

Methods and Applications

A Multi-Agent Framework for Real-Time Sentiment Monitoring and Predictive Analysis of Public Health Policies

Yanni Li¹; Yuqi Ma²; Qing Li^{3,†}; Mingming Liang^{4,†}

ABSTRACT

Introduction: Rapid policy rollouts can trigger localized dissatisfaction that is difficult to detect using text-only monitoring and single-pass large language model pipelines. This study aimed to evaluate whether a multimodal, multi-agent framework improves the accuracy, reliability, and early warning sensitivity of public response surveillance during a long-term care policy monitoring window.

Methods: This comparative evaluation study analyzed multimodal public discourse captured during a predefined monitoring window by integrating text with images and videos. The sentiment classification outputs were assessed against a human-consensus reference standard using the F1 score. Summarization reliability was quantified as the rate of unverifiable or fabricated claims in the generated policy feedback summaries. Temporal dynamics were characterized using sentiment trajectories, engagement acceleration, and topic subcluster tracking, with policy-relevant drivers estimated as shares of negative discourse volume.

Results: The multi-agent framework achieved a higher sentiment classification performance, with an F1 score of 0.89 compared with 0.82 for a single-pass baseline. Robustness improved most noticeably in sarcastic and implicit complaint content, where negative intent was consistently recovered despite superficially positive phrasing. Generative reliability improved sharply, with unverifiable or fabricated claims decreasing to 1.2% versus 14.0% from the baseline. Multimodal recovery increased the captured discourse volume by 34% and added 4,200 unique data points available only in the images and videos.

Conclusion: Multimodal multi-agent monitoring strengthened sentiment validity, reduced summary fabrication, and detected topic-level escalation signals in the observed monitoring window. The framework may support earlier identification of policy implementation issues, but its outputs should be

interpreted as decision support signals rather than as substitutes for formal policy evaluation.

The progressive digitization of public discourse has reshaped the epistemic foundations of public health surveillance and policy evaluation (1). As health-related communication increasingly migrates to online platforms, digital traces generated through social media interactions have emerged as critical substrates for understanding population-level perceptions, behavioral responses, and trust dynamics surrounding public health interventions (2). In this context, infodemiology provides a useful framework for studying the distribution and determinants of health information in electronic media and for using Internet-based data to identify health-related signals (3). For public health authorities, the capacity to interpret these signals in near-real time has become indispensable, particularly in the governance of complex and sensitive policies, where delayed comprehension or misaligned expectations can amplify social friction and undermine policy effectiveness (4).

Conventional approaches to policy feedback acquisition have been anchored in retrospective instruments such as structured surveys, focus groups, and deliberative consultations. Although these methods offer methodological rigor and interpretive clarity, their intrinsic latency severely constrains responsiveness, frequently yielding insights only after critical implementation windows have elapsed. Digital social listening techniques have been introduced to alleviate this temporal bottleneck; however, early computational systems dominated by static keyword matching and rudimentary sentiment lexicons exhibited limited semantic acuity (5). These tools struggle to resolve contextual nuances, rhetorical ambiguity, and culturally embedded expressions, thereby conflating substantive concerns with background noise and generating unstable signals that erode decision-maker confidence (6). More recent applications of large language models promise

richer semantic understanding; however, reliance on monolithic models introduces new challenges, including prohibitive computational overhead, constrained contextual awareness, and a well-documented tendency toward generative factual distortion.

This study employs an autonomous multi-agent system architecture for policy-oriented public opinion surveillance to reconcile the demand for real-time sensitivity and analytical robustness. Rather than entrusting the analytical pipeline to a single generalized model, the proposed framework decomposes cognition across multiple specialized agents responsible for information retrieval, critical validation, semantic analysis, and narrative synthesis. Through structured inter-agent collaboration, the system mitigates hallucination risk, enhances contextual grounding, and distributes the computational load more efficiently across continuous data streams. The methodological design was instantiated through an open-source multi-agent framework and operationalized to monitor the public discourse surrounding the implementation of China's long-term care insurance policy, a nationally piloted social security initiative characterized by iterative refinement and heterogeneous local responses.

METHODS

Study Design, Setting, and Observation Window

This infodemiology surveillance study monitored public discourse related to China's long-term care insurance policy implementation from January 1, 2025, to November 30, 2025. Public posts were collected from open-access channels in Sina Weibo and Xiaohongshu. The monitoring scope reflected national policy-related discourse and locally heterogeneous pilot implementation practices rather than a single-city intervention. When available, regional information was obtained from geolocation metadata, post content, hashtags, or account descriptions. This study focused on the public discussion of long-term care insurance implementation, including eligibility assessment, reimbursement procedures, benefit rules, service access, caregiver availability, and implementation-related complaints.

Model Architecture, Prompt Design, and Validation Procedures

The analytic pipeline uses a role-based multi-agent framework rather than a single-pass large language

model (LLM) workflow. It includes retrieval, multimodal parsing, sentiment classification, skeptical validation, evidence verification, and narrative synthesis agents.

The framework was not trained further using private user data. It uses a predefined, prompt-based workflow with fixed task instructions, sentiment rubrics, and evidence requirements. Each model-derived sentiment score or thematic label must include supporting evidence. Outputs without adequate evidence were flagged and excluded from the summary generation. The single-pass LLM baseline used the same input data but generated sentiment labels and summaries in one step without separate skeptical reviews or evidence verification.

Main Exposure: Policy-related Discourse Signals

The primary exposure was time-indexed long-term care insurance policy discourse signals derived from the retrieved posts and engagements. The retrieval process applies a seed vocabulary covering reimbursement, eligibility, benefit design, and service access, with iterative expansion based on stable co-occurrence to capture evolving hashtags and regional expressions. Exposure metrics included post volume and engagement intensity measured by likes, comments, and reposts, transformed using the natural logarithm of one plus count. Multimedia posts were included following optical character recognition, automated transcription, and standardized text normalization. Policy-related posts were retrieved from a predefined Chinese keyword list covering policy identification, eligibility and disability assessment, reimbursement and benefit rules, service access, caregiver availability, and implementation-related complaints. Detailed retrieval vocabulary and exclusion rules are provided in [Supplementary Table S1](#) (available at <https://weekly.chinacdc.cn/>).

Primary Outcome: Sentiment and Verified Thematic Insights

The primary outcomes were model-derived sentiment and thematic labels. Sentiment was quantified on a continuous scale from minus one, indicating maximum negativity, to plus one, indicating maximum positivity. Themes were assigned based on the normalized text, accompanied by evidence fields supporting each inference. A deliberative internal validation mechanism contrasted supportive and skeptical interpretations and reconciled them into final

sentiment values, topic labels, and anomalous flags.

Events were defined as abnormal increases in posting volume accompanied by meaningful sentiment shifts. The alerts were linked to the dominant themes driving each event. A previral signal is defined as a localized topic sub-cluster that shows three features before reaching a conventional trending threshold: a declining sentiment relative to baseline, accelerating engagement, and thematic coherence around a specific implementation issue. The early-warning lead time was calculated as the interval between the first framework-generated alert and the subsequent mainstream-trending event or volume surge.

A random sample of 1,000 posts was manually annotated to establish a human-consensus reference standard. Two trained researchers (YN L and MM L) independently annotated all posts using a predefined rubric that covered sentiment polarity, sentiment intensity, policy-related themes, and supporting evidence. The annotators were blinded to the outputs from both the proposed multi-agent framework and the single-pass LLM baseline. Any disagreements were resolved by consensus.

Covariates and Statistical Analysis

The covariates included the time of day, day of the week, platform source, content modality, and region when geolocation metadata were available. Content modality distinguishes native text posts from text converted from images, video lyrics, or other multimedia sources. The policy phases were identified from announcement- and implementation-related shifts in topic prevalence and posting volume.

Inter-annotator agreement was assessed using Cohen's kappa coefficient. The model performance was evaluated using precision, recall, and F1 scores. Monotonic trends were examined using the Mann-Kendall test. Analyses were conducted using Python (version 3.10, Python Software Foundation, Wilmington, Delaware, USA). Missing metadata resulted in the exclusion of analyses requiring those variables.

RESULTS

During the monitoring window, text-only retrieval captured 12,353 eligible policy-related observations from Weibo and Xiaohongshu. Multimodal parsing through OCR and video transcription recovered an additional 4,200 eligible observations in which policy-relevant information was available only through images

or videos. The proposed multi-agent framework outperformed the single-pass LLM baseline across the classification, reliability, and multimodal recovery indicators (Table 1). The sentiment classification F1 score increased from 0.82 to 0.89, compared to the human-consensus reference standard. In the manual error review, improvements were most evident in posts containing sarcasm, implicit dissatisfaction, or superficially neutral phrasing. After checking the evidence, the rate of unverifiable or fabricated claims in the generated policy feedback summaries decreased from 14.0% to 1.2%.

Applying this validated framework reveals distinct temporal patterns of public responses throughout the monitoring window. Table 2 shows that the early implementation phase, specifically days 1 to 15, was characterized by predominantly neutral sentiment alongside a rapid surge in information-seeking behavior. During this period, the frequency of interrogative keywords increased by approximately 400%, indicating that the early stage responses reflected comprehension deficits rather than overt dissatisfaction. Around day 45, the framework detected a localized negative cluster concerning disability assessment standards, whereas the aggregate sentiment remained stable. The cluster showed declining sentiment, increasing engagement velocity, and coherent discussions around assessment fairness before reaching a conventional volume-based trending threshold. The first alert occurred 72 h before the subsequent mainstream trending event, with the sharing rate increasing by approximately 50% per hour during the pre-trending period. This lead time was treated as an observed early-warning signal within the monitoring window rather than a generalizable prediction estimate.

The semantic decomposition of text data identified the primary mechanisms driving public dissatisfaction. Opacity in assessment procedures and perceived subjectivity in eligibility evaluation emerged as dominant negative drivers, accounting for approximately 28% of the total negative discourse volume. Reimbursement lags and caregiver scarcity followed, representing 15% and 12% of the total volume, respectively. These findings align with the structural ecosystem analysis presented in Figure 1, where integrated insurer-provider-digital care models effectively cluster in regions characterized by high policy standardization. In contrast, provider network management occupies a high-maturity but lower-standardization area. This positioning exposes the

TABLE 1. Comparative performance of the proposed multi-agent framework versus single-pass LLM baseline.

Evaluation domain	Metric definition	Proposed multi-agent framework	Single-pass LLM baseline
Sentiment Classification Performance	F1 score assessed against human-consensus reference standard	0.89, 95% <i>CI</i> : 0.87, 0.91	0.82, 95% <i>CI</i> : 0.80, 0.84
Robustness to Implicit Negativity	Error tendency in sarcasm and implicit complaint content	Substantially reduced misclassification; negative intent more consistently identified in superficially positive phrasing	Moderately improved versus keyword methods, but remained vulnerable to subtle irony
Generative Reliability in Summarization	Rate of unverifiable or fabricated claims in generated policy feedback summaries	1.2%, 95% <i>CI</i> : 0.5%, 1.9%	14.0%, 95% <i>CI</i> : 11.9%, 16.1%
Multimodal Discourse Recovery	Relative increase in captured discourse volume versus text-only scraping	+34%, descriptive recovery estimate	Not applicable
Additional Multimodal-Only Data Captured	Number of distinct data points recovered exclusively from images and videos	4,200, after OCR/transcription quality screening	Not applicable
Early-Warning Performance	Lead time before mainstream trending or volume surge	72 hours before mainstream trending event	Not detected by volume-threshold baseline

Note: The F1 score was assessed against human-consensus labels from 1,000 annotated posts.

Abbreviation: *CI*=confidence interval; LLM=large language model; OCR=optical character recognition.

TABLE 2. Temporal dynamics of public response and dominant negative drivers during the monitoring window.

Analytical dimension	Time anchor within study window	Observed discourse pattern	Quantified signal or estimated share	Policy-relevant interpretation
Early Implementation Phase	Days 1 to 15	Predominantly neutral sentiment accompanied by a rapid rise in information-seeking questions	Mean sentiment near 0.15 with dispersion around 0.20; interrogative keyword frequency increased by approximately 400%	Early-stage response reflected comprehension deficits rather than overt dissatisfaction, with pronounced interpretive ambiguity in lower-tier cities
Latent Crisis Emergence	Around Day 45	Stable aggregate sentiment masked a sharp negative shift within a localized topic cluster	Sub-cluster sentiment declined to approximately minus 0.65	Disability assessment standards became a focal grievance, amplified by emotionally salient multimedia evidence
Predictive Early Warning Performance	Seventy-two hours before trending event	Negative sentiment velocity and rising engagement acceleration preceded mainstream diffusion	Sharing rate increased by approximately 50% per hour despite low absolute volume	Early-warning capability captured pre-viral escalation that would not be detectable through volume thresholds alone
Dominant Negative Driver 1	Full study window	Assessment opacity and perceived subjectivity in eligibility evaluation	Approximately 28% of negative discourse volume	Lack of transparent assessment criteria contributed to distrust and perceived unfairness, especially in rural pilot regions
Dominant Negative Driver 2	Full study window	Reimbursement lag and dissatisfaction with administrative delays	Approximately 15% of negative discourse volume	Delayed fund transfer weakened perceived policy credibility and heightened frustration regarding benefit realization
Dominant Negative Driver 3	Full study window	Caregiver scarcity and anxiety over service availability	Approximately 12% of negative discourse volume	Concerns concentrated in large metropolitan areas where demand outpaced perceived supply of qualified caregivers

Note: Day 1 was defined as the first day of the monitoring window. The baseline period, trending-event threshold, and alert-triggering criteria are predefined.

sector to specific execution risks identified in the textual analysis, including workforce shortages and variations in provider quality. Stakeholder-specific recommendations corresponding to these implementation risks are summarized in [Figure 2](#).

The qualitative mapping illustrated in [Figure 3](#) highlights the significant divergence in sentiment distribution across different media channels. Policy documents and official media coverage maintained a predominantly supportive tone regarding long-term

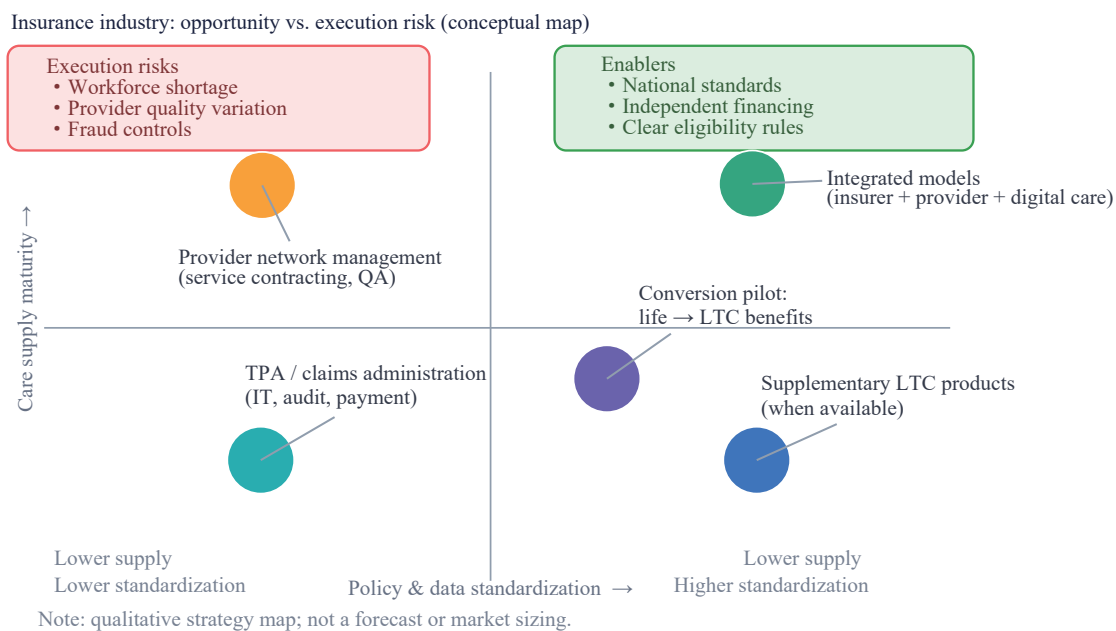


FIGURE 1. Opportunity–execution risk matrix of the long-term care insurance value chain.
Abbreviation: LTC=long-term care.

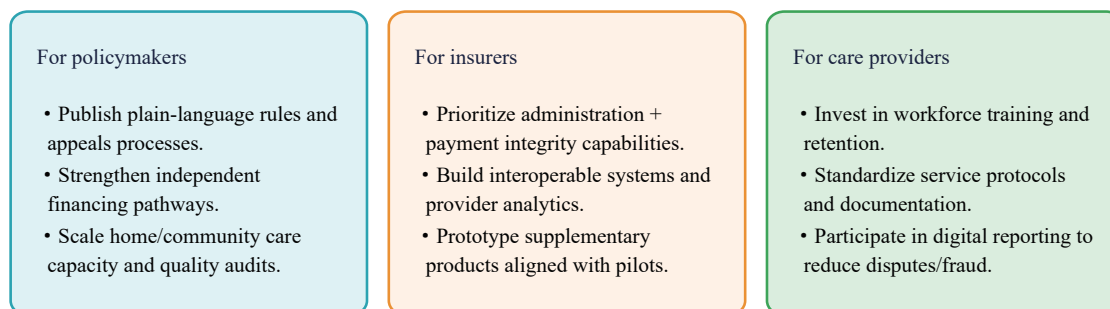


FIGURE 2. Stakeholder-specific recommendations to enable scalable LTC insurance implementation.
Abbreviation: LTC=long-term care.

care reforms. Conversely, incident reports provided a critical counter-narrative, whereas social media discussions exhibited mixed framing with high variability. This stark contrast between the uniform positivity of institutional channels and the complex friction-laden reality observed in social discourse validates the necessity of the proposed multi-agent approach for accurate policy impact assessment.

DISCUSSION

This study suggests that continuous policy-oriented infodemiology can be made both timely and analytically dependable when the monitoring pipeline is structured around specialized, mutually validating components rather than a single monolithic model. The central contribution is not merely improved

predictive or classification performance in isolation, but also the demonstration that reliability constraints that have historically limited social listening for governance can be reduced in an operational setting that resembles real-world surveillance. The observed gains in sentiment consistency, lower incidence of unsupported summary content, and earlier recognition of emerging risks collectively indicate that a multistage internally audited workflow can translate high-velocity public discourse into signals suitable for decision support. In the context of long-term care insurance, where implementation details are locally heterogeneous and public trust is easily destabilized by perceived unfairness, these results imply that digital traces can be used to complement slower feedback instruments with surveillance that is both responsive and traceable (7–8).

These findings align with a growing body of

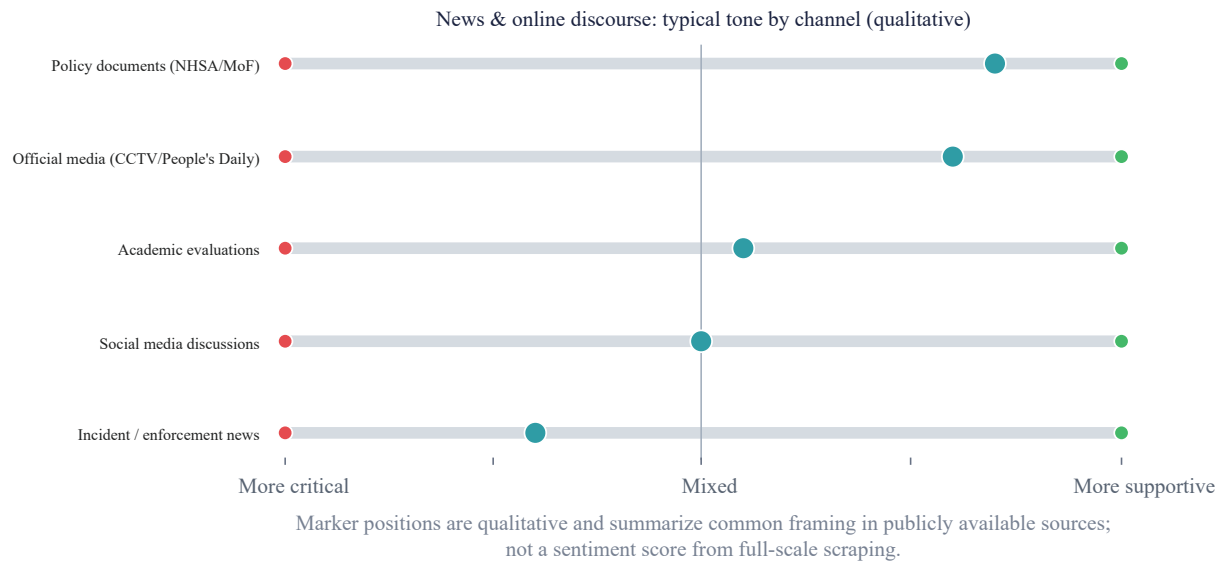


FIGURE 3. Qualitative positioning of typical media tone toward LTC policy by information channel. Abbreviation: LTC=long-term care.

literature showing that infodemiology can capture policy-relevant shifts faster than conventional surveys can, particularly during periods of rapid rollout or controversy, while reinforcing the concerns raised by prior studies on the brittleness of static keyword approaches and lexicon-based sentiment scoring (9–10). Earlier work on health and risk communication repeatedly documented that online discourse often relies on irony, indirect complaints, and culturally specific framing, which can degrade the validity of standard sentiment pipelines and produce signals that are difficult to interpret (11). The present study corroborates the claim that richer semantic modeling improves sensitivity to these forms of expression but also diverges from reports that large language models alone can provide stable monitoring under a continuous load. In practice, single-pass generation and classification can remain vulnerable to drift, context truncation, and unsupported narrative completion (12–13). The present results suggest that distributing the analytical burden across role-specialized agents and requiring internal critique before synthesis may be a pragmatic way to reduce these risks without abandoning high-throughput monitoring.

The practical implications of this study are twofold. For public health authorities and social policy administrators, this study indicates that near-real-time discourse surveillance can provide a structured complement to formal consultation processes by identifying which policy components generate confusion, grievances, or mobilization, and by doing so

within the temporal window when communication and operational adjustments are still feasible (12). Findings of infodemiology research suggest that evaluation should move beyond aggregate accuracy metrics and incorporate measures of generative fidelity, traceability, and resilience to linguistic adversarial language, as governance applications depend as much on credibility as on sensitivity (6). The study also raises the possibility that transparent internal logging and audit trails can increase institutional trust in algorithmic monitoring, as decision makers can trace outputs back to evidence segments rather than relying on opaque summary narratives.

This study had several limitations. The observation window covered an eleven-month implementation-related monitoring period and two open-access social media platforms, which may limit the generalizability to longer policy cycles, other platforms, or private discussion spaces. The dataset reflects open-access posts and therefore underrepresents private discussion spaces that may contain distinct forms of dissent or misinformation. Human labeling, although consensus-based, remains sensitive to annotator interpretation, particularly for sarcasm and moralized language, which may introduce uncertainty into the reference standard. Multimodal conversion introduces additional measurement errors through optical character recognition and speech transcription, which may disproportionately affect dialectal speech, low-quality imagery, and crowded screenshots. The forecasting module was evaluated against a small number of

prominent trending events, which constrained claims regarding sensitivity and specificity across diverse escalation types.

The discourse signals generated by this framework should be interpreted as prompts for follow-up investigations rather than as direct evidence of policy change. Repeated complaints about reimbursement delays may guide administrative process checks, whereas signals concerning caregiver scarcity may support local assessments of service capacity and workforce availability. The framework can help identify when and where confusion, dissatisfaction, or escalation may emerge, but should not be used as a substitute for formal policy evaluation, representative surveys, administrative data analysis, or stakeholder consultation. Policy responses should be based on the triangulation of digital discourse signals and conventional evidence sources.

The multimodal multi-agent framework improved sentiment classification, reduced unsupported generated claims, and identified topic-level escalation signals during the monitoring of long-term care insurance policy discourse. This may help public health and social policy administrators identify implementation issues that require further review, communication adjustments, or service capacity assessments. Its outputs should be used as decision support signals rather than as a substitute for formal policy evaluation.

Conflicts of interest: No conflicts of interest.

Acknowledgments: All researchers involved in manual annotation and data checking for their assistance.

Funding: Supported by the Open Program of the Hospital Management Institute of Anhui Medical University (Grant No. 2024gkjjwz05).

doi: 10.46234/ccdcw2026.139

Corresponding authors: Qing Li, liqing03027@ahmu.edu.cn; Mingming Liang, liangm9036@ahmu.edu.cn.

¹ Tuberculosis Prevention and Control Institute, Anhui Provincial Center for Disease Control and Prevention, Hefei City, Anhui Province, China; ² Department of Hepatology, The Second Hospital of Nanjing, Affiliated to Nanjing University of Chinese Medicine, Nanjing City, Jiangsu Province, China; ³ Department of Obstetrics and Gynecology, Anqing Municipal Hospital, Anqing City, Anhui Province, China; ⁴ School of Health Management, Anhui Medical University, Hefei City, Anhui Province, China.

Copyright © 2026 by Chinese Center for Disease Control and Prevention & Chinese Academy of Preventive Medicine. All content is distributed under a Creative Commons Attribution Non Commercial License 4.0 (CC BY-NC).

Submitted: January 18, 2026

Accepted: June 25, 2026

Issued: July 03, 2026

REFERENCES

1. Salm H, Schuler MK, Hentschel L, Richter S, Hohenberger P, Kasper B, et al. Preferences on treatment decision making in sarcoma patients: prevalence and associated factors - results from the PROSa study. *Oncol Res Treat* 2025;48(4):174 – 85. <https://doi.org/10.1159/000543456>.
2. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. *J Med Internet Res* 2021;23(1):e17187. <https://doi.org/10.2196/17187>.
3. Mavragani A. Infodemiology and infoveillance: scoping review. *J Med Internet Res* 2020;22(4):e16206. <https://doi.org/10.2196/16206>.
4. Holmes G, Tang BY, Gupta S, Venkatesh S, Christensen H, Whitton A. Applications of large language models in the field of suicide prevention: scoping review. *J Med Internet Res* 2025;27(1):e63126. <https://doi.org/10.2196/63126>.
5. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LTJ, Li YCJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci* 2025;32(1):45. <https://doi.org/10.1186/s12929-025-01131-z>.
6. Li HX, Jiang ZH, Guan ZY, Bao YQ, Liu YX, Hu TT, et al. Large language models for diabetes training: a prospective study. *Sci Bull* 2025;70(6):934 – 42. <https://doi.org/10.1016/j.scib.2025.01.034>.
7. Chen LH, Zhang L, Xu XC. Review of evolution of the public long-term care insurance (LTCI) system in different countries: influence and challenge. *BMC Health Serv Res* 2020;20(1):1057. <https://doi.org/10.1186/s12913-020-05878-z>.
8. Kim H, Yoon NH, Hahn Y, Hashimoto H. Explaining variations in long-term care use and expenditures under the public long-term care insurance systems: a case study comparison of Korea and Japan. *Int J Health Policy Manag* 2023;12(1):6640. <https://doi.org/10.34172/ijhpm.2022.6640>.
9. Takahashi M. Insurance coverage, long-term care utilization, and health outcomes. *Eur J Health Econ* 2023;24(8):1383 – 97. <https://doi.org/10.1007/s10198-022-01550-x>.
10. Guo ZJ, Lai A, Thygesen JH, Farrington J, Keen T, Li KZ. Large language models for mental health applications: systematic review. *JMIR Ment Health* 2024;11:e57400. <https://doi.org/10.2196/57400>.
11. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. *Nat Med* 2025;31(6):1873 – 81. <https://doi.org/10.1038/s41591-025-03626-6>.
12. Hershenhouse JS, Mokhtar D, Eppler MB, Rodler S, Storino Ramacciotti L, Ganjavi C, et al. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer Prostatic Dis* 2025;28(2):394 – 9. <https://doi.org/10.1038/s41391-024-00826-y>.
13. Hertz MF, Dierst-Davies R, Freire K, Verlenden JMV, Whitton L, Zimmerman J, et al. Design and implementation of an innovative, rapid data-monitoring strategy for public health emergencies: pilot of the united states school COVID-19 mitigation strategies project. *Public Health Rep* 2023;138(6):878 – 84. <https://doi.org/10.1177/00333549231190050>.

SUPPLEMENTARY MATERIALS

SUPPLEMENTARY TABLE S1. Keyword domains and retrieval strategy for long-term care insurance policy discourse.

Keyword domain	Example Chinese keywords or expressions	English meaning	Inclusion notes
Policy identification	长期护理保险; 长护险; 长期护理; 护理保险; 失能老人护理保险	Long-term care insurance; LTC insurance; long-term care; care insurance	Used to identify posts directly discussing the policy or its commonly used abbreviations
Eligibility and disability assessment	失能评估; 失能等级; 护理等级评估; 评估标准; 评估结果; 重新评估; 申请资格; 准入条件	Disability assessment; disability level; care-needs assessment; assessment standard; assessment result; reassessment; eligibility	Used to capture public discussion about qualification rules, assessment procedures, and disputes over eligibility
Reimbursement and benefit rules	报销; 报销比例; 待遇标准; 支付标准; 基金支付; 个人自付; 结算; 报销延迟; 申请材料	Reimbursement; reimbursement ratio; benefit standard; payment standard; fund payment; out-of-pocket payment; settlement; reimbursement delay; application documents	Used to capture concerns about benefit design, reimbursement procedures, administrative delays, and documentation requirements
Service access and care provision	居家护理; 机构护理; 上门护理; 护理服务; 护理员; 照护人员; 服务机构; 预约困难; 服务不足	Home-based care; institutional care; door-to-door care; care service; caregiver; care worker; service provider; appointment difficulty; service shortage	Used to identify posts related to availability, accessibility, and quality of long-term care services
Implementation complaints and public response	不透明; 不公平; 标准不清; 没人管; 办不下来; 流程复杂; 等太久; 服务不到位; 政策看不懂	Non-transparent; unfair; unclear standards; no one handles it; unable to complete application; complex process; waiting too long; inadequate service; policy hard to understand	Used to capture implicit complaints, dissatisfaction, and implementation-related friction
Platform-specific tags and regional expressions	#长护险#; #失能评估#; #长期护理保险#; 地方简称 + 长护险; 城市名 + 失能评估	Platform hashtags and regional expressions	Added only after repeated co-occurrence with seed terms and manual confirmation of policy relevance

Note: This keyword list was used only for the first-stage retrieval. Newly emerging hashtags, regional expressions, and colloquial terms were added after repeated co-occurrence with seed terms and manual confirmation of policy relevance. Commercial advertisements, general elderly-care marketing, recruitment notices, and unrelated personal narratives were excluded. The final sentiment and thematic labels were assigned after normalization, multimodal extraction, evidence checking, and model-based classification.

Abbreviation: AI=artificial intelligence; CI=confidence interval; F1=harmonic mean of precision and recall; LLM=large language model; LTC=long-term care; OCR=optical character recognition.