *JMIR infodemiology*, 4(1):e59641, 2024.

OU DENG and Qun Jin. Position: Public health systems should embrace a multi-layered epidemic early-warning with llm agents and local knowledge enhancement. 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Utsav Dhanuka, Soham Poddar, and Saptarshi Ghosh. Utilising large language models for generating effective counter arguments to anti-vaccine tweets. *arXiv preprint arXiv:2510.16359*, 2025.

Narendra M Dixit. Leveraging large language models for pandemic preparedness: Computational epidemiology. *Nature Computational Science*, 5(6):438–439, 2025.

Hongru Du, Yang Zhao, Jianan Zhao, Shaochong Xu, Xihong Lin, Yiran Chen, Lauren M Gardner, and Hao 'Frank' Yang. Advancing real-time infectious disease forecasting using large language models. *Nature Computational Science*, 5(6):467–480, 2025.

Carson Dudley, Reiden Magdaleno, Christopher Harding, Ananya Sharma, Emily Martin, and Marisa Eisenberg. Mantis: A simulation-grounded foundation model for disease forecasting. *arXiv preprint arXiv:2508.12260*, 2025.

Vinu Ekanayake, Md Sultan Al Nahian, and Ramakanth Kavuluru. Mining social media for barriers to opioid recovery with llms. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pp. 83–99, 2025.

Sarah Elmitwalli, John Mehegan, Allen Gallagher, and Rasha Alebshehy. Enhancing sentiment and intent analysis in public health via fine-tuned large language models on tobacco and e-cigarette-related tweets. *Frontiers in Big Data*, 7:1501154, 2024. doi: 10.3389/fdata.2024.1501154. URL https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1501154/full.

Laura Espinosa and Marcel Salathé. Use of large language models as a scalable approach to understanding public health discourse. *PLOS Digital Health*, 3(10):e0000631, 2024.

Laura Espinosa, Djilani Kebaili, Sergio Consoli, Kyriaki Kalimeri, Yelena Mejova, and Marcel Salathé. Open-source solution for evaluation and benchmarking of large language models for public health. *medRxiv*, pp. 2025–03, 2025.

Arthur J Funnell, Panayiotis Petousis, Fabrice Harel-Canada, Ruby Romero, Alex AT Bui, Adam Koncsol, Hritika Chaturvedi, Chelsea Shover, and David Goodman-Meza. Improving drug identification in overdose death surveillance by using clinical natural language processing models. *Journal of Forensic Sciences*, 2026.

Rodney A Gabriel, Onkar Litake, Sierra Simpson, Brittany N Burton, Ruth S Waterman, and Alvaro A Macias. On the development and validation of large language model-based classifiers for identifying social determinants of health. *Proceedings of the National Academy of Sciences*, 121(39):e2320716121, 2024.

Agasthya Gangavarapu. Introducing l2m3, a multilingual medical large language model to advance health equity in low-resource regions. *arXiv preprint arXiv:2404.08705*, 2024.

Vinicius G. Goecks and Nicholas R. Waytowich. Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. In *Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML)*, Honolulu, Hawaii, USA, 2023. URL https://openreview.net/forum?id=8Q7WLpjitU.

Andre R Goncalves, Jose Cadena Pico, Yeping Hu, David Schlessinger, John Greene, Liam O'suilleabhain, Heather Clancy, Michael Vollmer, Vincent Liu, Tom Bates, et al. Ai-enabled diagnostic prediction within electronic health records to enhance biosurveillance and early outbreak detection. *Medrxiv*, 2025.

Chenghua Gong, Rui Sun, Yuhao Zheng, Juyuan Zhang, Tianjun Gu, Liming Pan, and Linyuan Lv. Epillm: unlocking the potential of large language models in epidemic forecasting. *arXiv preprint arXiv:2505.12738*, 2025.

Caroline Govathson, Candice Chetty-Makkan, Ross Greener, Sasha Frade, Dino Rech, Sarah Morris, Yohann Richard, Rouella Mendonca, Natalie Maricich, Lawrence Long, et al. Breaking barriers: harnessing artificial intelligence for a stigma-free, efficient hiv prevention assessment among adults in south africa. *Frontiers in Digital Health*, 7:1731002, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zifan Gu, Lesi He, Awais Naeem, Pui Man Chan, Asim Mohamed, Hafsa Khalil, Yujia Guo, Jingwei Huang, Ismael Villanueva-Miranda, Ying Ding, et al. Sbdh-reader: a large language model-powered method for extracting social and behavioral determinants of health from clinical notes. *Journal of the American Medical Informatics Association*, 32(10): 1570–1580, 2025.

Lucinda Gullison and Feng Fu. Working with large language models to enhance messaging effectiveness for vaccine confidence. *arXiv preprint arXiv:2504.09857*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Batool Haider, Atmika Gorti, Aman Chadha, and Manas Gaur. Mental health equity in llms: Leveraging multi-hop question answering to detect amplified and silenced perspectives. *arXiv preprint arXiv:2506.18116*, 2025.

Joe B Hakim, Jeffery L Painter, Darmendra Ramcharran, Vijay Kara, Greg Powell, Paulina Sobczak, Chiho Sato, Andrew Bate, and Andrew Beam. The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings. *Scientific Reports*, 15(1):27886, 2025.

Eileen Han, Miao Feng, and Pamela Ling. Building an analytical framework for tobacco-related information on social media: an exploratory analysis with generative ai assistance. *BMC Public Health*, 25(1):3635, 2025.

Joshua Harris, Fan Grayson, Felix Feldman, Timothy Laurence, Toby Nonnenmacher, Oliver Higgins, Leo Loman, Selina Patel, Thomas Finnie, Samuel Collins, et al. Healthy llms? benchmarking llm knowledge of uk government public health information. *arXiv preprint arXiv:2505.06046*, 2025.

Georges Hattab, Christopher Irrgang, Nils Körber, Denise Kühnert, and Katharina Ladewig. The way forward to embrace artificial intelligence in public health, 2025.

Jaeff Hong, Duong Dung, Danielle Hutchinson, Zubair Akhtar, Rosalie Chen, Rebecca Dawson, Aditya Joshi, Samsung Lim, C Raina MacIntyre, and Deepti Gurdasani. Relation extraction from news articles (rena): A tool for epidemic surveillance. *arXiv preprint arXiv:2311.01472*, 2023.

Zhiyuan Hou, Zhengdong Wu, Zhiqiang Qu, Liubing Gong, Hui Peng, Mark Jit, Heidi J Larson, Joseph T Wu, and Leesa Lin. A vaccine chatbot intervention for parents to improve hpv vaccination uptake among middle school girls: a cluster randomized trial. *Nature Medicine*, 31(6):1855–1862, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jingyi Huang, Yuyi Yang, Mengmeng Ji, Charles Alba, Sheng Zhang, and Ruopeng An. Use of retrieval-augmented large language model agent for long-form covid-19 fact-checking. *arXiv preprint arXiv:2512.00007*, 2025a.

Weihong Huang, Wudi Wei, Xiaotao He, Baili Zhan, Xiaoting Xie, Meng Zhang, Shiyi Lai, Zongxiang Yuan, Jingzhen Lai, Rongfeng Chen, et al. Chatgpt-assisted deep learning models for influenza-like illness prediction in mainland china: time series analysis. *Journal of Medical Internet Research*, 27:e74423, 2025b.

Hilton Humphries, Lindani Msimango, Zimasa Tshawe, Natasha Gcelu, Kurt Ferreira, Jacqueline Pienaar, Elise M van der Elst, Danielle Giovenco, Don Operario, Eduard J Sanders, et al. A qualitative study assessing the acceptability of a multi-agent ai chatbot for providing hiv and mental health support among men who have sex with men and transgender women in kwazulu-natal, south africa. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 120(2):160–174, 2026.

Ayana Hussain, Patrick Zhao, and Nicholas Vincent. An audit and analysis of llm-assisted health misinformation jailbreaks against llms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 1290–1301, 2025.

Jumpei Ito, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras, and Kei Sato. A protein language model for exploring viral fitness landscapes. *Nature communications*, 16(1):4236, 2025.

Saidah Zahrotul Jannah, Elyanah Aco, Shaowen Peng, Shoko Wakamiya, and Eiji Aramaki. Multilingual symptom detection on social media: enhancing health-related fact-checking with llms. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 54–68, 2025.

Yuelyu Ji, Wenhe Ma, Sonish Sivarajkumar, Hang Zhang, Eugene M Sadhu, Zhuochun Li, Xizhi Wu, Shyam Visweswaran, and Yanshan Wang. Mitigating the risk of health inequity exacerbated by large language models. *npj Digital Medicine*, 8(1):246, 2025.

Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–16, 2023.

Jeena Joseph, Binny Jose, and Jobin Jose. The generative illusion: how chatgpt-like ai tools could reinforce misinformation and mistrust in public health communication. *Frontiers in Public Health*, 13:1683498, 2025.

Suprabhath Kalahasti, Benjamin Faucher, Boxuan Wang, Claudio Ascione, Ricardo Carbajal, Maxime Enault, Christophe Vincent Cassis, Titouan Launay, Caroline Guerrisi, Pierre-Yves Boëlle, et al. Foundation time series models for forecasting and policy evaluation in infectious disease epidemics. *medRxiv*, pp. 2025–02, 2025.

Jasleen Kaur and Zahid Ahmad Butt. Ai-driven epidemic intelligence: the future of outbreak detection and response. *Frontiers in Artificial Intelligence*, 8:1645467, 2025.

Vipina K Keloth, Salih Selek, Qingyu Chen, Christopher Gilman, Sunyang Fu, Yifang Dang, Xinghan Chen, Xinyue Hu, Yujia Zhou, Huan He, et al. Social determinants of health extraction from clinical notes across institutions using large language models. *npj Digital Medicine*, 8(1):287, 2025.

Sedigh Khademi, Christopher Palmer, Gerardo Luis Dimaguila, Muhammad Javed, and Jim Buttery. Exploring large language models for detecting online vaccine reactions. In *Health. Innovation. Community: It Starts With Us*, pp. 30–35. IOS Press, 2024.

Sedigh Khademi, Jim Black, Christopher Palmer, Muhammad Javed, Hazel Clothier, Jim Buttery, and Gerardo Luis Dimaguila. Enhancing vaccine safety surveillance: Extracting vaccine mentions from emergency department triage notes using fine-tuned large language models. *arXiv preprint arXiv:2507.07599*, 2025.

Kwanho Kim and Soojong Kim. Large language models' accuracy in emulating human experts' evaluation of public sentiments about heated tobacco products on social media: evaluation study. *Journal of Medical Internet Research*, 27:e63631, 2025.

Soojong Kim, Kwanho Kim, and Claire Wonjeong Jo. Accuracy of a large language model in distinguishing anti-and pro-vaccination messages on social media: The case of human papillomavirus vaccination. *Preventive Medicine Reports*, 42:102723, 2024.

Ondrej Klempir, Ladislav Dusek, Radim Krupicka, Gleb Donin, Jan Zigmond, Radka Storchova, and Ales Tichopad. Evaluating large language models for natural-language-to-code generation on aggregate czech public health data analysis. *medRxiv*, pp. 2025–12, 2025.

Kin On Kwok, Tom Huynh, Wan In Wei, Samuel YS Wong, Steven Riley, and Arthur Tang. Utilizing large language models in infectious disease transmission modelling for public health preparedness. *Computational and structural biotechnology journal*, 23:3254–3257, 2024.

Alfu Laily, Laura M Schwab-Reese, Megan Davish, Emily Cahue, Kathryn J LaRoche, Natalia M Rodriguez, Robert J Duncan, Randolph D Hubach, and Monica L Kasting. Examining artificial intelligence chatbots' responses in providing human papillomavirus vaccine information for young adults: Qualitative content analysis. *JMIR Public Health and Surveillance*, 12:e79720, 2026.

Max SY Lau, C Jessica E Metcalf, Zewen Liu, Bryan T Grenfell, and Wei Jin. Toward ai foundation models for epidemics: Promise, challenges, and paths forward. *Proceedings of the National Academy of Sciences*, 123(13):e2526192123, 2026.

Timothy Laurence, Joshua Harris, Leo Loman, Amy Douglas, Yung-Wai Chan, Luke Hounsome, Lesley Larkin, and Michael Borowitz. Review gide–restaurant review gastrointestinal illness detection and extraction with large language models. *arXiv preprint arXiv:2503.09743*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Chenxiang Li, Qiqiao Zhang, Yue Zhang, Bowen Zhao, Jule Yang, Li Qi, Jun Ding, and Dechao Tian. Fine-tuned large language models enhance influenza forecasting. *medRxiv*, pp. 2025–03, 2025a.

Hai Li, Jingyi Huang, Mengmeng Ji, Yuyi Yang, and Ruopeng An. Use of retrieval-augmented large language model for covid-19 fact-checking: development and usability study. *Journal of medical Internet research*, 27:e66098, 2025b.

Yiming Li, Jianfu Li, Jianping He, and Cui Tao. Ae-gpt: using large language models to extract adverse events from surveillance reports-a use case with influenza vaccine adverse events. *Plos one*, 19(3):e0300919, 2024.

Yiming Li, Deepthi Viswaroopan, William He, Jianfu Li, Xu Zuo, Hua Xu, and Cui Tao. Enhancing relation extraction for covid-19 vaccine shot-adverse event associations with large language models. *Research Square*, pp. rs–3, 2025c.

Zongjing Liang, Gongcheng Liang, Yun Kuang, Zhijie Li, and Kuang Yun. Application and comparative study of generative artificial intelligence for epidemic prediction of coronavirus disease. *Cureus*, 17(8), 2025.

Junyu Liu, Qian Niu, Momoko Nagai-Tanima, and Tomoki Aoyama. Understanding human papillomavirus vaccination hesitancy in japan using social media: content analysis. *Journal of Medical Internet Research*, 27:e68881, 2025a.

Ollie Liu, Sami Jaghouar, Johannes Hagemann, Shangshang Wang, Jason Wiemels, Jeff Kaufman, and Willie Neiswanger. Metagene-1: Metagenomic foundation model for pandemic monitoring. *arXiv preprint arXiv:2501.02045*, 2025b.

Siying Liu, Shisheng Zhang, and Indu Bala. Robust or suggestible? exploring non-clinical induction in llm drug-safety decisions. *arXiv preprint arXiv:2510.13931*, 2025c.

Xiaoyu Liu, Lu He, Eman Alanazi, Echu Liu, Arianna Goss, and Lionel Gumireddy. Assessing the accuracy and explainability of using chatgpt to evaluate the quality of health news. *BMC Public Health*, 25(1):2038, 2025d.

Yuqi Liu, Jing Li, Peihan Li, Yehong Yang, Kaiying Wang, Jinhui Li, Lang Yang, Jiangfeng Liu, Leili Jia, Aiping Wu, et al. Arnle model identifies prevalence potential of sars-cov-2 variants. *Nature Machine Intelligence*, 7(1):18–28, 2025e.

Zewen Liu, Guancheng Wan, B Aditya Prakash, Max SY Lau, and Wei Jin. A review of graph neural networks in epidemic modeling. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6577–6587, 2024.

Zewen Liu, Juntong Ni, Max SY Lau, and Wei Jin. Pre-training epidemic time series forecasters with compartmental prototypes. *arXiv preprint arXiv:2502.03393*, 2025f.

Linqi Lu, Yanshu Sybil Wang, Jiawei Liu, and Douglas M McLeod. Human–generative ai interactions and their effects on beliefs about health issues: Content analysis and experiment. *JMIR AI*, 5(1):e80270, 2026.

Samira Malek, Christopher Griffin, Robert D Fraleigh, Robert Lennon, Vishal Monga, and Lijiang Shen. Intervention in health misinformation using large language models for automated detection, thematic analysis, and inoculation: Case study on covid-19. *Journal of medical Internet research*, 28:e75500, 2026.

Kangkun Mao, Fang Xu, Jinru Ding, Yidong Jiang, Yujun Yao, Yirong Chen, Junming Liu, Xiaoqin Wu, Qian Wu, Xiaoyan Huang, et al. Epiplanagent: Agentic automated epidemic response planning. *arXiv preprint arXiv:2512.10313*, 2025.

Sarah Martinson, Lingkai Kong, Cheol Woo Kim, Aparna Taneja, and Milind Tambe. Llm-based agent simulation for maternal health interventions: uncertainty estimation and decision-focused evaluation. *arXiv preprint arXiv:2503.22719*, 2025.

Andrew J McMurry, Dylan Phelan, Brian E Dixon, Alon Geva, Daniel Gottlieb, James R Jones, Michael Terry, David E Taylor, Hannah Callaway, Sneha Manoharan, et al. Large language model symptom identification from clinical text: Multicenter study. *Journal of medical Internet research*, 27:e72984, 2025.

Vaishnavi Menon, Natnael Shimelash, Samuel Rutunda, Cyprien Nshimiyimana, Lucinda Archer, Mira Emmanuel-Fabula, Derbew Fikadu Berhe, Jaspret Gill, Emery Hezagira, Eric Remera, et al. Assessing the potential utility of large language models for assisting community health workers: protocol for a prospective, observational study in rwanda. *BMJ open*, 15(10):e110927, 2025.

Shravika Mittal, Hayoung Jung, Mai ElSherief, Tanushree Mitra, and Munmun De Choudhury. Online myths on opioid use disorder: A comparison of reddit and large language model. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 1224–1245, 2025.

Natansh D Modi, Bradley D Menz, Abdulhalim A Awaty, Cyril A Alex, Jessica M Logan, Ross A McKinnon, Andrew Rowland, Stephen Bacchi, Kacper Gradon, Michael J Sorich, et al. Assessing the system-instruction vulnerabilities of large language models to malicious conversion into health disinformation chatbots. *Annals of internal medicine*, 178(8): 1172–1180, 2025.

Jaeuk Moon, Jonghwa Shim, Eunbeen Kim, and Eenjun Hwang. Miflu: large language model-based multimodal influenza forecasting scheme. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Plinio P Morita, Matheus Lotto, Jasleen Kaur, Dmytro Chumachenko, Arlene Oetomo, Kristopher Dylan Espiritu, and Irfhana Zakir Hussain. What is the impact of artificial intelligence-based chatbots on infodemic management? *Frontiers in public health*, 12: 1310437, 2024.

Yida Mu, Mali Jin, Kalina Bontcheva, and Xingyi Song. Examining temporalities on stance detection towards covid-19 vaccination. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6732–6738, 2024.

Jacob Muller, Daniel Petti, Changying Li, Serap Gorucu, Matthew Pilz, and Bryan P Weichelt. Large language models for agricultural injury surveillance. *Safety*, 11(1):15, 2025.

Navapat Nananukul and Mayank Kejriwal. A large language model-based approach for analyzing covariates of health equity in registered research projects. *medRxiv*, pp. 2024–09, 2024.

Aditya Narayan, Michael Blasingame, Sam Warmuth, Gabriella Palmeri, India Halm, Ramin Bastani, Whitney Engeran-Cordova, Harold J Phillips, Leandro Mena, and Nirav R Shah. Ai-augmented communication improves hiv prep initiation and persistence in populations disproportionately impacted by hiv. *npj Digital Medicine*, 2026.

Yoo Jung Oh, Muhammad Ehab Rasul, Emily McKinley, and Christopher Calabrese. From digital traces to public vaccination behaviors: leveraging large language models for big data classification. *Frontiers in Artificial Intelligence*, 8:1602984, 2025.

Tobi Olatunji, Charles Nimo, Abraham Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Chinemelu Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, et al. Afrimed-qa: a pan-african, multi-specialty, medical question-answering benchmark dataset. *arXiv preprint arXiv:2411.15640*, 2024.

Elise Omaki, Felipe Restrepo, Wendy C Shields, and Alan Abrahams. Natural language processing tool for extracting information about opioid overdoses in the usa from case narratives in the violent death reporting system. *Injury Prevention*, 2025.

Jasmine Chiat Ling Ong, Benjamin Jun Jie Seng, Jeren Zheng Feng Law, Lian Leng Low, Andrea Lay Hoon Kwa, Kathleen M Giacomini, and Daniel Shu Wei Ting. Artificial intelligence, chatgpt, and other large language models for social determinants of health: Current state and future directions. *Cell Reports Medicine*, 5(1), 2024.

Jasmine Chiat Ling Ong, Yilin Ning, Rui Yang, Danielle S Bitterman, Xiaoxuan Liu, Yih Chung Tham, Gary S Collins, Michelle María Jiménez de Tavárez, Bilal A Mateen, Kwesi Nyan Amissah-Arthur, et al. Large language models in global health. *Nature Health*, 1(1):35–47, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Jeffery L Painter, Venkateswara Rao Chalamalasetti, Raymond Kassekert, and Andrew Bate. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language. *JAMIA open*, 8(1):ooaf003, 2025.

Jie Pan, Seungwon Lee, Cheligeer Cheligeer, Elliot A Martin, Kiarash Riazi, Hude Quan, and Na Li. Integrating large language models with human expertise for disease detection in electronic health records. *Computers in Biology and Medicine*, 191:110161, 2025.

Madhurima Panja, Ojas Modak, Grace Younes, and Tanujit Chakraborty. Zero-shot forecasting of epidemics. In *Recent Advances in Time Series Foundation Models Have We Reached the'BERT Moment'?*, 2025.

Devesh Pant, Rishi Raj Grandhe, Jatin Agrawal, Jushaan Singh Kalra, Sudhir Kumar, Saransh Khanna, Vipin Samaria, Mukul Paul, Satish V Khalikar, Vipin Garg, et al. Health sentinel: An ai pipeline for real-time disease outbreak detection. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pp. 23–42, 2025.

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

Susan T Parker. Supervised natural language processing classification of violent death narratives: Development and assessment of a compact large language model. *JMIR AI*, 4: e68212, 2025.

Jasmin Perret and Adrian Schmid. Application of openai gpt-4 for the retrospective detection of catheter-associated urinary tract infections in a fictitious and curated patient data set. *Infection Control & Hospital Epidemiology*, 45(1):96–99, 2024.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12):3590–3600, 2024.

Emma Pierson, Divya Shanmugam, Rajiv Movva, Jon Kleinberg, Monica Agrawal, Mark Dredze, Kadija Ferryman, Judy Wawira Gichoya, Dan Jurafsky, Pang Wei Koh, et al. Using large language models to promote health equity, 2025.

Joshua L Proctor and Guillaume Chabot-Couture. Democratizing infectious disease modeling: an ai assistant for generating, simulating, and analyzing dynamic models. *medRxiv*, pp. 2024–07, 2024.

Ashley Quigley, Mr Damian Honeyman, Ms Haley Stone, Rebecca Dawson, and C Raina MacIntyre. Epiwatch, an artificial intelligence early-warning system as a valuable tool in outbreak surveillance. *International Journal of Infectious Diseases*, 152:107579, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Shagoto Rahman, Cornelia Pechmann, and Ian G Harris. Enhancing detection of message intents in a mobile health smoking-cessation intervention using large language model fine-tuning, data downsampling, and error correction: Algorithm development and validation. *Journal of Medical Internet Research*, 28:e83437, 2026.

Pragnya Ramjee, Mehak Chhokar, Bhuvan Sachdeva, Mahendra Meena, Hamid Abdullah, Aditya Vashistha, Ruchit Nagar, and Mohit Jain. Ashabot: An llm-powered chatbot to support the informational needs of community health workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2025.

Florian Reis, Lea J Bayer, Claudius Malerczyk, Christian Lenz, and Christof von Eiff. Leveraging large language models to address common vaccination myths and misconceptions. *medRxiv*, pp. 2026–02, 2026.

Alberto Rizzo, Enrico Mensa, and Andrea Giacomelli. The future of large language models in fighting emerging outbreaks: lights and shadows. *The Lancet Microbe*, 5(11), 2024.

Tracy R Rosebrock, Zhen Yang, Lauren D'Arco, Tapan Pathak, Rebecca Vislay-Wade, Karen Fowler, John Diaz-Decaro, and Colin Kunzweiler. Using artificial intelligence methods to evaluate the effect of the national cytomegalovirus awareness month on the content and sentiment of social media posts: Infodemiology study. *JMIR infodemiology*, 6:e80922, 2026.

Melika Rostami and Suliman Hawamdeh. Debunk lists as external knowledge structures for health misinformation detection with generative ai. *Systems*, 13(10):882, 2025.

Hamid Reza Saeidnia, Shamim Jahani, Nasrin Ghiasi, and Hamid Keshavarz. Generative ai and health misinformation: production, propagation, and mitigation—a systematic review. *BMC Public Health*, 2026.

Mohammad Hosseini Samaei, Faryad Darabi Sahneh, Lee W Cohnstaedt, and Caterina M Scoglio. Epidemiqs: Prompt-to-paper llm agents for epidemic modeling and analysis. *IEEE Transactions on Artificial Intelligence*, 2026.

Niamh Sheridan et al. ILI surveillance from Twitter in Wales. In *Proceedings of the 10th Workshop on Noisy and User-generated Text (W-NUT)*, 2025. URL https://aclanthology.org/2025.wnut-1.1/.

Ziyi Shi, Xusen Guo, Hongliang Lu, Mingxing Peng, Haotian Wang, Zheng Zhu, Zhenning Li, Yuxuan Liang, Xinhu Zheng, and Hai Yang. Coordinated pandemic control with large language model agents as policymaking assistants. *arXiv preprint arXiv:2601.09264*, 2026.

Natnael Shimelash, Samuel Rutunda, Vaishnavi Menon, Mira Emmanuel-Fabula, Angel Uwimbabazi, Crystal Rugege, Cyprien Nshimiyimana, Ivan Rwema, Mouna Kandekwe, Derbew Fikadu Berhe, et al. A 'silent trial'assessing the accuracy of large language models for assisting community health workers in low-resource settings. *medRxiv*, pp. 2026–02, 2026.

Yafei Si, Yuyi Yang, Xi Wang, Jiaqi Zu, Xi Chen, Xiaojing Fan, Ruopeng An, and Sen Gong. Quality and accountability of chatgpt in health care in low-and middle-income countries: simulated patient study. *Journal of medical Internet research*, 26:e56121, 2024.

Grigori Sidorov, Muhammad Ahmad, Pierpaolo Basile, Muhammad Waqas, Rita Orji, and Ildar Batyrshin. Monitoring opioid-related social media chatter using natural language processing and large language models: Temporal analysis. *JMIR infodemiology*, 5(1):e77279, 2025.

Zalaya Simmons, Beti Evans, Tamsyn Harris, Harry Woolnough, Lauren Dunn, Jonathon Fuller, Kerry Cella, and Daphne Duval. Assessing the feasibility and acceptability of a bespoke large language model pipeline to extract data from different study designs for public health evidence reviews. *Cochrane Evidence Synthesis and Methods*, 3(6):e70061, 2025.

Nina Singh, Katharine Lawrence, Safiya Richardson, and Devin M Mann. Centering health equity in large language model deployment. *PLOS Digital Health*, 2(10):e0000367, 2023.

Xiaoying Song, Anirban Saha Anik, Dibakar Barua, Pengcheng Luo, Junhua Ding, and Lingzi Hong. Speaking at the right level: Literacy-controlled counterspeech generation with RAG-RL. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2812–2830, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.153. URL https://aclanthology.org/2025.findings-emnlp.153/.

Chayma Sridi and Salem Brigui. The use of chatgpt in occupational medicine: opportunities and threats. *Annals of occupational and environmental medicine*, 35:e42, 2023.

Dragan Stoll, Samuel Wehrli, and David Lätsch. Case reports unlocked: Harnessing large language models to advance research on child maltreatment. *Child Abuse & Neglect*, 160: 107202, 2025.

Rickard Stureborg, Sanxing Chen, Roy Xie, Aayushi Patel, Christopher Li, Chloe Zhu, Tingnan Hu, Jun Yang, and Bhuwan Dhingra. Tailoring vaccine messaging with common-ground opinions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2553–2575, 2024.

Jun Tao, Ellie Pavlick, Amaris Grondin, Josue D Bustamante, Harrison Martin, Hannah Parent, Natalie Fenn, Alexi Almonte, Amanda Maguire-Wilkerson, Mofan Gu, et al. Evaluation of an artificial intelligence conversational chatbot to enhance hiv preexposure prophylaxis uptake: Development and usability internal testing. *Journal of Medical Internet Research*, 28:e79671, 2026.

Conrad Testagrose, Sakshi Pandey, Mohammadali Serajian, Simone Marini, Mattia Prosperi, and Christina Boucher. Leveraging large language models to predict antibiotic resistance in mycobacterium tuberculosis. *Bioinformatics*, 41(Supplement_1):i40–i48, 2025.

Anushree Tiwari, Amit Kumar, Shailesh Jain, Kanika S Dhull, Arunkumar Sajjanar, Rahul Puthenkandathil, Kapil Paiwal, Ramanpal Singh, and Arun Sajjanar. Implications of chatgpt in public health dentistry: A systematic review. *Cureus*, 15(6), 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Albert Jan van Hoek, Sebastian Funk, Stefan Flasche, Billy J Quilty, Esther van Kleef, Anton Camacho, and Adam J Kucharski. Importance of investing time and money in integrating large language model-based agents into outbreak analytics pipelines. *The Lancet Microbe*, 5(8), 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Shresth Verma, Alayna Nguyen, Niclas Boehmer, Lingkai Kong, and Milind Tambe. Priority2reward: Incorporating healthworker preferences for resource allocation planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29709–29711, 2025.

Michiel Vos, Markus Göker, Richard Bendall, and Fabrizio Costa. Large language model-assisted text mining reveals bacterial pathogen diversity. *bioRxiv*, pp. 2025–07, 2025.

Song Wang, Yishu Wei, Haotian Ma, Max Lovitt, Kelly Deng, Yuan Meng, Zihan Xu, Jingze Zhang, Yunyu Xiao, Ying Ding, et al. A multi-stage large language model framework for extracting suicide-related social determinants of health. *Communications Medicine*, 5(1): 404, 2025a.

Yuanlong Wang, Pengqi Wang, Changchang Yin, and Ping Zhang. Sathealth: A multimodal public health dataset with satellite-based environmental factors. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5819–5830, 2025b.

Aniket Wattamwar and Sampson Akwafuo. Aries: A scalable multi-agent orchestration framework for real-time epidemiological surveillance and outbreak monitoring. *arXiv preprint arXiv:2601.01831*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Mingyang Wei, Dehai Min, Zewen Liu, Yuzhang Xie, Guanchen Wu, Carl Yang, Max SY Lau, Qi He, Lu Cheng, and Wei Jin. Epiqal: Benchmarking large language models in epidemiological question answering for enhanced alignment and reasoning. *arXiv preprint arXiv:2601.03471*, 2026.

Richard Wen, Anteneh Tesfaye Assalif, Andy Sze-Heng Lee, Rajeev Kamadod, Asha Behdinan, Ronald Carshon-Marsh, Catherine Meh, Thomas Kai Sze Ng, Patrick Brown, Prabhat Jha, et al. Computer assisted verbal autopsy: comparing large language models to physicians for assigning causes to 6939 deaths in sierra leone from 2019–2022. *BMC medicine*, 24(1):49, 2026.

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*, 2023.

Rory Wilson, Ciara M Weets, Amanda Rosner, and Rebecca Katz. Evaluating generative artificial intelligence's limitations in health policy identification and interpretation. *PLoS One*, 19(12):e0312078, 2024.

Jiaying Wu, Zihang Fu, Haonan Wang, Fanxiao Li, Jiafeng Guo, Preslav Nakov, and Min-Yen Kan. Beyond the crowd: Llm-augmented community notes for governing health misinformation. *arXiv preprint arXiv:2510.11423*, 2025a.

Julie T Wu, Bradley J Langford, Erica S Shenoy, Evan Carey, and Westyn Branch-Elliman. Chatting new territory: large language models for infection surveillance from pilot to deployment. *Infection Control & Hospital Epidemiology*, 46(3):224–226, 2025b.

Dengke Xia, Mengyao Song, and Tingshao Zhu. A comparison of the persuasiveness of human and chatgpt generated pro-vaccine messages for hpv. *Frontiers in public health*, 12: 1515871, 2025.

Jiacheng Xie, Ziyang Zhang, Shuai Zeng, Joel Hilliard, Guanghui An, Xiaoting Tang, Lei Jiang, Yang Yu, Xiufeng Wan, Dong Xu, et al. Leveraging large language models for infectious disease surveillance—using a web service for monitoring covid-19 patterns from self-reporting tweets: Content analysis. *Journal of Medical Internet Research*, 27(1): e63190, 2025.

Fengyi Xu, Jun Ma, Nan Li, and Jack CP Cheng. Large language model applications in disaster management: An interdisciplinary review. *International Journal of Disaster Risk Reduction*, 127:105642, 2025a.

Gelei Xu, Xueyang Li, Yixiong Chen, Yuying Duan, Shuqing Wu, Haoxinran Yu, Ching-Hao Chiu, Juntong Ni, Ningzhi Tang, Toby Jia-Jun Li, Alan Yuille, Wei Jin, and Yiyu Shi. A comprehensive survey of ai agents in healthcare. *Journal of Biomedical Informatics*, 179: 105045, 2026. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2026.105045. URL https://www.sciencedirect.com/science/article/pii/S1532046426000699.

Jiaxiang Xu, Zhengdong Wu, Lily Wass, Heidi J Larson, and Leesa Lin. Mapping global public perspectives on mrna vaccines and therapeutics. *npj Vaccines*, 9(1):218, 2024a.

Shan Xu, Zhaokun Yan, Chengxiao Dai, and Fan Wu. Mega-rag: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of llms in public health. *Frontiers in Public Health*, 13:1635381, 2025b.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(1):1–32, 2024b.

Chenghao Yang, Tuhin Chakrabarty, Karli Hochstatter, Melissa Slavin, Nabila El-Bassel, and Smaranda Muresan. Identifying self-disclosures of use, misuse and addiction in community-based social media posts. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2507–2521, 2024.

Kishwen K Yoga Ratnam. Generative artificial intelligence in public health research and scientific communication: A narrative review of real applications and future directions. *Digital Health*, 11:20552076251362070, 2025.

Zhihao Zhang, Yiran Zhang, Xiyue Zhou, Liting Huang, Imran Razzak, Preslav Nakov, and Usman Naseem. From generation to detection: A multimodal multi-task dataset for benchmarking health misinformation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24245–24260, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1316. URL https://aclanthology.org/2025.findings-emnlp.1316/.

Jiawei Zhou, Amy Z Chen, Darshi Shah, Laura M Schwab-Reese, and Munmun De Choudhury. A risk taxonomy and reflection tool for large language model adoption in public health. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–32, 2025a.

Xinyu Zhou, Jiaqi Zhou, Chiyu Wang, Qianqian Xie, Kaize Ding, Chengsheng Mao, Yuntian Liu, Zhiyuan Cao, Huangrui Chu, Xi Chen, et al. Ph-llm: public health large language models for infoveillance. *medRxiv*, 2025b.

Angelo Ziletti and Leonardo DAmbrosi. Retrieval augmented text-to-sql generation for epidemiological question answering using electronic health records. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pp. 47–53, 2024.

Ruohan Zong, Yang Zhang, and Dong Wang. Empowering llms to synthesize ai and human intelligence for explainable public health misinformation detection on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 2334–2348, 2025.

A Appendix

A.1 Additional background on large language models

Large language models: Additional background for PH professionals. A large language model (LLM) is a neural network with up to hundreds of billions of parameters, trained on large text corpora to predict and generate language. Modern LLMs are built on the **transformer** architecture (Vaswani et al., 2017), whose self-attention mechanism captures long-range dependencies and is well suited to public health tasks involving lengthy guidelines, surveillance reports, and heterogeneous evidence. Most systems in this survey begin with a **foundation model** (Bommasani et al., 2021) and adapt it to a specific public health task. Two main pre-training paradigms dominate: autoregressive models, such as the GPT family, predict the next token and are especially effective for generation and dialogue (Brown et al., 2020; Achiam et al., 2023), while masked language models such as BERT reconstruct missing tokens and are often used for classification and extraction (Devlin et al., 2019); encoder–decoder models such as T5 combine both directions (Raffel et al., 2020). Open-weight families such as LLaMA also matter for reproducibility and deployment in resource-constrained settings (Touvron et al., 2023). Since pre-training alone does not produce models that reliably follow instructions or avoid harmful outputs, **alignment** is typically added through supervised fine-tuning and reinforcement learning from human feedback (Ouyang et al., 2022). In practice, public health applications mainly rely on four adaptation strategies: prompting, including few-shot and chain-of-thought prompting (Brown et al., 2020; Wei et al., 2022); fine-tuning, including parameter-efficient methods such as LoRA (Hu et al., 2022); retrieval-augmented generation, which grounds responses in external documents (Lewis et al., 2020); and **agentic systems**, which connect LLMs to tools to execute more complex workflows.

A.2 Survey methodology and literature selection

Search strategy. We conducted a systematic literature search following PRISMA guidelines. We queried five major academic databases—PubMed, IEEE Xplore, ACM Digital Library,

Google Scholar, and Semantic Scholar—as well as preprint servers (arXiv, medRxiv) for papers published between January 2023 and early 2026. The search strategy combined two categories of keywords: model-centric terms (*Large Language Model, LLM, Foundation Model, GPT, Generative AI*) and domain-centric terms (*Public Health, Epidemiology, Surveillance, Pharmacovigilance, Infodemiology, Health Equity, Verbal Autopsy, Wastewater Surveillance*). To ensure comprehensive coverage, we supplemented the database queries with targeted venue-specific searches across both AI conferences (ICML, ICLR, NeurIPS, ACL, EMNLP, NAACL, AAAI, CHIL, ML4H, FAccT) and public health journals (Nature Medicine, The Lancet, AJPH, IJE, JMIR, BMC Public Health, Frontiers in Public Health). Forward and backward citation tracking was performed on key seed papers to capture additional relevant work. Our initial search yielded 281 candidate papers after deduplication.

Screening and boundary test. Because the boundary between individual-level clinical AI and population-level public health is often blurred, we applied a three-step boundary test to every candidate paper:

1. **Institutional attribution test:** Does the work's primary application reside within the public health system (e.g., CDC, WHO, health departments, schools of public health, epidemiological registries)?
2. **Core function test:** Does the core task fall within one of the recognized core public health functions—surveillance, disease prevention, health communication, health equity, population intervention, or governance—as defined by established frameworks such as the CDC's 10 Essential Public Health Services?
3. **Substitution test:** If removed from the public health context, would the work retain independent significance in another domain? If yes, its primary identity is likely outside public health.

A paper must satisfy at least the first two criteria. We explicitly exclude individual-level clinical applications (diagnosis, treatment, prognosis), drug discovery, consumer-level health assistants, and medical education unless targeting public health training or community health worker capacity building. Screening was performed independently by two reviewers, with disagreements resolved through discussion. This process led to the exclusion of 35 candidate papers and the reclassification of 10 borderline cases (including individual-level mental health prediction systems and personalized ADR prediction tools). A subsequent scope audit removed an additional 10 papers, yielding a final corpus of 192 papers. The six-category task taxonomy (T1–T6) and five LLM roles were then inductively developed from this final corpus, as described in Section 4. All paper URLs and metadata were manually verified for correctness.

Data extraction. For the 192 included papers, we performed systematic data extraction from full text. For each paper, we recorded structured metadata including: primary PH task (T1–T6), LLM functional role (Sensor, Predictor, Communicator, Simulator, Auditor), technique tags (Prompting, Fine-tuning, RAG, Agentic, etc.), base LLM(s), model access (Open/Proprietary), data source type, geographic focus, language(s), evaluation design, and deployment stage. We additionally recorded narrative fields—including a one-line summary, key results, and the specific public health challenges addressed—to support the qualitative synthesis in each task section. Task-specific fields (e.g., surveillance modality for T1, prediction horizon for T2, platform analyzed for T3) were recorded where applicable.

A.3 The public health data ecosystem

Table 3 summarizes the data modalities, language distribution, and evaluation benchmarks identified across the 192 surveyed papers.

Unlike clinical AI, which primarily operates on structured EHRs and medical imaging, public health applications draw on a remarkably heterogeneous data landscape. Three patterns stand out. First, **social media dominates** (25.5%), reflecting the field's emphasis on real-time population-level signal extraction—yet nearly all studies use Twitter/X or Reddit, leaving platforms with large global user bases (TikTok, WhatsApp, WeChat) almost entirely unexplored. Second, **several modalities have no clinical parallel**: epidemiological

Data source category	% of papers
Social media (Twitter/X, Reddit, Weibo)	25.5
Scientific literature & guidelines	16.1
Simulated / agent-based data	11.5
Clinical text (EHR notes, discharge summaries)	9.9
Official reports (WHO DONs, CDC bulletins)	9.4
Epidemiological time-series	6.8
Death certificates & verbal autopsy narratives	4.2
Policy documents (e.g., OxCGRT)	3.6
Genomic / environmental (wastewater, sequences)	3.1
Consumer reviews	2.6
Surveys & questionnaires	7.3
Language coverage	
English-only	80.7
Monolingual non-English	14.1
Genuinely multilingual ($\geq$3 languages)	5.2

Table 3: Distribution of data modalities and language coverage across 192 surveyed papers. Percentages for data sources sum to more than 100% because some papers use multiple modalities.

time-series, verbal autopsy narratives, wastewater metagenomics, and government policy documents are unique to public health and require specialized tokenization, alignment, or constrained decoding strategies that standard clinical NLP does not address. Third, the **linguistic imbalance is severe**: 80.7% of papers operate exclusively on English data despite public health being an inherently global endeavor, and only PH-LLM systematically evaluates across 29 languages, revealing significant performance degradation for Southeast Asian languages.