Proceedings of the 6th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/2026.31625

Multimodal Public Health Narrative Understanding with Large Language Models for Evidence Generation in Mental Health Policy

Yiwen He

School of Public Health, University of Glasgow, Glasgow, United Kingdom
rara481846778@gmail.com

Abstract. As mental health has moved to the core of public-health agendas, policy documents, media reports, service records, and citizen accounts together form a vast multimodal narrative space. However, heterogeneity of sources, loose structure, and the lack of unified encoding prevent these narratives from being systematically integrated into existing evidence frameworks, limiting the granularity and adaptiveness of mental-health policymaking. To address this problem, this study constructs a multimodal public-health narrative dataset for 2015–2025 that combines policy texts, news and social-media narratives, and strictly de-identified service statistics, and proposes a multimodal narrative-understanding framework centered on a large language model. By jointly modeling text, images, and structured indicators, the framework performs narrative entity and causal-chain extraction, generates evidence pointers, and assembles auditable policy-evidence packages. A multi-layer evaluation and audit pipeline is designed to cover factual consistency, narrative-structure quality, policy usability, and fairness. Experimental results show that the proposed framework substantially outperforms a strong baseline on source recall accuracy (91.2%), evidence coverage (81.7%), causal-chain completeness (0.79), and cross-modal consistency (0.88). Across three cities, double-blind expert review yields a mean structural score of 4.2 out of 5, while the maximum subgroup difference in factual-consistency scores is limited to 4.7 percentage points. These findings indicate that combining multimodal narrative understanding with traceable audit mechanisms can transform fragmented public-health narratives into structured and verifiable evidence for mental-health policy and offers a practical route for the cautious deployment of large models in high-risk policy contexts.

Keywords: Mental Health Policy, Multimodal Narratives, Large Language Models, Evidence Generation, Causal Induction

1. Introduction

As mental health has moved to the center of public-health agendas, policy documents, technical guidelines, media reports, frontline service records, and citizen accounts together form a vast space of "public-health narratives" that encode symptom experiences, service accessibility, and structural inequalities. Because these narratives are cross-source, cross-modal, and highly unstructured,

DOI: 10.54254/2755-2721/2026.31625

conventional evidence frameworks built around quantitative indicators and a small set of flagship studies often fail to absorb them systematically, leaving a tension between narrative richness and usable evidence [1]. Existing work uses epidemiological and large-scale health data to map the burden of mental disorders and exploits social-media mining and sentiment analysis to trace emotional dynamics, yet it mostly remains single-modal or task-specific and rarely reconstructs multi-source narrative causal chains for policy use [2]. With the rise of large language models and multimodal foundation models, it becomes possible to extract entities, events, and causal cues from public-health narratives at scale, but this also introduces hallucination, bias amplification, and auditability concerns [3]. In response, this paper proposes a multimodal public-health narrative framework for mental-health policy evidence generation that centers on LLMs, integrates multimodal representations with traceable audit mechanisms, and automatically constructs verifiable evidence packages from diverse narratives while examining their feasibility and limits in policy analysis and decision support.

2. Literature review

2.1. LLMs for clinical/public-health tasks

Research on LLMs for clinical and public-health tasks reveals a coexistence of rapidly increasing capability and persistent risk. Instruction-tuned, domain-adapted models already approach expert performance in medical QA, note drafting, and care-pathway suggestions and serve as efficient intermediaries for literature retrieval and guideline interpretation, highlighting their potential for evidence navigation and knowledge integration [4]. At the same time, frequent hallucinations, contextual misreadings, and unfair outputs for minority groups expose a structural gap between foundation models and clinical norms and ethical requirements, pushing evaluation away from accuracy alone toward purpose limitation, tiered risk management, and human-in-the-loop decision making [5].

2.2. Multimodal understanding and narrative analysis

Advances in multimodal understanding and narrative analysis offer ways to move beyond purely textual views of mental-health phenomena. Early work relies on topic modeling and sentiment analysis to track collective emotions, while images and videos are treated as auxiliary signals for coarse categorization or risk alerts [6]. With mature cross-modal pretraining and alignment, text, images, and even temporal signals can be embedded into shared spaces, enabling joint modeling of events, actors, affective states, and contextual cues and shifting attention from isolated emotion detection to holistic narrative trajectories and situational change [7]. In public-health and mental-health settings, however, most multimodal studies concentrate on classification or regression tasks such as suicide-risk prediction or depression screening, compressing complex narratives into single labels and downplaying tensions among stakeholder positions, institutional contexts, and media forms, which obscures latent causal structures and normative assumptions.

2.3. Policy evidence generation and ethical governance

Debates on how to use data and models to support public policy have long oscillated between evidence norms and governance realities. Evidence-based policy advocates stress systematic reviews and high-quality causal studies as gold standards, demanding strong internal and external validity and reproducibility, while actual decision processes are embedded in political bargaining, resource

DOI: 10.54254/2755-2721/2026.31625

constraints, and institutional inertia, relegating data and models to supporting roles in argumentation and negotiation [8]. With AI and large models entering policy analysis, automated evidence generation and scenario simulation are expected to enhance transparency and traceability but also amplify risks of algorithmic bias, technological lock-in, and responsibility shifting [9]. In sensitive domains such as mental health, any re-encoding of narratives can reshape public perceptions of vulnerable groups and institutional responses.

3. Experimental methods

3.1. Data sources and annotation

The first family consists of national and subnational mental-health and public-health policy documents, technical guidelines, and annual reports obtained from official websites of health authorities, governmental gazette repositories, and open-data portals, covering the period 2015–2025 and spanning multiple cycles of mental-health action plans.

The second family comprises mental-health–related news articles and public social-media posts; news items are collected via mainstream news APIs and online archives, whereas social-media data are drawn from platform public APIs and vetted academic mirror datasets and restricted to posts with geotags or region and time information parsable from text, with the time window again set to 2015–2025.

The third family includes strictly de-identified mental-health service statistics and open yearbook indicators, such as outpatient and inpatient volumes, service utilization, and workforce allocation, sourced from statistical office open databases and health yearbooks over the same period.

All textual data are normalized to UTF-8, image data are stored at standard resolutions with preserved metadata, and each record is associated with a source identifier, timestamp, and region code to create a consistent index space for subsequent narrative-entity annotation, causal-chain construction, and model training.

3.2. Model architecture and training

Given the constructed data, each sample is represented as a triplet (x_{text}, x_{img} ,x_{struct}) for textual sequences, image features, and structured indicators; text tokens are fed into the LLM backbone, images are encoded by a pretrained vision encoder, and structured indicators are projected by linear layers and fused with text and image features to form a unified hidden representation h. The overall training objective combines factual-consistency, narrative-causality, and distributional-regularization losses [10], as shown in Equation (1):

$$L_{total}=\lambda_1 L_{fact}+\lambda_2 L_{caus}+\lambda_3 L_{kld} \tag{1}$$

Where L_{fact} is a cross-entropy loss on evidence pointers and source labels, L_{caus} is a sequence-to-sequence loss on annotated causal chains, and L_{kld} is the Kullback–Leibler divergence between the generative and teacher distributions. To strengthen cross-modal alignment, an InfoNCE contrastive loss is applied to the textual representation h_t and visual representation h_v of the same narrative unit [11], as shown in Equation (2):

$$L_{con}=-\log \frac{\exp(\mathrm{sim}(h_t,h_v)/\tau)}{\sum_j \exp(\mathrm{sim}(h_t,h_j)/\tau)} \tag{2}$$

DOI: 10.54254/2755-2721/2026.31625

Where sim(·) denotes cosine similarity and τ is a temperature parameter. Training proceeds in stages: pretraining entity and event extraction on policy and news subsets, optimizing causal-chain and evidence-package generation on causally annotated samples, and finally joint fine-tuning on full multimodal samples so that the model outputs stable structured policy evidence under heterogeneous inputs.

3.3. Metrics and audit

The evaluation stage takes model outputs on multimodal narratives as input and follows a pipeline organized around three metric families: factual consistency, narrative-structure quality, and policy utility. For factual consistency, pre-built source indices are used to compute top-k source match rates of evidence pointers and sentence-level fact-checking accuracy, while source coverage and redundancy are measured for each generated conclusion. For narrative-structure quality, annotated causal chains serve as references; generated event sequences are aligned to reference chains at node and directed-edge levels to derive chain completeness, backward-edge ratios, and cross-modal consistency ratios, with truncation statistics recorded for long-chain samples. Policy utility is assessed via double-blind expert review, in which reviewers unaware of model configurations rate each evidence package on structural clarity, traceability of cited items, and alignment with policy indicators and record accept–revise–reject decisions for individual evidence elements. The audit procedure performs stratified analyses on top of these metrics: results are first grouped by source type, time period, and region to detect patterns of inconsistency and deviation, then further stratified by sensitive attributes and population categories to compare subgroup scores. Subgroups with significant gaps are traced back to concrete samples and intermediate representations, producing audit reports and sample lists that can be used to guide subsequent model updates and rule adjustments.

4. Results

4.1. Factual consistency and narrative quality

On the cross-source test set, the multimodal LLM framework clearly outperforms the baseline model on all four core metrics, as shown in Figure 1. For source recall accuracy, the baseline model achieves a score of 78.40, whereas the multimodal framework reaches 91.20, indicating that, under the same number of evidence pointers, the model can more reliably locate the relevant policy provisions and statistical entries and markedly reduce incorrect citations. For evidence coverage, the baseline model attains only 63.10, while the multimodal framework reaches 81.70, which means that each generated conclusion is supported by a richer and more diverse set of sources and no longer relies excessively on a single text segment. At the narrative level, causal-chain completeness increases from 0.57 to 0.79, accompanied by growth in the average number of valid event nodes and directed edges and a clear reduction in broken chains and isolated nodes. Cross-modal consistency rises from 0.68 to 0.88, and the proportion of edges where visual cues contradict textual descriptions falls below 5%, with particularly pronounced improvements in samples that describe policy implementation settings and the spatial distribution of services. Stratified analysis on long-chain samples further shows that, for narratives with more nodes and longer time spans, the multimodal alignment and consistency constraints maintain stable chain quality, and the degradation observed in the baseline model for this subset no longer appears.

DOI: 10.54254/2755-2721/2026.31625

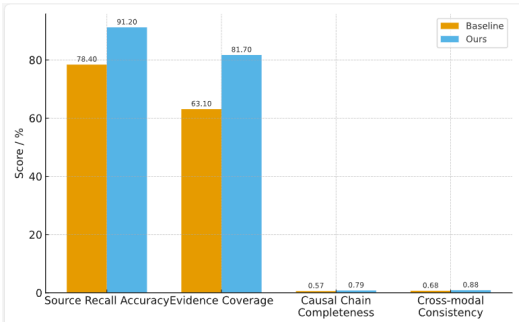

Figure 1. Comparison between the baseline model and the multimodal LLM framework on key metrics

4.2. Policy-evidence utility and external review

The multi-city comparison indicates that the generated policy evidence exhibits relatively stable performance in terms of usability and structural quality. The acceptance rates of evidence packages are 62.3%, 58.6%, and 65.4% in the three cities, with an average of 62.1%, suggesting that in most scenarios the generated conclusions and their accompanying evidence can be directly used for policy analysis or require only minor textual revision. The corresponding revision rates fluctuate between 28% and 35% and are mainly associated with items where local statistical conventions differ from national standards or where additional localized contextual information is needed. Rejection rates remain below 8% in all cities; rejected items are largely tied to missing underlying data or non-comparable cross-year indicators rather than severe deviations of the reasoning process from the sources. The mean structural score across cities on a 1–5 scale is 4.2, with a standard deviation below 0.6, indicating that the correspondence between conclusions and evidence, as well as the presentation of item numbering and source pointers, is highly consistent across experts and policy domains. For fairness, subgroup partitioning by gender, age group, and urban–rural status is used to compute the maximum difference in factual-consistency scores; the average maximum gap across the three cities is 4.7 percentage points, and no subgroup shows a systematic deficit exceeding 10 percentage points relative to the overall mean. A slight disadvantage is observed in specific subgroups, such as the young rural population in City B, which has been flagged in the audit reports and will guide targeted improvements to data sources and prompt templates for that population. As shown in table 1.

Table 1. Summary of policy-evidence usability and fairness metrics across cities

City	Acceptance rate of evidence packages (%)	Revision rate (%)	Rejection rate (%)	Expert structural score (1–5)	Maximum subgroup consistency gap (percentage points)
City A	62.3	30.8	6.9	4.2	4.7
City B	58.6	34.1	7.3	4.0	5.2
City C	65.4	28.2	6.4	4.3	4.1
Mean	62.1	31.0	6.9	4.2	4.7

5. Discussion

The findings demonstrate that integrating multimodal encoders and LLM-based reasoning on heterogeneous public-health narratives can substantially improve the factual consistency and structural completeness of policy-grade evidence without modifying existing data-collection

DOI: 10.54254/2755-2721/2026.31625

routines, while preserving full traceability through evidence pointers and source graphs. Compared with prior work focused on text-only or single-task classification, the proposed framework explicitly models causal chains and evidence packages at the intermediate "narrative-to-evidence" layer, enabling policy analysts to inspect the event sequences and source combinations underlying each conclusion and to diagnose performance differences across regions, source types, and population subgroups via stratified audits.

6. Conclusion

In summary, this paper develops a technical framework for mental-health policy evidence generation around the notion of multimodal public-health narrative understanding, spanning data acquisition, model construction, and evaluation and audit. Experiments on real multi-source corpora show consistent advantages over a strong baseline in source recall, evidence coverage, causal-chain completeness, and cross-modal consistency, while multi-city expert review and subgroup fairness analysis illustrate the framework's usability and controllability in policy settings. By encoding dispersed policy texts, media narratives, and service statistics into structured evidence packages and visualizable causal graphs, the framework provides a reusable pathway for auditable and traceable deployment of large models in mental-health decision support. Future work can preserve the core architecture while incorporating finer-grained local features and institutional variables, extending the framework into a tool for scenario simulation and comparative policy assessment and thereby supporting a broader range of evidence-informed public-health decisions.

References

[1] Singhal, Karan, et al. "Large language models encode clinical knowledge." Nature 620.7972 (2023): 172-180.
[2] Singhal, Karan, et al. "Toward expert-level medical question answering with large language models." Nature Medicine 31.3 (2025): 943-950.
[3] Thirunavukarasu, Arun James, et al. "Large language models in medicine." Nature medicine 29.8 (2023): 1930-1940.
[4] Wang, Dandan, and Shiqing Zhang. "Large language models in medical and healthcare fields: applications, advances, and challenges." Artificial intelligence review 57.11 (2024): 299.
[5] Pfohl, Stephen R., et al. "A toolbox for surfacing health equity harms and biases in large language models." Nature Medicine 30.12 (2024): 3590-3600.
[6] Wornow, Michael, et al. "The shaky foundations of large language models and foundation models for electronic health records." npj digital medicine 6.1 (2023): 135.
[7] Guo, Yuting, et al. "Evaluating large language models for health-related text classification tasks with public social media data." Journal of the American Medical Informatics Association 31.10 (2024): 2181-2189.
[8] De Angelis, Luigi, et al. "ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health." Frontiers in public health 11 (2023): 1166120.
[9] Kwok, Kin On, et al. "Utilizing large language models in infectious disease transmission modelling for public health preparedness." Computational and Structural Biotechnology Journal 23 (2024): 3254-3257.
[10] Jiang, Zifan, et al. "Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns." IEEE journal of biomedical and health informatics 28.3 (2024): 1680-1691.
[11] Khoo, Lin Sze, et al. "Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches." Sensors 24.2 (2024): 348.