An Agentic Framework for Auditable Public Health Policy Communication

Yansong Wen[1,2], Chengyun Fan[1,3], Jingya Yu[2], Yixuan Feng[2], Yueqi Jiang[1], Xiaodong An[6], Rebecka Rosenquist[2], Radha Pennotti[2], Aditi Vasan[2], Tara E Dechert[2], Stephanie Garcia[2], Jordan I Wood[2], Madison Langrin[2], Meredith Matone[2], Brian T Fisher[2], Susan Coffin[2], Jiasheng Shi[4], Lichao Sun[5], David Rubin[7], Kai Zhang[8*], Jing Huang[1,2*]

[1] University of Pennsylvania, Philadelphia, PA 19104, USA
[2] The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[3] The Hong Kong Polytechnic University, Hong Kong
[4] The Chinese University of Hong Kong, Shenzhen, Shenzhen, China
[5] Lehigh University, Bethlehem, PA 18015, USA
[6] Georgia Institute of Technology, Atlanta, GA 30332, USA
[7] University of California Office of the President, Oakland, CA 94607, USA
[8] Stanford University, Stanford, CA 94305, USA
**Corresponding authors*

Abstract

Public health policy communications translate disease surveillance data and evolving public health guidance into actionable recommendations for policymakers and communities. While recent advances in large language models (LLMs) have enabled a wide range of capabilities, including increasingly sophisticated text generation, complex data analysis, and information synthesis, their ability to generate accurate, accessible and evidence-based public health communications remains poorly understood. Real-world deployment requires outputs that are transparent, grounded in evidence, and readily auditable. We developed PolicyAgent, an agentic framework that integrates LLMs with planning, memory, retrieval, analysis and auditing tools to support the generation of evidence-grounded and auditable public health policy communications. We also developed PHPC-Bench, a benchmark comprising 61 domain-expert-written county-level COVID-19 policy communications paired with the corresponding surveillance data and CDC and WHO guidance available at the time of messaging. Across multiple LLMs, PolicyAgent matched or exceeded the expert-written communications on seven of nine quality dimensions, achieved expert-comparable alignment with public health guidance, and traced 94.9% of its quantitative claims to supporting evidence. PolicyAgent demonstrated robust performance across multiple LLM backbones, with stronger models improving communication fluency and numerical traceability. However, limitations in evidence selection, evidence weighting, and calibration of policy recommendations persisted even with state-of-the-art models. These findings suggest that current agentic AI systems can effectively support the drafting and auditing of public health policy communications while providing transparency into the evidence underlying generated content. Nevertheless, expert oversight remains essential for evaluating evidence relevance and determining the appropriate strength of policy recommendations.

Keywords: Agentic artificial intelligence; health policy communication; retrieval-augmented generation; factuality evaluation; large language models

Introduction

Effective public health policy communication is grounded in accurate, accessible, and evidence-based public health information.[1–3] By shaping how individuals and communities interpret risks, make decisions, and respond under uncertainty, public health communications can substantially influence the effectiveness of public health policies and interventions.[4,5] Conversely, shortcomings in communication may lead to misunderstanding of recommendations, increased uncertainty, reduced adherence to public health guidance, and erosion of trust in public health authorities.[6–8] Developing public health policy communications that are accurate, accessible, and evidence-based is challenging because it requires synthesizing complex and rapidly evolving evidence into information that is understandable, relevant, and actionable for diverse audiences.[9,10] During communicable disease outbreaks, communications often draw upon statistical and epidemiologic models that estimate disease transmission and evaluate the potential impact of interventions while accounting for local conditions and pathogen characteristics.[11,12] Translating these technical findings into public-facing messages requires substantial domain expertise, communication skills, and institutional resources, which may be limited in many public health agencies and local jurisdictions.[13,14]

Recent advances in artificial intelligence (AI), particularly large language models (LLMs), have enabled systems capable of synthesizing complex information, performing sophisticated data analyses, and generating high-quality natural language text.[15] These capabilities align closely with the demands of public health communication but are insufficient on their own to transform infectious disease surveillance and projection data into evidence-informed policy recommendations.[16,17] Such communication requires analyzing structured surveillance data, drawing situation-specific inferences, formulating evidence-informed recommendations, and considering trade-offs across multiple stakeholders.[5] Agentic AI systems extend the capabilities of LLMs through planning, memory, reasoning, and tool use, while multi-agent architectures enable specialized agents to collaborate on different components of this complex task.[15,18,19] To date, applications of AI in epidemic response have focused primarily on upstream tasks such as disease forecasting and intervention evaluation.[20,21] Comparatively little attention has been given to the downstream task of translating surveillance and projection data into public-facing communications that describe local risk and provide recommendations for households, schools, institutions, and communities. For example, a recent LLM multi-agent framework treats pandemic policy as a coordinated control problem, using regional agents to reallocate, suppress or screen inter-regional population inflows within a closed-loop epidemiological simulator.[22] Although this provides a testbed for counterfactual control, it leaves a downstream question: whether public-facing policy communications can faithfully represent local data, justify recommendations with evidence, and remain auditable against contemporaneous public health guidance. We hypothesize that an agentic system can assist in translating infectious disease surveillance and projection data into auditable public health communications that describe local risk and provide evidence-informed recommendations. A critical challenge, however, is ensuring that the generated communications remain faithful to the underlying data, grounded in evidence, and consistent with contemporaneous public health guidance.[23–26] Therefore, the trustworthiness of such systems must be assessed by evaluating whether their outputs are supported by expert judgment, public health evidence, and the source data from which the communications are derived.[17,27,28]

Here we introduce PolicyAgent (Fig. 1), an agentic framework designed to translate regional infectious disease surveillance and projection data into auditable public health policy communications. PolicyAgent integrates disease surveillance and projection data with local contextual information, including demographic, mobility, behavioral, health-system and news data relevant to the target communication date

and geographic region (Fig. 1a). The framework first synthesizes these heterogeneous data sources to characterize epidemic trends, evaluate national and regional disease activity, and assess local conditions relevant to public health decision-making (Fig. 1b1). A panel of specialized LLM agents representing epidemiologic, clinical, and public health management perspectives then deliberates on these situational assessments to formulate evidence-informed public health policy recommendations (Fig. 1b2). Finally, PolicyAgent generates a public health policy communication that summarizes the current epidemic situation, interprets surveillance and projection data in plain language, and provides policy guidance through an iterative drafting and review process designed to improve consistency with source data, public health evidence, and stakeholder perspectives.

To enable rigorous evaluation of AI-generated public health communications, we developed PHPC-Bench, a public health policy communication benchmark that comprises 61 expert-written COVID-19 policy blogs (May 20, 2020 – August 4, 2022),[29] paired with the corresponding county-level surveillance[12] and projection data[11] and an archived corpus of CDC and WHO guidance.[30,31] PHPC-Bench evaluates generated communications along three complementary dimensions: communication quality, assessed across nine dimensions using an automated grader calibrated against expert ratings; guidance alignment, measured by comparing recommended policies and public responses against contemporaneous CDC and WHO guidance using a retrieval-augmented[26] evaluation framework (Methods); and numerical accuracy, assessed by verifying communicated statistics against the underlying source data. Together, these evaluations assess whether generated communications are consistent with expert judgment, public health guidance, and source data.

We used PHPC-Bench to evaluate PolicyAgent and compare its performance with that of standalone LLMs. Our results revealed important limitations of single LLM generation, including weaker performance in emotional language use, numerical faithfulness, evidence grounding and consideration of multiple stakeholder perspectives. PolicyAgent substantially reduced these deficiencies. Its outputs were not found statistically different from expert-written communications on seven of nine expert-assessed quality dimensions, achieved an 88.8% agreement rate with CDC and WHO guidance (compared with 88.4% for expert-written blogs), and cited source statistics with 94.9% accuracy. These findings demonstrate the potential of agentic AI to support the drafting and verification of public health policy communications while highlighting evidence selection, attribution, and downstream audience evaluation as important areas for continued research and human oversight.

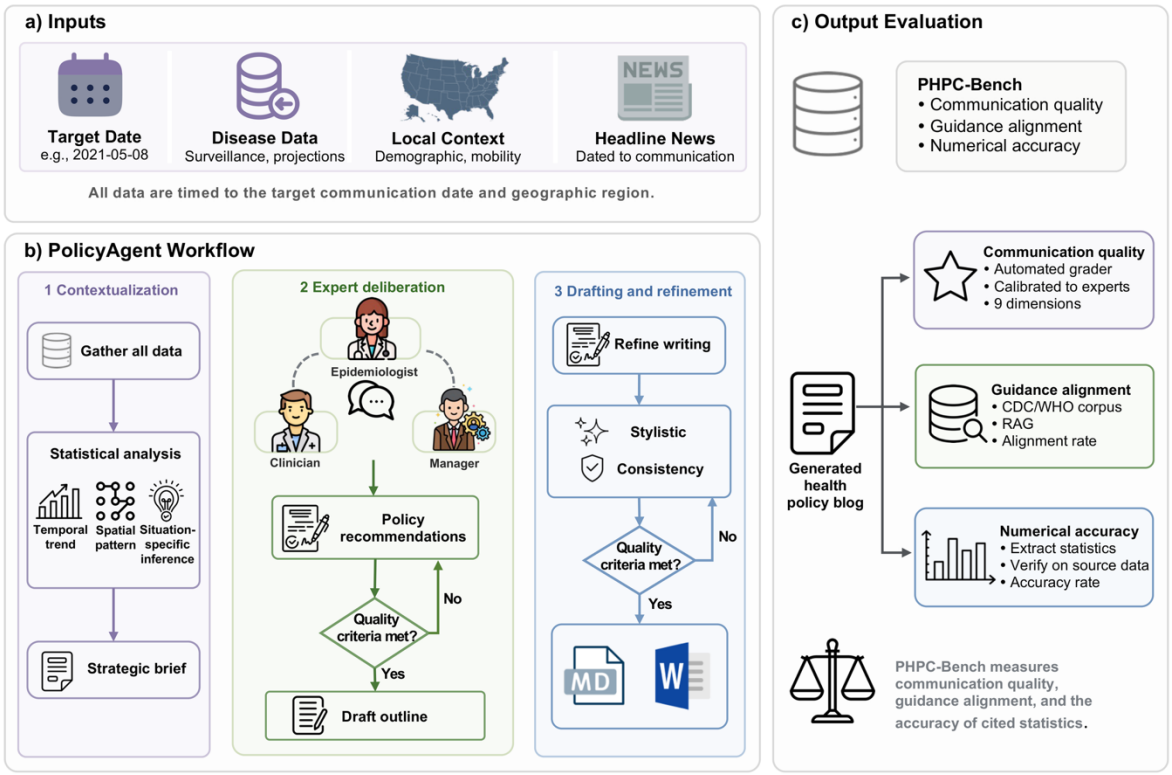


Figure1 | PolicyAgent framework and the PHPC-Bench. a, Inputs to PolicyAgent include disease surveillance and projection data, local contextual information (for example demographics, mobility, behavioral and health system characteristics), and headline news relevant to a target communication date. **b**, PolicyAgent operates through three stages. During situational assessment, surveillance, projection, and contextual data are synthesized to characterize epidemic trends, compare national and regional conditions, and assess local public health risks and needs. During expert deliberation, epidemiologist, clinician and public health manager agents evaluate the situation and formulate policy recommendations. These recommendations are incorporated into a public health policy communication that summarizes epidemic status, interprets surveillance and projection data, and provides policy guidance through an iterative review and refinement process. **c**, Generated public health policy communications are evaluated using PHPC-Bench across three independently auditable dimensions: communication quality, assessed using a human-aligned automated evaluator; guidance alignment, assessed against CDC and WHO guidance using retrieval-augmented evaluation; and numerical accuracy, assessed by verifying reported statistics against source surveillance and projection data.

Results

PolicyAgent and PHPC-Bench establish an auditable framework for public health policy communication

PolicyAgent is an agentic framework that translates infectious disease surveillance and projection data into public health policy communications (Fig. 1). The framework operates through a multi-stage workflow. First, it ingests county-level disease epidemiologic data, together with local demographic characteristic and contemporaneous news reports (Fig. 1a). It then performs statistical analyses to generate a situational assessment of local disease dynamics (Fig. 1b). Next, a panel of specialized agents, including epidemiologist, clinician and public health manager agents, deliberate on potential interventions and public health responses (Fig. 1b). The framework drafts a policy communication that undergoes iterative review, evidence verification, and stylistic refinement before release (Fig. 1b,c).

To evaluate PolicyAgent, we developed PHPC-Bench, a benchmark designed for time-aligned evaluation of public health policy communication (Fig. 1). PHPC-Bench pairs 61 expert-written COVID-19 policy communications between May 20, 2020 and August 04, 2022 with the disease surveillance and projection data and contemporaneous CDC and WHO guidance that were available at the time each communication was written (Fig. 2a, Extended Data Fig. 1). By reconstructing the evidentiary environment available to communication authors, the benchmark enables generated communications, source data, public health guidance, and expert-written references to be compared within a common temporal context.

PHPC-Bench evaluates three independently auditable dimensions: communication quality, assessed using a nine-dimension expert rubric (Fig. 2b), guidance alignment, assessed by comparing policy recommendations with contemporaneous CDC and WHO guidance, and numerical accuracy, assessed by verifying quantitative claims against the underlying surveillance and projection data. Together, these evaluations not only compare the quality of PolicyAgent-generated communications with expert-written benchmarks but also assess two auditable dimensions: consistency with contemporaneous guidance and traceability to source data.

PolicyAgent improves representational breadth but remains limited in evidence quality

The nine dimension communication quality rubric spans three constructs: scientific soundness, decision relevance, and representational breadth (Fig. 2b). Rubric scores were generated using a calibrated LLM-as-judge whose outputs were aligned to human expert ratings (Fig. 2b; Fig. 4a; Methods). Figure 2d compares the full PolicyAgent workflow across three backbone models (Gemini 2.5 Pro, Gemini 3 Flash, and Gemini 3.1 Pro). A worked example from the publication week of 13 January 2021 illustrates rubric annotations in matched expert-written and PolicyAgent-generated communications (Fig. 2c).

Agentic 3.1 Pro, the full PolicyAgent workflow running on the Gemini 3.1 Pro backbone, reached or exceeded the expert-written reference on seven of the nine rubric dimensions, with its deficits concentrated in evidence quality and policy intensity (Figure 2d; Extended Data Table 1). Its composite score exceeded that of the expert-written reference (4.44 versus 4.15; $P=4.7\times10^{-5}$), although this aggregate masked important differences across dimensions.

The strongest gains were in representational breadth. Vulnerability identification and stakeholder balance exceeded the expert-written reference (both $P<10^{-14}$), whereas geographic representation was statistically

indistinguishable from the reference, indicating that the agentic workflow consistently represented a broader set of affected groups and institutional perspectives. The remaining deficits occurred in dimensions that requires greater judgement rather than broader coverage. Within scientific soundness, the agent matched the expert-written reference on transparency and scored higher on benefit-cost balance, but evidence quality showed the largest gap on the rubric (3.15 versus 3.90). Within decision-relevance, emotional language and ideological framing matched or exceeded the reference, but policy intensity was lower, suggesting that expert authors more consistently translated similar epidemiological situation into direct recommendations. Overall, Agentic 3.1 Pro broadened representation while remaining weaker in evidence weighting and calibration of recommendation strength (per-dimension means, confidence intervals and tests in Extended Data Table 1).

The overall performance profile was consistent across backbone models. Representational breadth exceeded the expert-written reference across all three backbones, whereas evidence quality remained below the reference for every backbone examined. Among the three backbones, composite performance increased with language model capability from Gemini 2.5 Pro to Gemini 3 Flash to Gemini 3.1 Pro (4.03, 4.29 and 4.44, respectively), surpassing the expert-written reference between Gemini 2.5 Pro and 3 Flash. These gains were concentrated in presentation-related dimensions, including transparency and emotional language, whereas evidence quality changed little and remained far below the reference. Thus, stronger language models improved fluency and overall communication quality but did not substantially reduce the evidence-quality gap. A simplified workflow variant that omitted retrieval memory and iterative revision, Agentic 2.5 Pro (base), showed a similar pattern and is described in the Methods and Supplementary Information.

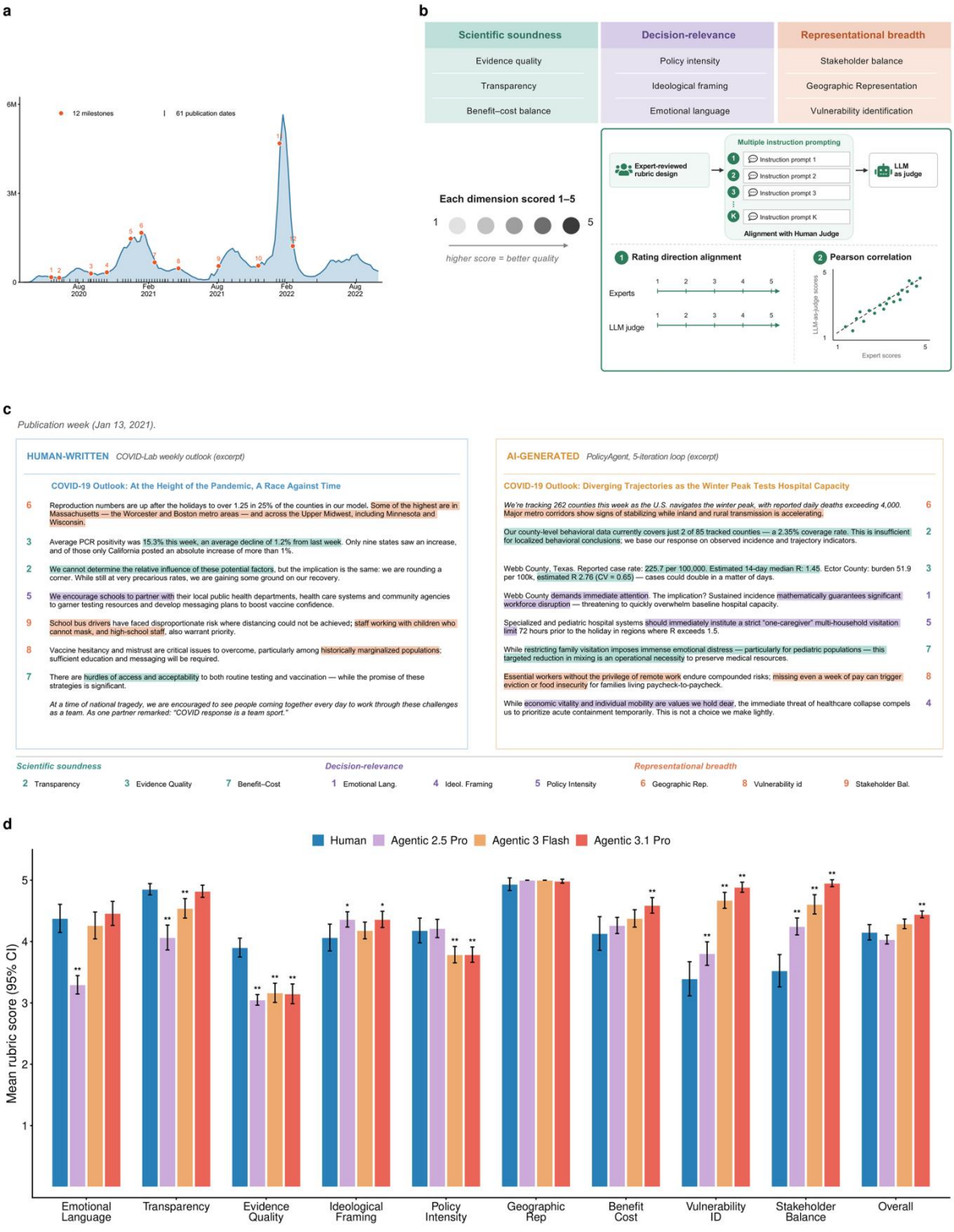


Figure 2. | Quality evaluation of expert-written and agent-generated public health policy communications. **a**, Evaluation timeline showing the 61 publication dates included in PHPC-Bench and the 12 milestone dates independently evaluated by human experts. **b**, Nine-dimension communication quality rubric grouped into three constructs. Human expert ratings were used to calibrate the LLM-as-judge. **c**, Example rubric annotations in matched expert-written and agent-generated communications from the publication week of January 13, 2021. **d**,

Communication quality scores for expert-written communications and PolicyAgent implemented on three backbone models (Gemini 2.5 Pro, Gemini 3 Flash, and Gemini 3.1 Pro). Bars represent mean scores across the 61 publication dates; error bars indicate 95% confidence intervals. P values were calculated using paired two-sided t-tests against the expert-written reference (*P<0.05, **P<0.01).

Guidance alignment remains comparable to expert-written communications despite the evidence-quality gap

The rubric analysis identified a persistent deficit in evidence quality. To better understand the source of this discrepancy, we examined whether lower evidence-quality scores reflected disagreement with contemporaneous public health guidance or weaker selection and weighting of otherwise guidance-consistent evidence. To distinguish between these possibilities, we compared policy claims against contemporaneous CDC and WHO guidance using a retrieval-augmented evaluation framework (Methods). Alignment was summarized at the claim level using the Claim Agreement Rate (CAR) and at the evidence-passage level using the Evidence Alignment Rate (EAR).

Contrary to the hypothesis that lower evidence-quality scores reflected disagreement with public health guidance, PolicyAgent and the expert-written reference achieved nearly identical claim-level agreement in the full 61-date benchmark (CAR 88.8% versus 88.4%; Fig. 3a). Passage-level agreement was also high, although modestly lower for PolicyAgent (EAR 90.9% versus 93.6%; Fig. 3b). Similar results were observed in the 12 milestone dates, where PolicyAgent exceeded the expert-written reference on both CAR (91.7% versus 82.5%) and EAR (95.9% versus 93.4%). Thus, the evidence-quality deficit observed in the rubric was not accompanied by a higher rate of contradiction with contemporaneous CDC/WHO guidance.

To characterize the remaining disagreements, we examined the frequency and retrieval rank of contradicting evidence passages. Contradictions were uncommon across workflow variants and backbone models. Among claims with any contradicting evidence, most were associated with only one or two contradicting passages among the five retrieved passages (Fig. 3c), and these passages were distributed across retrieval ranks rather than concentrated among the highest-ranked results (Fig. 3d). This pattern is more consistent with sparse, context-dependent retrieval conflicts, such as differences in population, timing, or recommendation granularity, than with systematic disagreement with highly relevant public health guidance. Taken together, these findings suggest that the evidence-quality deficit primarily reflects limitations in evidence selection, weighting, and framing rather than systematic conflict with contemporaneous guidance.

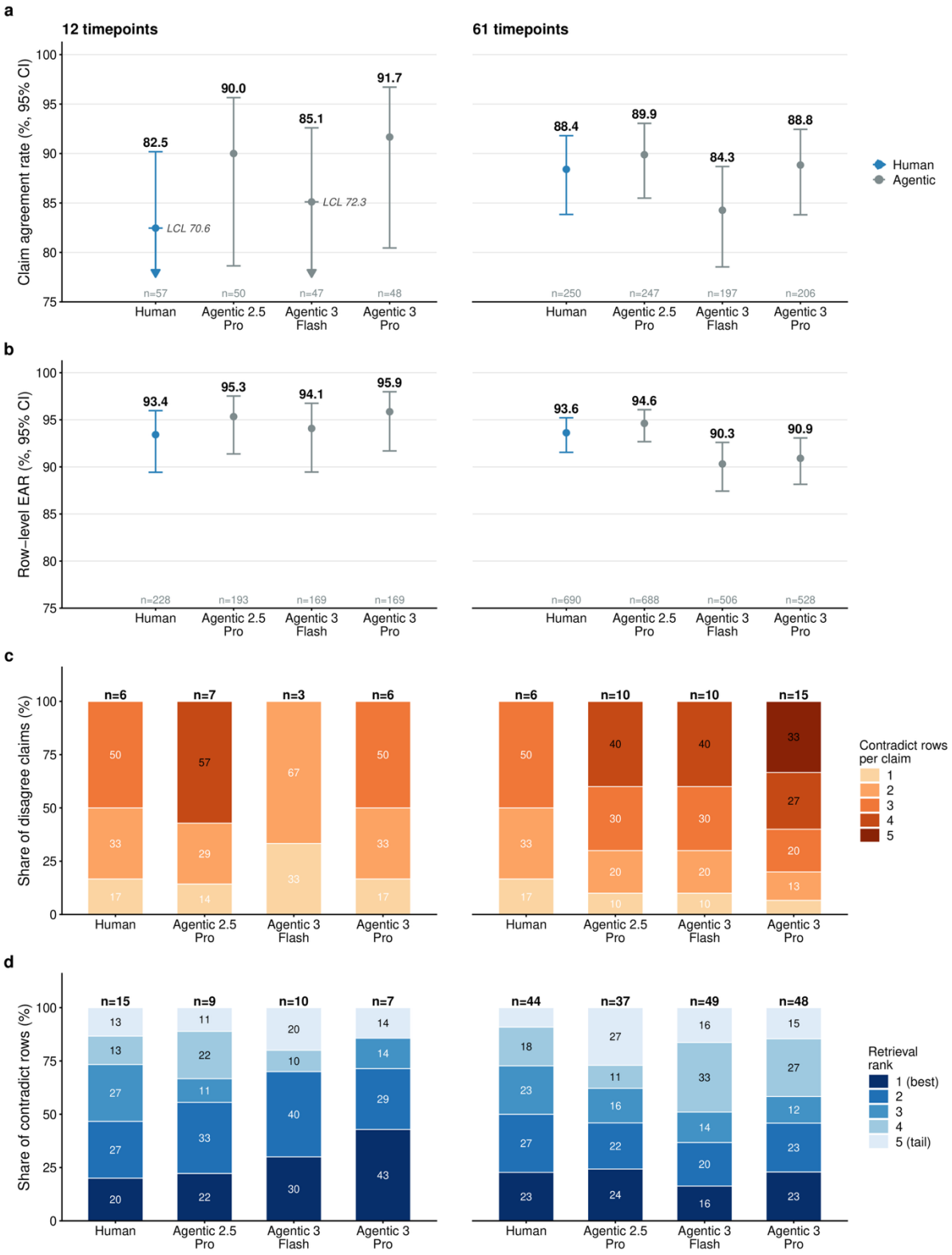


Figure 3. Alignment of policy claims with contemporaneous CDC and WHO guidance. Qualitative and policy claims extracted from expert-written and PolicyAgent-generated communications were matched to contemporaneous CDC and WHO guidance by retrieving the five most relevant guidance passages per claim. Each claim-passage pair was adjudicated as aligned, contradicting, or not addressing the claim (Methods). Results are

shown for the 12 milestone publication dates (left) and the full 61-date benchmark (right). Numbers (n) indicate the number of claims or claim-passage pairs evaluated in each group. Results of three agent runs (Agentic 2.5 Pro, Agentic 3 Flash and Agentic 3.1 Pro, the full workflow on Gemini 2.5 Pro, 3 Flash and 3.1 Pro) were included. **a**, Claim Agreement Rate (CAR), defined as the proportion of claims supported by at least one aligned passage and no contradicting passages among the top five retrieved passages. **b**, Evidence Alignment Rate (EAR), defined as the proportion of aligned claim–passage pairs among all aligned and contradicting pairs. Error bars in a and b indicate Wilson 95% confidence intervals. **c**, Distribution of claims containing at least one contradicting passage, stratified by the number of contradicting passages (1-5) among the five retrieved passages. **d**, Distribution of contradicting passages by retrieval rank, from rank 1 (most relevant) to rank 5.

PolicyAgent increases the traceability of quantitative claims to source data

We next examined whether the evidence-quality deficit reflected weak grounding of quantitative claims in the underlying surveillance data. This analysis evaluated one auditable aspect of evidence use: whether quantitative statements could be traced to surveillance values available to the system at generation time. Each quantitative claim was compared against the cached surveillance data available to the pipeline and classified as verified or not found (Methods). A not-found claim indicates a failure of source traceability or a mismatch between the claim and available source data; it does not necessarily indicate a factual error.

The data did not support the hypothesis that lower evidence-quality scores reflected poor numerical grounding. Across both Gemini 2.5 Pro and Gemini 3.1 Pro backbones, the full PolicyAgent workflow achieved higher source traceability than direct generation using the backbone model alone, reaching 94.9% claim-level verification accuracy for Agentic 3.1 Pro (Table 1). Importantly, this improvement did not arise from generating fewer quantitative claims. On the Gemini 3.1 Pro backbone, Agentic 3.1 Pro produced 894 quantitative claims compared with 301claims from the corresponding single-call baseline, while simultaneously reducing the proportion of claims that could not be verified. Thus, despite lower rubric scores for evidence quality, the full workflow generated quantitative claims that were both more numerous and more traceable to source data, arguing against the interpretation that the evidence-quality deficit is driven primarily by fabricated or unsupported numerical claims.

The increased volume of quantitative claims also affected date-level verification metrics. A publication date was considered fully accurate only if all quantitative claims generated on that date were successfully verified. Consequently, systems generating larger numbers of quantitative statements had more opportunities for at least one claim to be classified as not found. Although Agentic 3.1 Pro achieved the highest claim-level verification accuracy, it produced fewer fully accurate publication dates than the single-call baseline (33 of 61 versus 41 of 61) because a small number of unverified claims (46) were distributed across a larger set of dates. Overall, the full workflow improved source traceability at the claim level despite generating fewer dates with perfect verification.

Table 1. Evidence verification of quantitative claims against source surveillance data. Quantitative claims generated by single-call LLMs (Raw) and PolicyAgent (Agent) using Gemini 2.5 Pro and Gemini 3.1 Pro were compared against the surveillance and projection data available at generation time. Claims were classified as verified if the reported value matched the source data within a predefined tolerance and as not found otherwise (Methods). Results are aggregated across all 61 publication dates. Full-accuracy dates denote publication dates for which all quantitative claims were successfully verified. Overall accuracy is reported at the claim level as the proportion of verified quantitative claims among all quantitative claims.

	Gemini 2.5 Pro		**Gemini 3.1 Pro**	
	Raw	Agent	Raw	Agent

	Gemini 2.5 Pro		Gemini 3.1 Pro	
Dates with verifiable claims	58	60	47	61
Total quantitative claims	521	304	301	894
Verified	464	276	269	848
Not found	57	28	32	46
Full accurate dates	46/61	45/61	41/61	33/61
Overall accuracy	89.1%	90.8%	89.4%	94.9%

Evaluator reliability supports group-level comparisons

Because both the communication-quality assessment and the guidance-alignment audit relied on language-model evaluators, we evaluated whether these tools were sufficiently reliable to support the group-level comparisons reported above. Specifically, we examined agreement with human expert ratings, performance of the guidance verifier, and potential self-preference by the rubric judge.

On the 12 milestone publication dates independently evaluated by human experts (Fig. 2a), expert ratings identified strengths of PolicyAgent in stakeholder balance and vulnerability identification, weaknesses in emotional language, transparency, evidence quality, and ideological framing, and comparable performance on the remaining dimensions (Fig. 4a). The calibrated rubric judge recovered this same overall pattern of strengths and weaknesses. Because the rubric dimensions contain inherently subjective and normative elements, expert ratings of individual communications varied substantially (Supplementary Table S4). Accordingly, we use the calibrated judge to compare groups of communications across the benchmark rather than to replace human judgment for any individual communication or rubric dimension.

We next evaluated the guidance-alignment verifier against an independently adjudicated claim–evidence reference set. Across 500 claim-evidence pairs, the three evaluation components, i.e. population matching, polarity matching, and mechanism matching, each exceeded the predefined performance threshold for this audit (mean F1 = 0.774; mean accuracy = 0.726; Fig. 4b). Mechanism matching showed the lowest performance (F1 = 0.727), indicating that the verifier is suitable for aggregate analyses but should not be interpreted as definitive for any single claim-evidence comparison.

Finally, because the primary rubric judge shared a model family with the agent generator, we evaluated potential self-preference using four judges from independent model families. We found no evidence that the primary judge systematically favored PolicyAgent outputs. The difference between the primary judge and the peer-judge mean fell within the predefined equivalence margin (difference-in-differences = +0.018; two-sided P = 0.77; Fig. 4c). ChatGPT 5.5 likewise showed no significant directional bias, whereas Claude and Grok exhibited significant deviations in opposite directions that were not consistent across rubric dimensions (Methods). Together, these results support the use of PHPC-Bench for group-level evaluation of communication quality and guidance alignment, while recognizing that individual communication scores and claim-evidence judgments remain inherently noisier than aggregate benchmark comparisons.

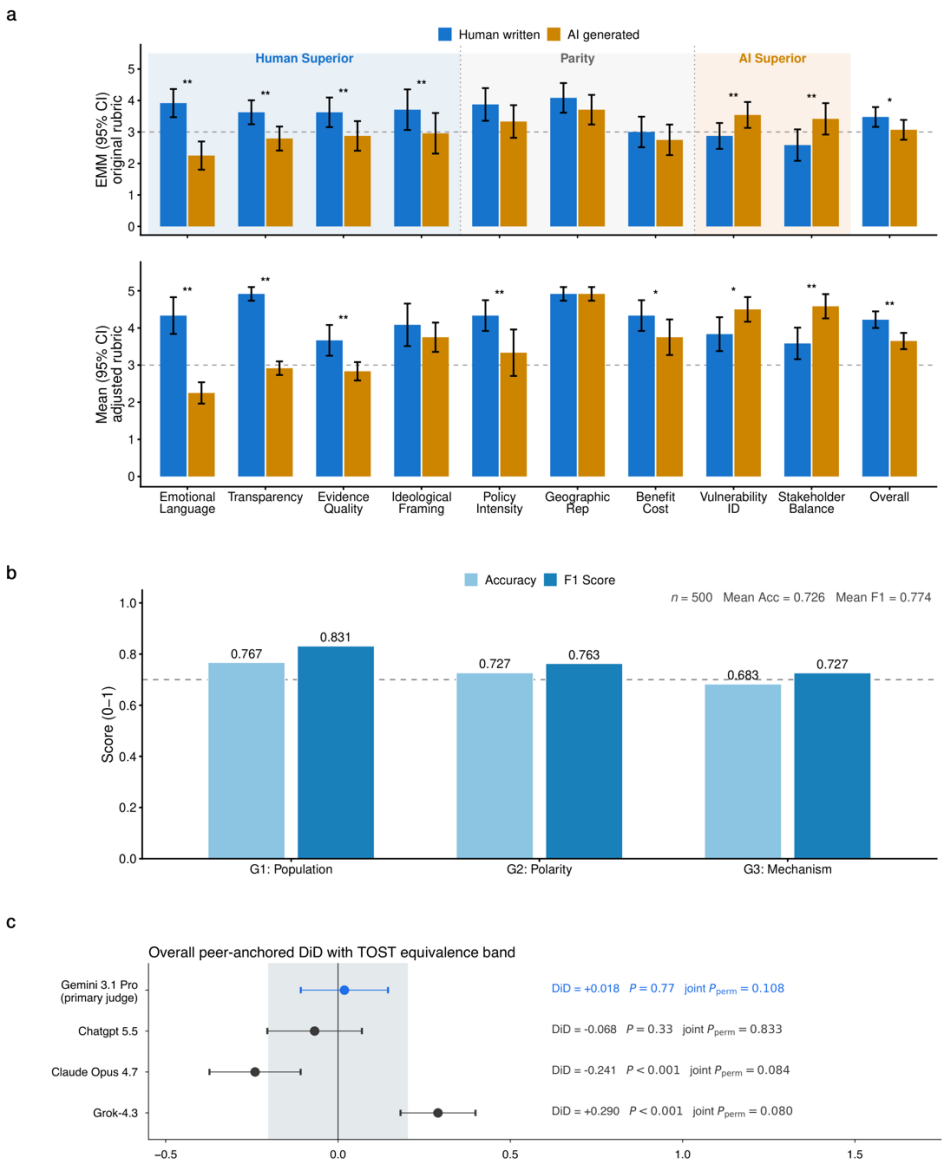


Figure 4. | Reliability of the PHPC-Bench evaluation framework. **a**, Comparison of human expert ratings and calibrated LLM-judge scores for expert-written and PolicyAgent-generated communications across the nine rubric dimensions. The upper panel shows estimated marginal means (EMMs) from a linear mixed-effects model of human expert ratings, adjusted for rater effects, with 95% confidence intervals. The lower panel shows mean scores assigned by the calibrated LLM judge across the 61 publication dates, with 95% confidence intervals. Asterisks indicate significant differences between expert-written and PolicyAgent-generated communications (*P<0.05, **P<0.01). **b**, Performance of the guidance-alignment verifier against an independently human-annotated validation set (n = 500 claim–evidence pairs). Accuracy and F1 scores are shown for the three evaluation components: population matching, polarity matching, and mechanism matching. The dashed line indicates the predefined performance threshold (0.7). **c**, Assessment of potential judge self-preference. Peer-anchored difference-in-differences (DiD) compares each judge's PolicyAgent-expert score difference with the mean difference across the remaining judges. Points indicate DiD estimates and error bars indicate 95% confidence intervals. Shaded regions denote the predefined equivalence bounds (±0.20) used for the two one-sided tests (TOST). Annotated P values test the null hypothesis of no bias (DiD = 0); joint permutation P values were obtained using a sign-flip permutation test

across the nine rubric dimensions (Methods).

Discussion

Across three independently auditable dimensions, PolicyAgent generated public health policy communications that approached expert-written references while maintaining strong alignment with contemporaneous public health guidance and underlying surveillance data. Under a nine-dimension communication rubric, the strongest configuration (Agentic 3.1 Pro) matched or exceeded the expert-written reference on seven dimensions, with gains concentrated in stakeholder balance and vulnerability identification. It also aligned with contemporaneous CDC and WHO guidance at rates comparable to the expert-written reference and achieved high traceability of quantitative claims to source data. Together, these findings demonstrate the feasibility of using agentic workflows to translate infectious disease surveillance and projection data into auditable public health communications. Importantly, this capability should be viewed as a draft-and-audit layer rather than autonomous policy authorship.

A central finding of this study is that factual grounding and expert-quality evidence use are not equivalent. Despite strong guidance alignment and numerical traceability, PolicyAgent consistently underperformed expert-written communications on evidence quality. This gap persisted across backbone models and workflow variants, suggesting that it does not primarily arise from model capability limitations. Rather, evidence quality appears to depend on selecting, prioritizing, and contextualizing relevant evidence, and on clearly communicating the rationale underlying recommendations. Factual correctness is therefore necessary but not sufficient for expert-quality evidence use.

The dimensions showing the largest gains, stakeholder balance and vulnerability identification, are also those most likely to benefit from multi-agent deliberation.[19] Public health policy communication requires integrating perspectives from epidemiology, clinical medicine, and public health operations while considering the needs of diverse populations. By distributing these responsibilities across specialized agents, PolicyAgent consistently represented a broader range of affected groups and institutional stakeholders than the expert-written reference. These findings suggest that multi-agent systems may be particularly valuable for expanding representational breadth, even when higher-level judgment tasks such as evidence weighting and recommendation calibration remain challenging.[32]

The lower policy-intensity scores should be interpreted in a similar context. The deficit does not imply that stronger recommendations are preferable, but rather that PolicyAgent was less calibrated in matching the strength, conditionality, and rights-sensitivity of recommendations to the underlying epidemiologic situation. Because these judgments involve normative considerations on which experts themselves may disagree, PHPC-Bench should be viewed as identifying areas requiring expert review rather than as defining a single correct recommendation. The benchmark is therefore most useful for revealing where human oversight remains necessary, not for delegating recommendation strength to an AI system.

The guidance-alignment and numerical-verification analyses help explain the nature of the remaining evidence-quality gap. PolicyAgent's recommendations remained highly consistent with contemporaneous CDC and WHO guidance, and quantitative claims were more traceable to source data than those generated by single-call LLM baselines. These findings argue against the interpretation that lower evidence-quality scores were driven primarily by fabricated numbers or systematic disagreement with public health guidance.

Instead, the remaining gap appears to arise from higher-order challenges in evidence selection, weighting, framing, and recommendation justification.

Because PHPC-Bench relies on LLM-based evaluators,[28,33] we explicitly assessed evaluator reliability rather than assuming it. The rubric judge reproduced the overall pattern of strengths and weaknesses identified by human experts, the guidance verifier achieved acceptable agreement with an independently annotated reference set, and we found no evidence that the primary judge systematically favored outputs generated by its own model family.[27] These analyses support the use of PHPC-Bench for aggregate comparisons between communication systems, while reinforcing that individual communication scores and claim–evidence judgments remain inherently noisier than benchmark-level conclusions.

COVID-19 provided an unusually well-instrumented testbed for evaluating public health communication systems. Dense county-level surveillance data, archived forecasting outputs, time-stamped CDC and WHO guidance, and a contemporaneous corpus of expert-written communications enabled generated text, source data, guidance, and reference communications to be evaluated within a common temporal context. This is both a strength and a limitation. Many public health domains, including chronic disease, environmental health, vaccine hesitancy, and healthcare operations, lack similarly complete surveillance, guidance, and communication archives.[4,34] Generalizability is further limited by the retrospective COVID-19 setting, the English-language U.S. corpus, and the use of a single expert-written reference source.

The scope of PHPC-Bench is also intentionally narrow. The benchmark evaluates whether communications are auditable, evidence-grounded, and aligned with contemporaneous guidance; it does not assess whether readers understand the communications, trust them, act on them, or perceive them as useful. Future work should evaluate reader comprehension, trust calibration, decision usefulness, and institutional adoption under expert-in-the-loop deployment settings.[9,35]

Within these bounds, our findings support a practical operating model for AI-assisted public health communication. Agentic systems can help structure the drafting process, synthesize surveillance data, surface stakeholder considerations, and expose an auditable evidence trail. However, evidence selection, recommendation calibration, and final approval remain responsibilities that require human expertise. The goal is therefore not to make public health communication automatic, but to make it more transparent, auditable, and scalable.

Reference

1. Ratzan, S. C. *et al.* The quality health information for all commission: reinventing health communication for the digital era. *Nat. Med.* **31**, 22–23 (2025).
2. Gostin, L. O., Ratzan, S. C. & Batista, C. Quality health information for all is a fundamental determinant of health. *Nat. Med.* https://doi.org/10.1038/s41591-026-04320-x (2026) doi:10.1038/s41591-026-04320-x.

3. Centers For Disease Control And Prevention. *Crisis and Emergency Risk Communication (CERC): Introduction*. (U.S. Department of Health and Human Services, 2018).
4. Van Bavel, J. J. & Others. Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460–471 (2020).
5. World Health Organization. *Communicating Risk in Public Health Emergencies: A WHO Guideline for Emergency Risk Communication (ERC) Policy and Practice*. (World Health Organization, Geneva, 2017).
6. Sauer, M. A., Truelove, S., Gerste, A. K. & Limaye, R. J. A failure to communicate? How public messaging has strained the COVID-19 response in the United States. *HEALTH Secur.* **19**, 65–74 (2021).
7. Zarocostas, J. How to fight an infodemic. *Lancet* **395**, 676 (2020).
8. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
9. Shoemaker, S. J., Wolf, M. S. & Brach, C. Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ. Couns.* **96**, 395–403 (2014).
10. Fraumann, G. & Colavizza, G. The role of blogs and news sites in science communication during the COVID-19 pandemic. *Front. Res. Metrics Anal.* **7**, 824538 (2022).
11. Rubin, D. *et al.* Association of social distancing, population density, and temperature with the instantaneous reproduction number of SARS-CoV-2 in counties across the United States. *JAMA Netw. Open* **3**, e2016099 (2020).
12. Reinhart, A. *et al.* An open repository of real-time COVID-19 indicators. *Proc. Natl. Acad. Sci.* **118**, e2111452118 (2021).
13. Patel, K. *et al.* Sustainability of the growth of the local public health workforce during the COVID-19 pandemic, 2019–2022. *Am. J. Public Health* **115**, 1271–1277 (2025).
14. Bishai, D. M. *et al.* Being accountable for capability—getting public health reform right this time.

Am. J. Public Health **112**, 1374–1378 (2022).

15. Wang, L. *et al.* A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345 (2024).
16. Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
17. Tam, T. Y. C. *et al.* A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Med.* **7**, 258 (2024).
18. Zhang, Z. *et al.* A survey on the memory mechanism of large language model-based agents. *ACM Trans. Inf. Syst.* **43**, 1–47 (2025).
19. Chen, X. *et al.* Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Med.* **8**, 159 (2025).
20. Du, H. *et al.* Advancing real-time infectious disease forecasting using large language models. *Nat. Comput. Sci.* **5**, 467–480 (2025).
21. Song, S., Liu, X., Li, Y. & Yu, Y. Pandemic policy assessment by artificial intelligence. *Sci. Rep.* **12**, 13843 (2022).
22. Shi, Z. *et al.* Coordinated pandemic control with large language model agents as policymaking assistants. (2026).
23. Min, S. *et al.* FActScore: fine-grained atomic evaluation of factual precision in long form text generation. in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 12076–12100 (Association for Computational Linguistics, Singapore, 2023). doi:10.18653/v1/2023.emnlp-main.741.
24. Rashkin, H. *et al.* Measuring attribution in natural language generation models. *Comput. Linguist.* **49**, 777–840 (2023).
25. Wang, Z. *et al.* Compliance and factuality of large language models for clinical research document generation. *J. Am. Med. Inf. Assoc.* **33**, 563–572 (2026).
26. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. in *Advances in*

Neural Information Processing Systems vol. 33 9459–9474 (2020).

27. Panickssery, A., Bowman, S. R. & Feng, S. LLM evaluators recognize and favor their own generations. in *Advances in Neural Information Processing Systems* vol. 37 (2024).
28. Liu, Y. *et al.* G-eval: NLG evaluation using GPT-4 with better human alignment. in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 2511–2522 (Association for Computational Linguistics, Singapore, 2023). doi:10.18653/v1/2023.emnlp-main.153.
29. Children's Hospital Of Philadelphia, PolicyLab. Responding to COVID-19. (2020).
30. U.S. Centers For Disease Control And Prevention. CDC expands eligibility for COVID-19 booster shots to all adults. (2021).
31. Tangcharoensathien, V. *et al.* Framework for managing the COVID-19 infodemic: methods and results of an online, crowdsourced WHO technical consultation. *J. Med. Internet Res.* **22**, e19659 (2020).
32. Cemri, M. *et al.* Why do multi-agent LLM systems fail? (2025).
33. Zheng, L. *et al.* Judging LLM-as-a-judge with MT-bench and chatbot arena. in *Advances in Neural Information Processing Systems* (2023). doi:10.52202/075280-2020.
34. Khoury, M. J., Iademarco, M. F. & Riley, W. T. Precision public health for the era of precision medicine. *Am. J. Prev. Med.* **50**, 398–401 (2016).
35. Tworek, H., Beacock, I. & Ojo, E. *Democratic Health Communications during Covid-19: A RAPID Response*. https://democracy.ubc.ca/policy-report/democratic-health-communications-during-covid-19-a-rapid-response/ (2020).
36. Mathieu, E. *et al.* Coronavirus pandemic (COVID-19). (2020).
37. Centers For Disease Control And Prevention. CDC museum COVID-19 timeline. (2024).
38. U.S. Census Bureau. American community survey 5-year data (2015–2019). (2019).
39. Getis, A. & Ord, J. K. The analysis of spatial association by use of distance statistics. *Geogr. Anal.* **24**, 189–206 (1992).
40. U.S. Department Of Agriculture, Economic Research Service. Rural–urban continuum codes (2023

release). (2023).

41. LangChain Inc. LangGraph: build resilient language agents as graphs. (2024).

42. Gemini Team, Google DeepMind. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. (2025).

43. Google DeepMind. *Gemini 3 Flash Model Card*. https://deepmind.google/models/model-cards/ (2025).

44. Google DeepMind. *Gemini 3.1 pro Model Card*. https://deepmind.google/models/model-cards/gemini-3-1-pro/ (2026).

45. Lenth, R. V. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. (2024).

46. Robinson, G. K. That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* **6**, 15–32 (1991).

47. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).

48. Anthropic. *Claude Opus 4.7 System Card*. https://www.anthropic.com/claude-opus-4-7-system-card (2026).

49. xAI. *Grok 4.3 Model Card*. https://docs.x.ai/developers/models/grok-4.3 (2026).

50. OpenAI. *GPT-5.5 System Card*. https://openai.com/index/gpt-5-5-system-card/ (2026).

51. Card, D. & Krueger, A. B. Minimum wages and employment: a case study of the fast food industry in new jersey and pennsylvania. Working Paper at https://doi.org/10.3386/w4509 (1993).

52. Schuirmann, D. J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **15**, 657–680 (1987).

53. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

54. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).

55. Xiao, S. *et al.* C-pack: packed resources for general chinese embeddings. in *Proceedings of the 47th*

International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24) 641–649 (2024). doi:10.1145/3626772.3657878.

56. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare* **3**, 1–23 (2021).

57. Hayes, A. F. & Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* **1**, 77–89 (2007).

58. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).

59. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).

60. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, New York, 2013). doi:10.4324/9780203771587.

61. Good, P. I. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. (Springer, New York, 2005).

62. Chroma. *Chroma: The Open-Source Embedding (Vector) Database*. (2026).

63. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).

64. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Software* **4**, 1686 (2019).

65. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016). doi:10.1007/978-3-319-24277-4.

Methods

Study design and evaluation framework

We conducted a retrospective, *source-relative* evaluation of PolicyAgent, an agentic workflow for drafting and auditing COVID-19 public health policy communication from temporal surveillance data. By source-relative we mean that generated claims are judged against the surveillance records available to the pipeline at each date, not against external world knowledge. The workflow synthesizes epidemiological surveillance

signals across 745 U.S. counties into draft policy commentaries, which we evaluate against an expert-written reference set rather than treating it as a substitute for expert authorship.

The design compared expert-written blogs from a single publishing source (n=61)[29] against AI-generated blogs for the same dates (n=61), scored on nine quality dimensions (range from 1-5) along two complementary tracks: a panel of nine recruited human experts (eight retained after outlier exclusion) and an LLM-as-judge[33] calibrated against their ratings. Reliability was then assessed through three audit layers of increasing interpretive demand: the human-calibrated quality judge, programmatic verification of every quantitative claim against the surveillance records[12,36] the pipeline could access, and retrieval-augmented[26] alignment of policy claims with contemporaneous CDC and WHO guidance.[30,31,36,37] Programmatic verification is the most direct external check; the rubric judge and the RAG verifier are calibrated screening instruments, interpreted alongside human validation and sensitivity analyses.

Benchmark corpus and surveillance data

Reference corpus and evaluation dates. The expert-written reference corpus comprises 61 expert-written COVID-19 policy blogs published by a single organization between May 2020 and August 2022, spanning six distinct pandemic waves (Supplementary Table S1).[29] Two nested date sets are analyzed. The primary evaluator-validation set comprises 12 CDC milestone dates[37] selected to cover major pandemic inflection points between May 2020 and February 2022, including reopening, summer and winter surges, vaccine rollout, Alpha and Delta emergence, booster recommendations, Omicron, and the transition toward endemic management; each date contributes one expert-written and one AI-generated blog (Supplementary Table S2). The extended architectural comparison set comprises all 61 publication dates. For a report dated T, only data with observation date $\leq$ T were admissible; news-cache, statistics-cache, API-query and prediction dates were each checked, and violations raised a *TemporalValidationError* with audit logging.

Surveillance data sources. Our system integrates epidemiological surveillance data from multiple sources through a unified pipeline. Real-time indicators are retrieved from the Delphi COVIDcast API,[12] including hospital admissions and community-level disease activity. (Test positivity is not surfaced to the drafting prompt and is therefore treated as unsupported at verification; see Programmatic verification of quantitative claims.) Demographic covariates, including urban-rural classifications, age structure, socioeconomic indicators and population statistics, are derived from the 2019 American Community Survey.[38]

For each analysis time point, we retrieve 28-day transmission trajectories (14 days retrospective and 14 days prospective) generated by the COVID-Lab prediction model, together with social-distancing metrics within a two-week window of the query date.[11] We compute three primary epidemiological indicators: the median effective reproduction number (R), the temporal coefficient of variation of R over 14 days, and the 14-day cumulative incidence per 100,000 population. Spatial clustering analysis employs the *Getis-Ord* G_i^* statistic[39] to identify transmission hotspots and coldspots.

Static cached data include epidemiological predictions for 745 counties, social-distancing metrics from mobility patterns, and geographic metadata from the 2019 U.S. Census, Rural–Urban Continuum Codes, and FIPS proximity matrices.[40] Data quality was assessed using coverage thresholds: GOOD ($\geq$ 70%), FAIR ($\geq$ 40%), and Limited ($<$ 40%). Temporal matching employed a ±14-day window for local data and ±120 days for external APIs, with COVIDcast fallback mode activated when offline coverage fell below 50%.

PolicyAgent workflow

PolicyAgent is a 10-node directed acyclic graph (DAG) implemented in LangGraph[41] (Fig. 1), with one conditional loop-back edge that triggers up to three revision iterations. The pipeline proceeds through ten stages in a fixed order: it first gathers all data, then performs a situational analysis, creates a strategic brief, convenes an expert debate, issues a data request, carries out collaborative drafting, reviews and revises the draft, polishes it for style, runs a final cleanup, and finally saves the blog. Independent subtasks within a stage execute asynchronously. The architecturally significant components are those whose effects are visible in the Results.

Expert debate. The expert-debate stage convenes the three persona agents shown in Fig. 1b (an Epidemiologist, a Clinical Physician and an Institutional Manager) in a two-phase, asymmetric advisors-to-decision-maker structure. In Phase 1 the Clinical Physician (patient safety and capacity) and the Institutional Manager (implementation feasibility and stakeholder conflict) act as advisors and run in parallel; each emits a memo following a structured output contract specifying role, core concern, source-typed evidence, recommended action with actor/population/burden, stakeholders most helped and most burdened, main uncertainty, and explicit deferrals. In Phase 2 the Epidemiologist acts as decision-maker, synthesizing both memos into a recommendation under its own structured contract (shared ground, main disagreements, decision, rationale, trade-offs, operationalization, reversal conditions, residual uncertainty). Four institutional scenarios, namely hospital, government (the default), research and rural, modulate persona constraints and stakeholder lists.

Drafting and revision loop. The collaborative-drafting stage assembles the debate output, strategic brief and runtime evidence, and prompts the backbone model to produce a draft under a mandatory geographic progression that descends from the national level through the regional and state levels to the county level, with prescribed quantitative-density targets. The review-and-revision stage then performs evidence verification against the runtime statistics cache and triggers up to three priority-sorted revision iterations when evidence-validation errors, excessive warnings, word-count deviations, or deterministic-guard violations are detected.

Reference-style transfer and polish. The style-polishing stage applies a reference-blog style encoding learnt from a small reference corpus, combining a small number of few-shot expert-written blogs, contrastive before/after transformations correcting overstatement, and critical-enhancement rules covering meta-commentary removal, explicit value trade-off statements and stakeholder-empathetic language.

Hallucination prevention. Four layered defenses guard against fabricated statistics: (i) a learning log of past data-validation errors prepended to the decision-maker prompt as operational protocols; (ii) structured output contracts that require every cited statistic to carry a source-type label, one of observed, model-based, survey-based or contextual, and explicitly prohibit fabrication; (iii) a post-draft LLM hallucination[16] detector complemented by deterministic regex guards for hot-spot misuse and causal behavioral overclaims; and (iv) a polish-stage rule that prohibits introducing any numerical value not present in the upstream evidence.

Model configuration. Primary analysis, synthesis and review use the backbone model at temperature 0.3; drafting at temperature 0.7; the LLM-as-judge runs at temperature 0.1 deterministically. Backbone choice is the principal ablation dimension (see Computer configurations).

Additional implementation details, including prompt templates and structured-output contracts (Supplementary Table S9), generation and revision settings and retrieval parameters (Supplementary

Table S10), and data-source availability (Supplementary Fig. S2), are provided in the Supplementary Information; lower-level implementation specifications are provided in the released source code.

Comparator configurations

The main-text analyses contrast a single-call raw baseline against four agentic configurations, all evaluated on the same dates under the same rubric and LLM-as-judge configuration (see LLM-as-judge calibration and sensitivity analysis). The *raw baseline* (labelled "Raw" in the tables) is a single LLM invocation that bypasses the agentic workflow entirely, evaluated on the Gemini 2.5 Pro and 3.1 Pro backbones. The four agentic configurations vary two factors, using the same names as the Results:

- **Workflow form** — the *base* workflow, which omits the revision loop and retrieval/temporal memory, versus the *full* workflow, which adds all defensive layers (memory management, the up-to-three-iteration revision loop and hallucination detection; see Hallucination prevention).
- **Backbone model** — the underlying LLM the workflow runs on: Gemini 2.5 Pro, Gemini 3 Flash or Gemini 3.1 Pro.

Crossing these factors yields the four runs reported in the main text, each named by backbone with a *(base)* suffix marking the early workflow:

- **Agentic 2.5 Pro (base)** — base workflow on Gemini 2.5 Pro;[42]
- **Agentic 2.5 Pro** — full workflow on Gemini 2.5 Pro;
- **Agentic 3 Flash** — full workflow on Gemini 3 Flash;[43]
- **Agentic 3.1 Pro** — full workflow on Gemini 3.1 Pro.[44]

We designate **Agentic 3.1 Pro**, the full workflow on the strongest backbone, as the primary configuration; the remaining runs test whether the rubric pattern holds across backbones and across the base-versus-full workflow contrast. Two contrasts isolate the design factors: the *Raw → agent* step at a fixed backbone isolates the contribution of the agentic workflow, while the *Agentic 2.5 Pro (base) → Agentic 3.1 Pro* step captures the combined effect of upgrading the backbone and adding the full set of defensive layers, rather than isolating either component alone. Additional exploratory configurations spanning seven design dimensions (backbone model, multi-agent debate, temporal memory, writing-style polish, revision loop, institutional framing and source attribution) were examined during development to inform design choices and were not used for confirmatory inference.

Human rating and consensus-score estimation

Nine-dimension rubric. Content was evaluated on a 1-5 scale across nine dimensions: policy intensity, evidence quality, ideological framing, emotional language, stakeholder balance, geographic representation, transparency, vulnerability identification and benefit-cost balance. Full definitions with score anchors for each dimension are provided in Supplementary Table S3 (Public Health Blog Quality Rubric, adjusted version).

Rater expertise, calibration, and scoring procedure. The nine recruited raters were policy researchers working in health policy, each holding a PhD or master's degree in a relevant discipline. Prior to formal evaluation, the raters participated *together* in a single 2-hour in-person calibration session designed to align rubric interpretation across the group. During this session, raters independently scored sample articles and then discussed every dimension's score anchor and ambiguous cases in simultaneous group discussion, continuing until convergence on a shared reading of each anchor and dimension definition. Formal evaluation took place in a subsequent in-person session, during which raters scored independently and

without consultation.

Rater assignment and outlier handling. Nine annotators were initially recruited to evaluate 24 unique articles (12 expert-written, 12 AI-generated) using an incomplete crossed design: eight raters each evaluated 6 articles and one rater evaluated 3 articles, with deliberate overlap in article assignments to enable estimation of rater-specific bias. The 24 articles correspond to the 12 CDC milestone dates listed under Reference corpus and evaluation dates (one expert-written and one AI-generated blog per date); the remaining 49 dates in the 61-date extended corpus were not human-rated and are evaluated by the LLM-as-judge only (see LLM-as-judge calibration and sensitivity analysis).

During quality control review, the ninth rater (who completed only 3 articles) was identified as an outlier. This rater assigned predominantly low scores (score of 2) across dimensions and content types, exhibited substantially lower within-rater score variability, and showed less discrimination among articles and rubric dimensions than the other raters. To improve rating reliability, this rater was excluded from subsequent analyses. Consensus Estimated Marginal Means (EMMs),[45] best linear unbiased predictor (BLUP)[46] estimates, and variance decomposition were therefore calculated using the remaining eight raters (n=48 rating observations; Supplementary Fig. S1; Supplementary Table S4). Across the eight retained raters, rater stringency was estimated via BLUPs[46] from a linear mixed model[47] (see Statistical analysis) and spanned a modest lenient-to-stringent range (Supplementary Table S4).

Score aggregation. To account for rater heterogeneity, we fitted a linear mixed model jointly across all 24 articles within each dimension (see Statistical Analysis). Consensus scores were computed as the sum of the global intercept and shrinkage-corrected article effects, removing systematic rater bias from final estimates. Articles were subsequently stratified by authorship condition (expert-written vs. AI-generated) for comparative analysis.

LLM-as-judge calibration and sensitivity analysis

Automated quality scoring followed the LLM-as-judge paradigm[33], in which a large language model rates generated text against an explicit rubric. Initial automated evaluation used Gemini 2.5 Pro with temperature 0.1 for deterministic scoring. Calibration assessed human–AI alignment by computing, within each rubric dimension, the Pearson correlation (r) between the LLM-as-judge score and the human-rater estimated marginal mean (EMM) at the article level. EMMs are derived from linear mixed-effects models with rater and date as random effects (see Statistical analysis). Each dimension's correlation is therefore taken across n=24 articles (the 12 expert-written + 12 AI-generated blogs from the milestone-date calibration set; significance tested against the null r=0 with df=22). The rubric was developed iteratively. The original rubric revealed sign reversals on four dimensions, indicating systematic directional disagreement. We then evaluated three candidate grading models (Gemini 2.5 Pro, Gemini 3.1 Pro, Gemini 3 Flash) and Gemini 3.1 Pro achieved the highest overall alignment. The adjusted rubric was developed by targeted prompt calibration, incorporating pattern-based scoring guidelines for Transparency (TYPE A/TYPE B epistemic honesty framework), epistemic calibration penalties for Evidence Quality, and spatial trade-off reasoning for Benefit–Cost Balance. The adjusted rubric was finalized before scoring the 61-date extended corpus and before any of the architectural comparisons were performed, so the calibration could not have been tuned to those downstream results. Under the adjusted rubric, Gemini 3.1 Pro reached statistically significant alignment with human annotators ($P<0.05$) on the overall composite and six of the nine rubric dimensions. The three that did not reach significance (geographic representation, vulnerability identification and benefit–cost balance) reflect construct-validity limits (a geographic ceiling effect, and

implicit-versus-explicit framing for the other two) and are retained but interpreted with caution. For dimensions vulnerable to surface fluency or unsupported specificity, we applied conservative scoring rules that penalized unqualified claims, missing trade-off reasoning, generic stakeholder lists and unsupported policy specificity regardless of authorship condition, so that observed score increases reflect genuine improvements rather than artefactual leniency (Supplementary Table S3). This Gemini 3.1 Pro / adjusted-rubric configuration is the judge used for all main-text scores reported in Fig. 2 and Table (Extended Data Table 1), for the n=61 paired t-tests in Supplementary Table S6, and for the *Human* reference reported in those displays (the same judge applied to the 61 expert-written blogs). The judge shares its Gemini 3.1 Pro backbone with the Agentic 3.1 Pro generator, introducing a potential judge–generator confound. Because model-based judging may introduce non-additive, dimension-specific bias, we do not interpret LLM-as-judge scores as unbiased estimates of communication quality; instead, we treat them as calibrated screening measures and triangulate them with three further checks (human-EMM calibration, programmatic quantitative verification and RAG-based claim audits) together with the cross-judge sensitivity analysis below. The calibration reported above, which reached significant agreement with the human EMM on the overall composite and six of the nine dimensions, establishes partial agreement with human graders; the four-judge sensitivity analysis bounds the residual same-family confound directly.

Judge model sensitivity: design. To test the judge–generator confound directly, rather than only through the three indirect checks above (human-EMM calibration, programmatic verification and RAG audits), the same 24 milestone blogs (12 generated by Agentic 3.1 Pro, 12 expert-written; see Reference corpus and evaluation dates) were re-scored by four flagship judges of disjoint training lineage: Gemini 3.1 Pro (the primary judge), Anthropic Claude Opus 4.7,[48] xAI Grok-4.3[49] and OpenAI ChatGPT (gpt-5.5).[50] All four received byte-identical prompt packets (the adjusted v3.4 rubric and the extracted blog text), so any between-judge variation reflects the scoring model alone; the first three ran at temperature 0.1, whereas ChatGPT was held at the provider default of 1.0 by API constraint and so carries slightly higher run-to-run variance. The inferential unit is a peer-anchored difference-in-differences (DiD):[51] for each judge we take its agent-minus-human score gap, $24\Delta_j = s_{j,AI} - s_{j,Human}$, average it over the 108 cells (12 dates × 9 dimensions), and compare it by paired two-sided t test against the mean gap of the other three judges (the peer baseline).

Equivalence margin and its sensitivity. Equivalence to the peer baseline was assessed by two one-sided tests (TOST):[52] a judge counts as equivalent when the two-sided 95% CI of its DiD lies entirely within ± (these CIs are the error bars in Fig. 4c; reading equivalence off the 95% CI is a conservative TOST at α=0.025 per side). The margin =0.20 was not pre-registered but chosen *post hoc* as the smallest effect size of interest, less than one-fifth of a single step on the 1–5 rubric. A sweep across candidate margins shows this choice to be neither too strict nor a rubber stamp: at =0.20 only the primary Gemini judge is equivalent (DiD=+0.018, 95% CI [-0.108,+0.145]); Claude and Grok fall clearly outside (DiD=-0.241 and +0.290, both P<0.001), and ChatGPT falls just outside (CI reaching -0.205 despite a near-zero gap, DiD=-0.068, P=0.33). A stricter =0.10 would reject even the primary judge, whereas a looser =0.25 would begin to admit ChatGPT. Because the margin was fixed after the estimates were seen, we read the equivalence results as a sensitivity bound on the confound, not as a confirmatory test.

Joint test across dimensions. To complement the cell-level test, the nine per-dimension DiDs were tested jointly: the 12 per-date DiD 9-vectors were stacked into a 12 × 9 matrix and summarized by the Wald statistic $W=\widehat{\beta^T}V^{-1}\beta$ from the column means $\hat{\beta}$ and date-row covariance $\hat{V}$. Because n=12 against a 9-dimensional mean lies far outside the $X^2(9)$ asymptotic regime, W was referred to an exact sign-flip permutation distribution over all 2^{12}=4096 ±1 assignments of the date rows, and per-dimension paired t

statistics were Holm-corrected[53] within each judge across the nine dimensions.

Secondary analyses. Two further analyses were also examined: a human-anchored DiD that replaces the peer baseline with the eight-rater human EMM (see Human rating and consensus-score estimation), and single-peer DiDs comparing Gemini against each peer individually. The main text reports the peer-anchored overall DiD for the primary judge together with the joint Pperm.

Programmatic verification of quantitative claims

Beyond the rubric-based judgement of writing quality, every quantitative claim made in each generated blog (Table 1) was checked programmatically against the surveillance data the pipeline had access to. We used four ground-truth sources, routed by claim type:

- *Epidemiological predictions* (R value, 14-day cumulative incidence per 100,000, daily case counts, case rates): the corresponding (county, date) entry from the *pred_obs* prediction cache joined to the 2019 Census population denominator. A claim was *verified* if it matched the cached value within rounding (one decimal place for R).
- *COVIDcast behavioural and healthcare signals* (mask wearing, vaccine acceptance, COVID-like illness, hospital admissions): the matching COVIDcast signal at ±2 percentage-point tolerance.
- *COVIDcast JHU–CSSE epidemiological feed*[54] *(confirmed_7dav_incidence_prop) for 2022+ dates*: at ±20% tolerance. The broader window reflects the 2022 fallback protocol, which substitutes the JHU–CSSE feed for the discontinued *pred_obs* predictions (cutoff 2021-12-07).
- *Test positivity*: positivity is *not* provided in the LLM prompt context, so any positivity claim is classified as unsupported by the provided prompt context.

A claim that fell outside its tolerance window, or that invoked a metric absent from the LLM prompt context, was labelled *not found*. A claim referring to a metric for which no ground-truth source existed on the given date (e.g.\ an R value for a 2022 date with no *pred_obs* coverage) was labelled *unverifiable* and excluded from the accuracy denominator. *Full-accuracy dates* are dates with zero not-found claims. *Overall accuracy*, as reported in Table 1, is Verified/(Verified + Not found). Quantitative claims were extracted from each blog by the verification pipeline and cross-referenced against the source-typed labels emitted by the in-debate output contracts (see Hallucination prevention); the per-date, per-configuration breakdown for all 61 dates is released as part of the source-data archive (Data availability).

Qualitative case study. As an illustrative complement, outputs from four configurations for one matched date were compared across five qualitative themes: analytical depth, multi-metric evidence integration, spatial reasoning, actionable recommendations and calibrated uncertainty (Supplementary Table S8).

Retrieval-augmented verification against CDC and WHO guidance

Public health guidance evolves rapidly and often includes population-specific qualifications. To assess factual alignment of generated policy blogs with contemporaneous CDC and WHO guidance, we developed a retrieval-augmented verification pipeline combining a temporally versioned knowledge base, hybrid dense retrieval, and a three-gate LLM judge calibrated against adjudicated human annotations.

Curated knowledge base and temporal versioning. We compiled a corpus of 60 policy documents from the CDC (n = 37) and WHO (n = 23), spanning five document categories (Extended Data Fig. 1). The category counts were 29 current-guidance documents, 6 archived protocols, 2 technical documents, 22 topic-specific WHO guidance documents and 1 infection-prevention-and-control document. Timestamped "current" and "archived" snapshots were maintained for each document. For each claim, only documents

with publication date ≤ T were eligible for retrieval, ensuring that retrieved evidence reflects the policy stance in effect at the time of the claim.

Hybrid retrieval and condensation. We benchmarked 10 dense retrieval configurations on a held-out 10-claim retrieval test set spanning the five knowledge-based document categories, selecting BGE-Large-en-v1.5[55] with 2800-token chunks (NDCG@5 = 0.947, Hit Rate@5 = 1.00); the full claim-level verification pipeline was then validated against the N=500 adjudicated human gold standard described below, so that the small retrieval-stage set is used only to select the retriever model, not to validate downstream claim alignment. The superiority of BGE-Large over domain-specialized models such as PubMedBERT[56] reflects the mixed administrative and clinical register of CDC/WHO documents, which requires broader semantic modelling than biomedical abstracts alone. The advantage of 2800-token chunks further support preserving contiguous policy conditions for retrieval coherence. Candidate evidence passages are condensed to minimal semantically complete sentences, cached using hashed fingerprints, and the top 5 ranked candidates passed to the gate-based evaluation module. **Gate-based evaluation.** Each candidate evidence passage is evaluated by a three-gate LLM judge:

- **G1 (Population compatibility):** Does the evidence apply to the same population or setting as the claim?
- **G2 (Polarity alignment):** Does the evidence support the same direction or modality as the claim?
- **G3 (Mechanism consistency):** Does the evidence share the same causal mechanism or rationale as the claim?

Gate semantics. The three gates play asymmetric roles that motivate the verdict rules below. G1 is a *relevance gate*: a passage about the wrong population cannot support or refute the claim, so any G1=0 pattern is off topic. G2 sets the verdict *direction* (support vs. opposition). G3 acts as a *causal engagement check* that confirms the evidence is reasoning about the same causal claim rather than a different mechanism that happens to share a policy direction. Consequently, ALIGNED accepts G3=0 because directional agreement is self-supporting, whereas CONTRADICTS requires G3=1 because directional opposition is only refutation once the same mechanism is in play.

Verdict assignment. Claim-level verdicts are assigned from gate outputs [G1, G2, G3] \0,1\³:

- **NOT ADDRESSED** when G1=0 (patterns [0, *, *]): the evidence is off topic regardless of G2 or G3.
- **ALIGNED** when G1=G2=1 — [1,1,1] (strong; mechanism confirmed) or [1,1,0] (weak; direction-only). G3 distinguishes strength of support but does not flip the verdict.
- **CONTRADICTS** when [G1, G2, G3] = [1,0,1]: same population, same causal mechanism, opposing direction.
- **NOT ADDRESSED** when [G1, G2, G3] = [1,0,0]: opposing direction but a different mechanism, so the evidence engages a different causal claim and cannot refute this one.

At the claim level, CONTRADICTS takes priority: if any retrieved passage yields [1,0,1], the claim is marked CONTRADICTS regardless of co-occurring supporting passages. Otherwise, ALIGNED is assigned if any passage yields [1,1,1] or [1,1,0]. Otherwise, NOT ADDRESSED is assigned. Alignment is summarised at two grains. At the claim level, the Claim Agreement Rate (CAR) is the proportion of claims with at least one aligned passage and no contradicting passage among their top-five retrieved rows (Fig. 3a). At the passage level, the Evidence Alignment Rate (EAR) pools over the retrieved rows themselves (Fig. 3b):

$$EAR = (Aligned)/ (Aligned + Contradicting) \times 100\%,$$

where Aligned and Contradicting count passage rows rather than claims. Because EAR conditions on a claim being adjudicable, it does not capture the proportion of claims that fall outside the retrievable evidence corpus or that are framed in ways the cascade cannot match. We therefore distinguish two complementary metrics: *addressability* (the share of claims receiving either an ALIGNED or CONTRADICTED verdict) and the *NOT-ADDRESSED rate* (its complement), so that contradiction risk can be distinguished from attribution risk.

Human reference standard for RAG calibration. The hold-out test set comprised 500 claim–evidence pairs drawn from the 12 milestone-date blogs analyzed in Fig. 2: 250 pairs derived from expert-written blogs and 250 pairs derived from the matched Agentic 2.5 Pro (base) outputs, enabling balanced evaluation of the verification pipeline across content types. Annotation was carried out by a panel of four STEM-trained reviewers, each holding a master's degree in a relevant discipline. This panel is distinct from the eight Fig. 2 raters of Methods; the two annotation efforts were conducted independently to keep blog-quality scoring and claim-level evidence adjudication on separate evaluator pools. A two-session protocol was used. In session 1, the four annotators were organized into two pairs, and the 500 claims were split into two halves of 250 with one half assigned to each pair; within each pair, the two annotators independently rated every claim in their assigned half across the three gates (G1 population, G2 polarity, G3 mechanism), so every claim received two independent labels per gate. In session 2, each pair re-examined the items on which they disagreed and resolved them by discussion; the small number of items still in disagreement after pair discussion were referred to a fifth annotator, who provided a tie-breaking grade that the pair adopted. Each claim therefore carries at least two and at most three human labels per gate, with the post-adjudication consensus adopted as the gold standard. Pre-adjudication agreement was summarized using pairwise Cohen's κ and three-rater Krippendorff's α.[57] Pairwise estimates, exact per-comparison sample sizes and the three-rater α values are reported in Supplementary Table S7.

Calibration and validation. Gate-level performance was evaluated on this hold-out test set of N=500 claims against the human gold standard. All three gates exceeded the F1 = 0.7 reliability threshold, with F1 decreasing from G1 to G3 (per-gate accuracy and F1 in Fig. 4b). This ordering reflects the increasing inferential complexity of each gate: population matching is largely lexical, polarity alignment requires statement-level entailment, and mechanism consistency requires causal reasoning across heterogeneous policy documents.

Statistical analysis

Primary comparison. Analyses were paired at two scales. On the 12 CDC milestone dates, each timepoint yielded one expert-written and one AI-generated blog, each scored both by the human expert panel and by the LLM-as-judge. Human-rated quality was compared between conditions (Fig. 4a, top) using estimated marginal means (EMM) derived from linear mixed-effects models fitted separately for each dimension:

$$score_{ij} = \beta_0 + \beta_1 \, Title_j + u_t + v_i + \varepsilon_{ij}$$

where i indexes raters, j indexes articles, $u_t \sim N(0, \sigma^2_{time})$ represents date-level random effects, and $v_i \sim N(0, \sigma^2_{rater})$ represents rater-level random effects. *Title* (expert-written vs. AI-generated) is the fixed effect of interest. EMMs were extracted using the *emmeans* package and pairwise contrasts were tested using Wald tests with Satterthwaite degrees of freedom via *lmerTest*; error bars are model-based 95% confidence intervals from the EMM estimates. LLM-as-judge means on the same 12 dates are shown beneath the

human estimates (Fig. 4a, bottom), with significance from two-sided paired t-tests (n=12 pairs).

Linear mixed model for consensus scores. Bias-corrected consensus scores (the human EMM plotted in Fig. 4a) were derived from the same linear mixed-effects model structure, with date and rater as crossed random effects:

$$score_{ij} = \beta_0 + u_t + v_i + \varepsilon_{ij}$$

where β_0 is the global intercept, $u_t \sim N(0, \sigma^2_{time})$ represents date-level variation in content difficulty or epidemiological context, $v_i \sim N(0, \sigma^2_{rater})$ represents rater stringency, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is residual error. Consensus scores were computed as $\beta_0 + u_t$ (shrinkage-corrected timepoint effects removing systematic rater bias).

Intraclass correlation coefficients (ICC).[58] Inter-rater reliability and temporal stability were quantified using ICC derived from variance components:

$$ICC_{time} = \frac{\sigma^2_{time}}{\sigma^2_{time}+\sigma^2_{rater}+\sigma^2_{residual}},\ ICC_{rater} = \frac{\sigma^2_{rater}}{\sigma^2_{time}+\sigma^2_{rater}+\sigma^2_{residual}},$$

ICC_{time} quantifies the proportion of variance attributable to pandemic timepoint (data-dependent dimensions); ICC_{rater} quantifies variance attributable to individual rater differences (subjective dimensions). Full variance component estimates are reported in Supplementary Table S4.

Agreement between each evaluator configuration and the human expert estimated marginal means was assessed using Pearson correlations across the 24 author-by-timepoint conditions; these calibration results are reported in Supplementary Table S5.

For the 61-date extended comparison (Fig. 2d), paired two-sided t-tests compared each agent configuration against the expert-written reference, paired by publication date (n=61 pairs); error bars are 95% confidence intervals computed as $x \pm t_{0.975,\ n-1} \times SE$.

RAG factual accuracy. The Evidence Alignment Rate (EAR) is the binomial proportion aligned/(aligned+contradicting). The Claim Agreement Rate (CAR) and row-level EAR (Fig. 3a, b) are reported with Wilson score 95% confidence intervals,[59] where n is the number of adjudicated claims or rows; the Wilson interval is used in preference to the normal (Wald) approximation because rates lie close to the 0–1 boundary. The composition of disagreeing claims by number of contradicting rows and the retrieval-rank distribution of contradicting rows (Fig. 3c, d) are descriptive 100%-stacked proportions and are not accompanied by inferential tests.

Effect sizes and P values. Effect sizes were calculated using Cohen's d.[60] Dimension-level comparisons used two-sided paired t-tests at $\alpha = 0.05$. Because the nine rubric dimensions are conceptually distinct but statistically correlated, dimension-level P values are reported descriptively in the Supplementary information and interpreted jointly with effect sizes, confidence intervals and sensitivity analyses rather than as a single confirmatory family. Holm-corrected results are reported for the judge-sensitivity analysis, where all nine dimensions are tested under one explicit joint hypothesis (see Judge-sensitivity equivalence Cohen's dtesting).

Judge-sensitivity equivalence testing. Between-judge bias was quantified by a peer-anchored difference-in-differences (DiD): for each judge, the agent-minus-human score gap was contrasted with the mean gap of the remaining judges across the 108 (date × dimension) cells, tested two-sided against DiD=0 by paired

t test. The nine-dimension joint hypothesis was assessed with a Wald statistic referred to an exact 4,096-fold sign-flip permutation[61] distribution over the 12 date rows, rather than to an asymptotic $X^2(9)$, because n=12 lies well outside the asymptotic regime. Equivalence was evaluated by two one-sided tests (TOST) against a post-hoc margin of ± 0.20 rubric points, the smallest effect size of interest; the choice of margin and its sensitivity analysis are detailed under Equivalence margin and its sensitivity. Equivalence findings are interpreted as a sensitivity bound rather than a confirmatory test.

Computing environment

Hardware. All generation and LLM-as-judge calls were dispatched through the Gemini provider API, so no local GPU was required. Data ingestion, the agent workflow and statistical analyses were executed on a single macOS workstation with 8 CPU cores and 8 GB of system RAM. Under the parallelism settings used for at-scale generation, total runtime for 12 blogs decreased from ~8 hours (serial) to ~1 hour.

Software. The agent workflow is implemented in Python 3.13 and orchestrated by LangGraph, with large state objects offloaded to an in-process data store to reduce memory pressure and independent subtasks executed concurrently via *asyncio.gather*. Long-term memory uses a ChromaDB[62] vector store; the retrieval-augmented verification pipeline (see Retrieval-augmented verification against CDC and WHO guidance) uses the BGE-Large-en-v1.5 embedding model with 2800-token chunks. Surveillance ingestion queried the Delphi COVIDcast API; LLM generation and judging used the temperature settings specified under Model configuration. Mixed-effects modelling estimated marginal means and pairwise tests were performed in R (v4.3) using *lme4*,[63] *lmerTest*,[47] *emmeans*[45] and *tidyverse*; [64] figures were produced with *ggplot2*.[65] Concurrency, retry and rate-limit settings, the ChromaDB configuration, graph-state offloading, and data-source-priority and scale-up resource settings are specified in the released source code.

Author contributions

Y.W., K.Z. and J.H. conceived the study. Y.W. and C.F. developed the methods, software and evaluation framework. Y.W. curated the data and performed the formal analyses. X.A. contributed to visualization. J.Y., Y.F. and Y.J. performed the claim-level evidence annotation and validation. R.R., R.P., A.V., T.E.D., S.G., J.I.W., M.L., M.M., B.T.F., S.C. and D.R. contributed domain expertise, expert evaluation and interpretation of the findings. J.S. and L.S. provided methodological guidance. Y.W. drafted the manuscript. All authors reviewed and edited the manuscript. K.Z. and J.H. supervised the study.

Competing interests

The authors declare no competing interests.

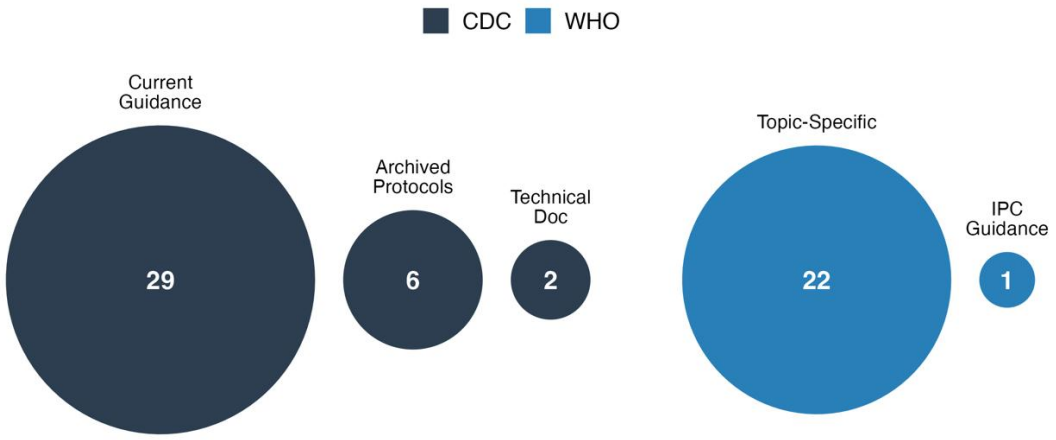
Data availability

The data and reproducibility materials supporting this study have been deposited in an unpublished Zenodo draft. A private read-only reviewer-access link that permits viewing and downloading is provided through the manuscript submission system. The deposit includes the curated CDC and WHO guidance corpus, human evaluation data, LLM-as-judge outputs, quantitative-verification inputs and per-date audit results used in the reported analyses. The record will be made publicly available upon publication. Data originating from the Delphi COVIDcast API remain available from the original provider.

Code availability

The PolicyAgent source code, LangGraph workflow definitions, prompt templates and evaluation scripts are included in the private reviewer-access Zenodo package and correspond to the Git commit recorded in the deposit. The GitHub repository and a versioned Zenodo software archive will be made publicly available upon publication. Model weights are not redistributed, and API access is required for new model generation. The published numerical audit and figures can be reproduced from the archived inputs without rerunning proprietary model calls.

Extended Data



Extended Data Figure 1. |Composition of the curated CDC/WHO knowledge base used for retrieval-augmented verification. Document counts for the contemporary CDC and WHO guidance corpus against which extracted policy claims were matched by the three-gate retrieval verifier (see Retrieval-augmented verification against CDC and WHO guidance). The knowledge base comprises 60 temporally versioned documents across five categories, coloured by issuing authority. CDC contributed 37 documents: current guidance (n = 29), archived protocols (n = 6) and technical documentation (n = 2). WHO contributed 23 documents: topic-specific guidance (n = 22) and infection-prevention-and-control guidance (n = 1).

Extended Data Table 1. | Rubric-dimension scores with 95% CI and paired tests against the expert-written reference. This table is the source for the means and P values reported in the main text and visualized in Fig. 2d.

Dimension	Configuration	Mean (95% CI)	P vs.\ human
Evidence quality	Human	3.90 [3.75, 4.05]	—
	Agentic 2.5 Pro (base)	2.77 [2.66, 2.88]	2.0×10^{-16}**
	Agentic 2.5 Pro	3.05 [2.96, 3.14]	4.5×10^{-14}**
	Agentic 3 Flash	3.16 [3.01, 3.32]	6.8×10^{-9}**
	Agentic 3.1 Pro	3.15 [2.99, 3.31]	8.4×10^{-10}**
Transparency	Human	4.85 [4.76, 4.94]	—
	Agentic 2.5 Pro (base)	3.11 [3.00, 3.23]	2.4×10^{-33}**
	Agentic 2.5 Pro	4.07 [3.86, 4.27]	3.4×10^{-10}**
	Agentic 3 Flash	4.54 [4.38, 4.70]	0.0018**
	Agentic 3.1 Pro	4.82 [4.72, 4.92]	0.641

Dimension	Configuration	Mean (95% CI)	P vs.\ human
Benefit–cost balance	Human	4.13 [3.86, 4.41]	—
	Agentic 2.5 Pro (base)	3.93 [3.73, 4.14]	0.238
	Agentic 2.5 Pro	4.26 [4.13, 4.39]	0.437
	Agentic 3 Flash	4.38 [4.24, 4.52]	0.108
	Agentic 3.1 Pro	4.59 [4.46, 4.72]	0.0038**
Policy intensity	Human	4.18 [3.98, 4.38]	—
	Agentic 2.5 Pro (base)	3.61 [3.38, 3.83]	6.2×10^{-5}**
	Agentic 2.5 Pro	4.21 [4.06, 4.36]	0.784
	Agentic 3 Flash	3.79 [3.65, 3.92]	0.0017**
	Agentic 3.1 Pro	3.79 [3.66, 3.91]	0.0007**
Ideological framing	Human	4.07 [3.85, 4.28]	—
	Agentic 2.5 Pro (base)	3.90 [3.72, 4.09]	0.221
	Agentic 2.5 Pro	4.36 [4.24, 4.48]	0.030*
	Agentic 3 Flash	4.18 [4.04, 4.32]	0.397
	Agentic 3.1 Pro	4.36 [4.23, 4.49]	0.033*
Emotional language	Human	4.38 [4.15, 4.61]	—
	Agentic 2.5 Pro (base)	2.56 [2.41, 2.71]	2.2×10^{-23}**
	Agentic 2.5 Pro	3.30 [3.14, 3.45]	2.8×10^{-11}**
	Agentic 3 Flash	4.26 [4.04, 4.48]	0.411
	Agentic 3.1 Pro	4.46 [4.26, 4.66]	0.587
Stakeholder balance	Human	3.52 [3.26, 3.79]	—
	Agentic 2.5 Pro (base)	4.54 [4.41, 4.67]	1.2×10^{-9}**
	Agentic 2.5 Pro	4.25 [4.11, 4.38]	5.3×10^{-6}**
	Agentic 3 Flash	4.61 [4.45, 4.76]	8.0×10^{-10}**
	Agentic 3.1 Pro	4.95 [4.89, 5.01]	8.0×10^{-16}**
Geographic representation	Human	4.93 [4.83, 5.04]	—
	Agentic 2.5 Pro (base)	4.90 [4.82, 4.98]	0.621
	Agentic 2.5 Pro	5.00 [5.00, 5.00]	0.209
	Agentic 3 Flash	5.00 [5.00, 5.00]	0.209
	Agentic 3.1 Pro	4.98 [4.95, 5.02]	0.370
Vulnerability identification	Human	3.39 [3.12, 3.67]	—
	Agentic 2.5 Pro (base)	4.44 [4.31, 4.58]	3.0×10^{-10}**
	Agentic 2.5 Pro	3.80 [3.61, 4.00]	0.0092**
	Agentic 3 Flash	4.67 [4.54, 4.80]	6.2×10^{-12}**
	Agentic 3.1 Pro	4.89 [4.80, 4.97]	7.8×10^{-15}**
Overall composite	Human	4.15 [4.03, 4.28]	—
	Agentic 2.5 Pro (base)	3.75 [3.66, 3.84]	2.5×10^{-7}**
	Agentic 2.5 Pro	4.03 [3.96, 4.11]	0.115

Dimension	Configuration	Mean (95% CI)	P vs.\ human
	Agentic 3 Flash	4.29 [4.21, 4.37]	0.059
	Agentic 3.1 Pro	4.44 [4.39, 4.50]	4.7×10^{-5}**

Extended Data Table 2. | Ablation of the agentic structure (Gemini 2.5 Pro). Evidence verification accuracy for the raw baseline (Raw LLM), the base workflow without the revision loop and memory (Agentic 2.5 Pro (base)), and the full agentic workflow (Agentic 2.5 Pro), across all 61 publication dates. Adding the revision loop and retrieval/temporal memory increased verification accuracy from 71.0% in the base workflow to 90.8% in the full workflow.

	Raw LLM	**Agentic (base)**	**Agentic**
Dates with verifiable claims	58	51	60
Total quantitative claims	521	217	304
Verified	464	154	276
Not found	57	63	28
Full-accuracy dates	46/61	25/61	45/61
Overall accuracy	89.1%	71.0%	90.8%

Extended Data Table 3. | Rubric-dimension scores for raw-LLM and agentic configurations. Mean rubric-dimension scores (1–5) for raw-LLM and agentic outputs under Gemini 2.5 Pro and Gemini 3.1 Pro. P values are from paired two-sided t-tests; bold marks scores and P values for dimensions with the largest agentic gains.

	Gemini 2.5 Pro		**Gemini 3.1 Pro**		**P (Agent vs. Raw)**	
Dimension	Raw	Agent	Raw	Agent	2.5 Pro	3.1 Pro
Policy intensity	3.61	3.61	4.08	3.79	>0.99	0.015
Evidence quality	2.98	2.77	2.87	3.15	4.9×10^{-4}	0.0055
Ideological framing	3.52	3.90	4.38	4.36	0.0025	0.89
Emotional language	2.62	2.56	2.84	**4.46**	0.51	3.6×10^{-21}
Stakeholder balance	2.62	**4.54**	3.62	**4.95**	4.3×10^{-25}	4.7×10^{-25}
Geographic representation	4.98	4.90	5.00	4.98	0.058	0.32
Transparency	3.00	3.11	2.98	**4.82**	0.070	5.9×10^{-44}
Vulnerability identification	1.98	**4.44**	3.36	**4.89**	4.3×10^{-27}	3.9×10^{-22}
Benefit–cost balance	3.30	3.93	4.56	4.59	4.9×10^{-4}	0.78
Overall composite	3.18	3.75	3.74	4.44	3.9×10^{-13}	9.8×10^{-23}