Automated Multi-Tier Tagging of Chinese Online Health Education Materials Using a Large Language Model: Development and Validation Study

Jialin Meng, Ruiming Dai, Xiaolan Huang, Yi Gu, Shixing Yan, Xiaoke Wang, Jingrong Gao, Tiantian Zhang

Submitted to: Journal of Medical Internet Research
on: August 29, 2025

Table of Contents

Automated Multi-Tier Tagging of Chinese Online Health Education Materials Using a Large Language Model: Development and Validation Study

Jialin Meng[1] MPH; Ruiming Dai[2] MPH; Xiaolan Huang[3] MPH; Yi Gu[3] MPH; Shixing Yan[2] MS; Xiaoke Wang[2] BS; Jingrong Gao[3] MPH; Tiantian Zhang[1] PhD

[1] Department of Public Health Fudan University Shanghai CN
[2] Shanghai Center for Emerging Technologies Governance in Medicine and Public Health Shanghai CN
[3] Shanghai Municipal Center for Health Promotion Shanghai CN

Corresponding Author:
Tiantian Zhang PhD

Department of Public Health
Fudan University
Dong 'an Road No.130
Xuhui District
Shanghai
CN

Abstract

Background: Effective precision health education and promotion depend on the efficient dissemination of health information. However, current health communication encounters structural bottlenecks, including information overload with insufficient precision matching, variable quality of health resources, and a lack of personalized services. These challenges impede large-scale targeted distribution and audience access. This study aimed to develop and validate an automated tagging system using a large language model (LLM) to enhance the efficiency and equity of health communication and promotion.

Objective: This study aimed to develop, deploy, and validate an artificial intelligence-driven, multi-tier, automated content tagging system to address the core challenges in managing Chinese health education resources and provide a technical foundation for scalable precision health communication.

Methods: We developed a health promotion taxonomy with 10 primary, 34 secondary, and 90,562 tertiary tags using a hybrid method combining a top-down approach (aligned with national standards and expert knowledge) and a bottom-up approach (corpus mining). Subsequently, we constructed an automated tagging system for health promotion materials by fine-tuning a Baichuan2-7B LLM with Low-Rank Adaptation (LoRA), then integrated it with a named entity recognition model and a vector database (Chroma DB), and evaluated its performance.

Results: The final taxonomy included all 16 national priority health domains. The model achieved an overall tag automation rate of 94.8% on the test set, with rates of 97.38% for text-only resources and 89.55% for nontext resources. In a comparative analysis, the model-generated tags demonstrated a higher thematic relevance to the source content than the original manual annotations.

Conclusions: A fine-tuned LLM can efficiently automate the assignment of a granular multilevel tagging system for Chinese health promotion resources. This approach provides a scalable solution to a key bottleneck in health-information management, establishing a technical foundation for advancing precise health communication and improving equitable access to health information.

(JMIR Preprints 29/08/2025:83219)
DOI: https://doi.org/10.2196/preprints.83219

Preprint Settings

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

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

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

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Automated Multi-Tier Tagging of Chinese Online Health Education Materials Using a Large Language Model: Development and Validation Study

Abstract

Background: Effective precision health education and promotion depend on the efficient dissemination of health information. However, current health communication encounters structural bottlenecks, including information overload with insufficient precision matching, variable quality of health resources, and a lack of personalized services. These challenges impede large-scale targeted distribution and audience access. This study aimed to develop and validate an automated tagging system using a large language model (LLM) to enhance the efficiency and equity of health communication and promotion.

Objective: This study aimed to develop, deploy, and validate an artificial intelligence-driven, multi-tier, automated content tagging system to address the core challenges in managing Chinese health education resources and provide a technical foundation for scalable precision health communication.

Methods: We developed a health promotion taxonomy with 10 primary, 34 secondary, and 90,562 tertiary tags using a hybrid method combining a top-down approach (aligned with national standards and expert knowledge) and a bottom-up approach (corpus mining). Subsequently, we constructed an automated tagging system for health promotion materials by fine-tuning a Baichuan2-7B LLM with Low-Rank Adaptation (LoRA), then integrated it with a named entity recognition model and a vector database (Chroma DB), and evaluated its performance.

Results: The final taxonomy included all 16 national priority health domains. The model achieved an overall tag automation rate of 94.8% on the test set, with rates of 97.38% for text-only resources and 89.55% for nontext resources. In a comparative analysis, the model-generated tags demonstrated a higher thematic relevance to the source content than the original manual annotations.

Conclusions: A fine-tuned LLM can efficiently automate the assignment of a granular multilevel tagging system for Chinese health promotion resources. This approach provides a scalable solution to a key bottleneck in health-information management, establishing a technical foundation for advancing precise health communication and improving equitable access to health information.

Keywords: health promotion; large language model (LLM); natural language processing (NLP); tagging; digital health; China; Named entity recognition (NER)

Introduction

Background

Health communication plays a pivotal role in improving health literacy, reducing health disparities, and advancing public health goals [1]. However, in many settings, particularly in low- and middle-income countries, such as China, the digital infrastructure supporting structured health communication remains underdeveloped. A vast and growing repository of publicly available health education materials, spanning articles, videos, and audio, often lacks the standardized metadata and consistent content labeling necessary for effective organization and dissemination[2]. This gap undermines the efficient retrieval, recommendation, and reuse of valuable content, exacerbating the problem of information overload for both clinicians and the public[3].

This challenge is particularly acute in Precision Health Promotion, which aims to tailor health information according to an individual's specific risk profile, behavior, and preferences[1]. Achieving this level of personalization requires a foundation for high-quality finely structured content. However, most health promotion platforms still rely on manual or simple rule-based tagging methods, which are labor-intensive, inconsistent, and difficult to scale[4], particularly across diverse and multimedia formats[5]. Although traditional machine-learning pipelines have been applied to text classification, they often struggle to handle the implicit semantics and domain-specific nuances present in large heterogeneous health corpora[6].

Recent advances in large language models (LLMs) have demonstrated powerful capabilities in abstractive summarization, entity recognition, and domain adaptation, thereby offering promising solutions to this structured labeling bottleneck[7]. When coupled with expert-defined domain ontologies or taxonomies, LLMs have the potential to automate the complex task of assigning relevant and standardized tags to content. However, few empirical studies have validated the performance and real-world utility of such systems, particularly in the context of non-English (eg, Chinese) health promotion resources and deployment within public-sector health agencies[8-10].

To address this gap, the present study developed, implemented, and evaluated an AI-powered, multi-tier tagging system for a large corpus of Chinese language health communication materials. The aim was to enable high coverage, high consistency, and real-time tag assignment across multiple content modalities, thereby laying the technical foundation for a scalable, precision-driven health education delivery system.

Objectives

This study aimed to develop a multi-tier tagging system using a fine-tuned LLM based on health education materials from the Shanghai Municipal Center for Health Promotion. The objectives were as follows: (1) to design a three-level tagging system for Chinese health promotion resources, covering multiple resource modalities by leveraging their textual metadata; (2) to build and fine-tune an LLM-based pipeline to automate the assignment of these tags, targeting an automation rate of at least 90% to enhance efficiency and minimize manual annotation efforts; and (3) to conduct a preliminary assessment of the system’s potential to improve retrieval efficiency and information equity, with the understanding that appropriate retrieval and user-experience metrics will be applied once sufficient usage data become available.

Related Work

Tailored Health Communication Frameworks

Tailored health communication refers to the strategic customization of health messages based on an individual's characteristics, behaviors, and needs. This approach is known to enhance message salience, engagement, and behavioral outcomes[11]. Foundational theories such as the Elaboration Likelihood Model (ELM) and the Health Belief Model (HBM) have demonstrated the importance of aligning content with users’ cognitive and motivational profiles[12]. Systematic reviews have demonstrated that computer-tailored health communication (CTHC) can effectively increase physical activity, improve medication adherence, and empower patients to manage various chronic conditions[13]. However, although these theories support personalization at the message level, their practical application on digital platforms increasingly depends on structured metadata frameworks that enable real-time automated content delivery at scale[14].

Existing Mobile Health Tagging Efforts

Previous tagging efforts in mobile- and web-based health interventions have relied heavily on manually curated taxonomies or rule-based keyword systems. Similarly, Zhang et al applied an ontology-driven annotation model to label health educational materials but noted the difficulty of maintaining semantic consistency across heterogeneous modalities[15]. Although these approaches improve the metadata structure, they typically require substantial manual effort and lack generalizability, particularly in Chinese-language health communication contexts[15].

Limitations of Rule-Based and Shallow Machine Learning Approaches

Rule-based systems and conventional machine learning (ML) models (eg, support vector machines and decision trees) have been applied to health content classification but exhibit several well-documented limitations, including

limited adaptability, domain specificity, and poor performance on short or noisy text inputs[16]. These systems often depend on rigid feature engineering and fail to capture implicit contexts, a critical need in tagging multiformat, user-facing health content. Conversely, large language models (LLMs) have shown promise for abstractive summarization, named entity recognition, and domain adaptation[17]. However, end-to-end validation of LLM-powered tagging pipelines, particularly within the public sector health infrastructure and multilingual multimedia settings, remains limited[16,18].

Methods

Taxonomy and Tagging System Design

We created a three-level taxonomy for Chinese health promotion resources by combining a top-down review of national public health standards with a bottom-up corpus-mining approach. The development process involved two main stages. An initial candidate pool of terms was generated by screening 22 official terminology sets and 15 peer-reviewed studies. The initial list was iteratively refined over the course of two Delphi rounds by a multidisciplinary expert panel (n=20; public health, health education, informatics/AI, clinical specialties, and behavioral science). The panel's task was to reach a consensus on the final hierarchical structure of the taxonomy. The degree of consensus was assessed using the scale-level Content Validity Index (CVI) and Kendall's W coefficient. A detailed description of the screening criteria, expert demographics, and item-level indices is provided in **Multimedia Appendix 1**.

Data Collection and Annotation

A dataset of 10,000 health communication resources was assembled in collaboration with the Shanghai Municipal Centre for Health Promotion (SMCHP). The corpus included a variety of modalities such as text articles, short-form videos, audio clips, and slide decks, each linked to human-curated key phrases provided by non-expert staff. After preprocessing (deduplication, white-space trimming, and removal of off-topic tags), the corpus was partitioned into training (n=7,000), validation (n=2,000), and testing (n=1,000) sets.

The hold-out test set was further stratified by content modality to facilitate a nuanced performance evaluation. This resulted in two subsets: text-rich samples (n=689), comprising articles with abundant textual information, and text-sparse samples (n=311), consisting of multimedia resources with limited textual metadata, such as titles and brief descriptions.

Model Development and Deployment

Large-language-model adaptation

The generative component was Baichuan2-7B, a 7-billion-parameter transformer model pretrained on a 2.6-trillion-token corpus. This model was selected for its state-of-the-art performance on Chinese language benchmarks and its open-source license, permitting research use. To facilitate efficient adaptation of the model to the health communication domain, we utilized Low-Rank Adaptation (LoRA)[19], which is a parameter-efficient fine-tuning (PEFT) method. LoRA freezes the pretrained model weights and injects small, trainable, low-rank matrices into the transformer layers, reducing the number of trainable parameters to less than 0.2% and reducing memory use by approximately 40%. We configured the LoRA adaptation matrices with a rank (r) of 16 and a scaling factor (α) of 32. The training was performed on a single NVIDIA V100 GPU (80 GB) with mixed precision (fp16) for approximately 2 h. We used the AdamW optimizer with an initial learning rate of 1×10^{-4}, linear decay, and a batch size of 16.

Hybrid inference pipeline

At runtime, the system executes a three-stage pipeline to ensure both thematic breadth and terminological precision: (1) the fine-tuned LLM summarizes the input text and proposes a set of candidate tags, (2) a domain-specific Named Entity Recognition (NER) model filters these candidates to retain core thematic terms, and (3) the surviving terms are mapped to canonical labels in the taxonomy via a cosine similarity search within a Chroma DB vector store. This vector database was preloaded with embeddings for all concepts in the taxonomy to ensure that all final outputs were standardized. See Figure 1.

[Insert Figure 1 here]

System Deployment

The final model was packaged as a web service and deployed as a RESTful API for integration into the live content management platform of the SMCHP, utilizing Docker for containerization, Flask for API handling, and CUDA for GPU acceleration. The model development process is detailed in **Multimedia Appendix 2.**

Evaluation Metrics and Statistical Analysis

The evaluation methodology consisted of the following metrics:

1. Inter-Rater Reliability: To measure the degree of consensus, we employed two complementary metrics. Cohen's Kappa (κ) was used to measure the level of agreement beyond that which would be expected by chance. The calculation was performed by treating humans and AI as two raters across all 483,000 possible document-label pairs (1,000 documents × 483 unique labels) to construct the required contingency table. The Jaccard Similarity

Coefficient, a complementary metric well-suited for multi-label tasks, was calculated to measure the overlap between the sets of labels assigned by the human and AI rater for each document. The final score was macro-averaged across 1,000 documents in the test set.

2. Analysis of Disagreement Patterns: To understand the nature of the disagreements, we conducted a granular analysis by categorizing the annotation patterns of all 1,000 documents into four distinct types: (1) Complete Agreement, (2) AI Additive, (3) Human Additive, and (4) Partial Agreement/Conflict.

3. Expert Adjudication of AI-Additive Annotations: To assess the quality and validity of the additional information provided by the AI system, the "AI Additive" cases underwent expert adjudication. A random sample of 50 documents from this category was selected. For each case, a domain expert (senior author of this study) reviewed the source material and additional tags generated by the AI to determine whether they were correct and relevant, followed by calculation of adjudicated precision.

System Robustness and Feasibility

To validate real-world applicability, the final model was packaged as a web service and deployed as a RESTful API for integration into the Shanghai Municipal Center for Health Promotion (SMCHP) content management platform. The service architecture utilizes Docker for containerization, a Flask-based API for handling inference requests, and CUDA for GPU acceleration. The API workflow is structured as follows: (1) the front-end initiates a labeling request and sends it to the back-end, (2) the back-end invokes an automated tagging web service, (3) the tagging model returns standardized labels, and (4) the back-end processes these labels and renders them on the front-end user interface (Figure 1). This workflow enables the platform user interface (workers from the SMCHP) to send a resource to the backend, which invokes the tagging service and returns standardized labels for processing and display.

Results

Taxonomy Development and Validation

The expert-led development process generated a comprehensive three-level hierarchical taxonomy of public health information. After the Delphi procedure, the final tag library comprised ten first-level domains (L1), 34 second-level domains (L2), and 90 562 third-level concepts (L3). The vocabulary was derived from ten national public health standards, a set of 20 official nomenclatures, and the SMCHP's existing three-level practice taxonomy, all of which were harmonized through expert Delphi rounds before being uploaded to the Chroma vector store (See Table 1). Content validity was S-CVI/Ave = 0.91; inter-rater agreement reached Kendall's W = 0.78, meeting the prespecified stopping rule.

Table 1. Overview of the Three-Tier Health Education Taxonomy.

Primary Category (L1)	Examples of Secondary Categories (L2)
Disease Prevention	Chronic non-communicable disease prevention; Communicable disease prevention; Screening and early detection
Disease Care and Treatment	Symptoms and signs; Diseases/conditions; Medications; Diagnostic and therapeutic procedures (tests, surgery)
Rehabilitation and Convalescence	Rehabilitation care; Convalescent care; Functional training
Healthy Living	Nutrition and diet; Smoking/alcohol cessation; Physical activity; Mental health
Health Skills	First aid and emergency; Health management; Self-monitoring skills
Health Policy and Administration	Health policy and regulation; Health education/promotion programs
Health Culture	Folk health knowledge; Traditional health practices
Population Health	Population Health
Specific Groups	Sex/gender groups; age groups; Occupational groups; Special populations
Professional Groups	Public health; Clinical medicine; Nursing; Traditional Chinese Medicine

***** The taxonomy comprises 10 primary categories, 34 secondary categories, and 90,562 tertiary terms. It was developed through two Delphi rounds with 20 experts, and achieved a scale-level Content Validity Index of 0.91.

System Performance and Inter-Rater Reliability

To establish a baseline understanding of the relationship between the AI system outputs and manual annotations from non-expert staff, we first quantified their levels of agreement by treating them as two independent raters. This approach avoids the erroneous assumptions of the gold standard and provides a more appropriate performance assessment.

The analysis, performed on a full test set of 1,000 documents, yielded an overall Cohen's kappa (κ) coefficient of

0.540 and a macro-averaged Jaccard similarity of 0.482. These results indicate a "Moderate Agreement" between the human and AI raters, suggesting a solid foundation of shared logic but also highlighting significant disagreements that require deeper analysis.

To investigate the impact of content modality, we stratified the analysis by the textual richness of the source material. As shown in Table 2, agreement was substantially higher for text-rich articles than for text-sparse multimedia items, which was an expected finding, given the limited semantic context available in the latter.

Table 2. Overall Performance Metrics of the System on the Test Set.

Metric	Text-Rich Samples (n=689)	Text-Sparse Samples (n=311)	Overall (n=1000)
Overall Agreement	99.76%	99.73%	99.73%
Cohen's Kappa (κ)	0.615	0.441	0.54
Jaccard Similarity (Macro-Avg)	0.558	0.389	0.482

To further understand the nature of these disagreements, we categorized the annotation patterns for all 1,000 documents based on the relationship between the tag sets produced by the human annotator and the AI system. The distributions of these patterns are presented in Table 3.

Table 3. Patterns of Agreement and Disagreement Across 1000 Samples.

Annotation Pattern	Sample Count (N=1000)	Percentage	Description
Complete Agreement	583	58.30%	The AI system and the human annotator produced identical sets of tags.
AI Additive	159	15.90%	The AI system included all human-assigned tags and added at least one new, relevant tag.
Human Additive	85	8.50%	The human annotator's tag set was a superset of the AI's.
Partial Agreement / Conflict	173	17.30%	The tag sets had some overlap but also contained unique or conflicting tags.

The analysis revealed that in 58.3% of cases, the AI and human annotators were in complete agreement. More importantly, it highlights that in a substantial proportion of cases (15.9%), the AI system provided more comprehensive thematic coverage by identifying additional relevant topics that were missed in the initial manual annotation. This finding suggests that an AI system may possess a greater recall capacity than a non-expert human annotator.

Qualitative Case Study and Human Comparison

To validate the quality of the additional tags provided by the AI system in the "AI Additive" cases, an expert adjudication was performed. A random sample of 50 documents (out of the 159 "AI Additive" cases) was selected. For each case, a domain expert (senior author of this study) reviewed the source material and additional tags generated by the AI to determine whether they were correct and relevant. Table 4 presents the adjudication results. The high adjudicated precision demonstrates that the model's ability to identify missing themes is highly reliable.

Table 4. Adjudicated Evaluation of AI-Additive Annotations.

Evaluation Category	Sample Size for Adjudication	Adjudicated as Correct and Relevant	Adjudicated Precision
AI-Additive Cases	50	45	90%

An expert review determined that in 90.0% of the sampled cases, the additional tags provided by the AI system were both correct and relevant to the health topic. This high adjudicated precision strongly indicates that the AI system not only provides broader thematic coverage but also has a high degree of accuracy. This finding reframes the role of the system from a simple automation tool to a powerful augmentation instrument capable of enhancing the depth and quality of health content annotation beyond the level achieved by non-expert human staff.

System Deployment and Operational Feasibility

The final model was successfully packaged as a web service and deployed as a RESTful API for integration into the live content management platform of the SMCHP. In the production environment, the system demonstrated high

operational robustness, achieving an overall Tag Automation Rate (TAR) of 94.8% for the test set. Performance varied by modality, with a TAR of 97.38% (671/689) for text-rich samples and 89.06% (277/311) for text-sparse multimedia samples; this difference was significant (two-sample proportion z-test, $P<.001$).

Furthermore, the system demonstrated high technical feasibility, achieving a median end-to-end processing latency of less than one second per resource. This performance met the operational requirements for real-time content ingestion and processing workflows, confirming the system's readiness for practical large-scale applications in a public health setting.

Discussion

Principal Findings

This study successfully developed, validated, and deployed a hybrid AI pipeline for the automated, multilevel tagging of health communication resources within a live municipal public health platform. The system demonstrated high operational robustness, achieving an overall TAR of 94.8% across a diverse test set of 1,000 resources, with strong performance for both text-rich (97.4%) and text-sparse multimedia (89.1%) items.

Recognizing that existing manual annotations were produced by non-expert staff and did not constitute a gold standard, we adopted a more appropriate evaluation framework centered on inter-rater reliability. Our analysis revealed a moderate level of agreement between the AI system and the human annotators (Cohen's Kappa $\kappa = 0.540$; macro-averaged Jaccard similarity = 0.482), indicating that the AI model had learned a significant portion of the logic applied in the manual workflow, while also exhibiting important points of divergence.

The true value of the AI system was determined through a granular analysis of these disagreements. Critically, in 15.9% of all cases, the AI system provided more comprehensive thematic coverage by identifying additional relevant topics that were missed by the human annotator (the "AI Additive" pattern). To validate the quality of these supplementary tags, expert adjudication was performed using a random sample of the patients. The results were compelling; the expert confirmed that the AI's additional tags were correct and relevant in 90.0% of cases. This high adjudicated precision provides strong evidence that the system functions not merely as an automation tool but also as a powerful augmentation instrument capable of enhancing the depth and quality of health content annotation beyond the level of non-expert human staff.

Strengths and Innovations

The primary strength of this study is its end-to-end design, ranging from the rigorous expert-led development of a large-scale taxonomy to the successful deployment and methodologically sound evaluation of an AI system in a real-world operational environment. This study presents several key innovations. First, the hybrid technical architecture, which combines a parameter-efficient fine-tuned LLM with a vector database for standardization, was proven to be both effective and efficient for a large-scale ontology of over 90,000 terms. Second, our use of PEFT via LoRA demonstrated a computationally feasible approach for adapting a 7-billion-parameter model for a highly specific task, making the system maintainable for public health agencies with limited GPU capacity.

Third, and most significantly, this study contributes a robust evaluation framework for validating AI systems in real-world settings where perfect ground truth data are often unavailable, a common challenge that is frequently overlooked in benchmark-focused research; however, it is critical for translating AI from theory to practice. By moving beyond flawed accuracy metrics and using inter-rater reliability analysis coupled with expert adjudication, we provide a more honest and insightful assessment of the true value of AI as an augmentation tool.

Implications for Health Promotion and Information Equity

The successful deployment and validation of this AI-powered tagging system have significant implications for the fields of health education, management, and promotion. By transforming a vast, unstructured repository of health resources into a structured, machine-readable knowledge base, our system serves as a foundational pillar for enabling **Precision Health Promotion**[20]. This approach moves beyond one-size-fits-all communication strategies to deliver tailored health information to individuals based on their specific needs, risks, and preferences. For instance, the granular, multi-level tags generated by our system can power sophisticated recommendation engines, ensuring that a user searching for information on diabetes receives content specifically relevant to their context, such as "gestational diabetes diet" or "type 2 diabetes in adolescents."

Furthermore, this study provides a potential solution to the challenge of health information equity[21]. The digital divide concerns not only access to technology but also the ability to navigate and find relevant information within an overwhelming sea of content[22]. By creating a consistently and comprehensively tagged corpus, our system enhances the discovery of vital health resources, making it easier for public health platforms to serve diverse populations, including those with lower digital or health literacy. These structured data are also invaluable for public health surveillance, allowing health authorities to analyze information-seeking trends, identify community health concerns in near real time, and deploy educational resources more effectively[23,24].

Finally, our findings redefine the role of AI in the public health workflow, not as a replacement for human expertise but as a collaborative partner. The American Medical Association conceptualizes this as Augmented Intelligence,

where the role of AI is to enhance human capabilities. Our system exemplifies this by automating the laborious task of initial tagging with high reliability, freeing human editors from focusing on higher-level strategic tasks such as content validation, campaign design, and addressing complex, nuanced health queries. This human-in-the-loop model is critical for building trust and ensuring the responsible deployment of AI in safety-critical domains such as public health.

Comparison With Prior Work

This study advances the field of automated health content annotation in three ways. First, in contrast to prior studies that largely relied on narrow datasets or flat label sets, we developed and validated our system against a comprehensive three-level taxonomy of over 90,000 terms aligned with national public health standards. This represents an ontology that is one order of magnitude larger and deeper than those used in previous LLM tagging studies, such as ICD coding tasks in Med-PaLM and sentence-level tasks in BioGPT[25,26]

Although other domain-adapted models often rely on full-parameter fine-tuning, which constrains scalability, our use of PEFT via LoRA demonstrates a computationally feasible approach for adapting a 7-billion-parameter model for a highly specific task[19]. This makes the system maintainable and incrementally updatable for local health agencies with limited GPU capacity, which is a critical factor for real-world sustainability that is often highlighted as a barrier to AI adoption in healthcare.

Most significantly, this study moves beyond offline benchmarks to report the end-to-end deployment of an LLM-assisted tagging service within a governmental health promotion platform. To our knowledge, this is one of the first studies to document the operational feasibility, including sub-second latency and integration via a RESTful API, of such a system in a non-English, multimedia public health context[27].

Limitations

This study has several limitations. First, the training and test data were sourced from a single municipal corpus, which may limit the generalizability of our findings to other regions or health systems with different content characteristics. Second, the performance of multimedia resources was constrained by the minimal descriptive metadata available; richer metadata could further improve accuracy. Third, while the expert adjudication was rigorous, it was performed on a random sample (n=50) of the "AI Additive" cases; a larger-scale adjudication could further strengthen these findings. Finally, this study focused on the technical validation of the tagging system. Its downstream impact on user engagement, health information retrieval efficiency, and health literacy outcomes have not yet been evaluated.

Conclusions

This study demonstrated the successful development and real-world implementation of a scalable LLM-powered system for automated health content tagging. By adopting a robust evaluation framework that does not rely on an imperfect gold standard, we have shown that the AI system not only aligns moderately with existing human workflows but, more importantly, serves as a powerful augmentation tool that reliably identifies relevant health topics missed by manual annotators with a high degree of expert-validated accuracy. This study provides both technical and methodological blueprints for implementing AI to enhance, not just automate, precision health communication in public health settings, ultimately paving the way for more efficient, consistent, and equitable access to health information.

Acknowledgments

We would like to acknowledge the raters from the Shanghai Municipal Center for Health Promotion, School of Public Health, Fudan University, and Zhongshan Hospital, Fudan University.

TZ and **JG** were responsible for the study's conceptualization and supervision. **TZ** designed the methodology and was the corresponding author, responsible for writing and revising the final manuscript. **JM** was responsible for the core research implementation, including designing taxonomy and Delphi questionnaires, data curation, performing formal analysis,and drafting the initial manuscript. **RD**, **SY**, and **XW** made significant contributions to the system's development and deployment, which involved packaging the automated tagging system as a web service and RESTful API for deployment within the platform. **XH** and **YG** provided and processed the dataset and offered crucial institutional support and insights for the taxonomy design and validation. All authors read, reviewed, and approved the final manuscript and agree to be accountable for all aspects of the work.

We used the generative AI tool ChatGPT by OpenAI* to review and give some commnets for the draft manuscript for reference, and did not followed all the suggestions from it. The original ChatGPT transcripts are made available as Multimedia Appendix 3.

Data Availability

The real-world corpus curated by the Shanghai Municipal Center for Health Promotion is not publicly available because of third-party ownership and data use agreements but can be obtained from the corresponding author on

reasonable request. Access will require the execution of a data use agreement with the institution and institutional review board. De-identified resources that support reproducibility (eg, taxonomy files, data dictionaries, and synthetic samples), together with code and LoRA adapter weights, are openly available in the Harvard Dataverse. https://doi.org/doi:10.7910/DVN/0W2BMF

Conflicts of Interest
None declared.

Abbreviations
- LLM: Large Language Model
- NER: Named Entity Recognition
- LoRA: Low-Rank Adaptation
- API: Application Programming Interface

Figure and Table List
- **Figure 1**: System architecture workflow.

References

1. Antao EM, Rasheed A, Näher AF, Wieler LH. Reason and responsibility as a path toward ethical AI for (global) public health. NPJ Digit Med. 2025;8(1):329. doi:10.1038/s41746-025-01707-x

2. Oktaviana RS, Handayani PW, Hidayanto AN. Health organization challenges in health data governance implementation: a systematic review. J Infras Policy Dev. 2024;8(6):3892. doi: 10.24294/jipd.v8i6.3892

3. Data overload, EHR and physician burnout | athenahealth. Available from: https://www.athenahealth.com/resources/blog/ehr-usability-information-overload [accessed 2025-05-24]

4. Kucuk E, Cicek I, Kucukakcali Z, Yetis C. Comparative analysis of machine learning algorithms for biomedical text document classification: A case study on cancer-related publications. Med Sci. 2024;13(1):174-4. doi: 10.5455/medscience.2023.10.209

5. Chiche A, Yitagesu B. Part of speech tagging: a systematic review of deep learning and machine learning approaches. J Big Data 2022;9(1):10. doi.org/10.1186/s40537-022-00561-y

6. Sakai H, Lam SS. Large language models for healthcare text classification: a systematic review. *arXiv* [preprint]. March 3, 2025; https://arxiv.org/abs/2503.01159 [accessed 2025-03-15]

7. Goel A, Gueta A, Gilon O, et al. LLMs accelerate annotation for medical information extraction. arXiv [preprint]. December 4, 2023 http://arxiv.org/abs/2312.02296 [accessed 2024-09-13]

8. Zhang Z, Jin L, Huang Y, Li W. Research on chinese medical named entity recognition based on ALBERT and IDCNN. In: Qiu D, Jiao Y, Yeoh W, editors. Proceedings of the 2022 International Conference on Bigdata Blockchain and Economy Management (ICBBEM 2022) Dordrecht: Atlantis Press International BV; 2023. p. 977–986. Available from: https://www.atlantis-press.com/doi/10.2991/978-94-6463-030-5_96 [accessed Aug 28, 2025]ISBN:978-94-6463-029-9

9. Liu F, Liu M, Li M, Xin Y, Gao D, Wu J, Zhu J. Automatic knowledge extraction from Chinese electronic medical records and rheumatoid arthritis knowledge graph construction. Quant Imaging Med Surg 2023 Jun;13(6):3873–3890. doi: 10.21037/qims-22-1158

10. Wei L, Zhao D, Qin L, Liu Y, Shen Y, Ye C. Medical text classification model integrating medical entity label semantics. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi = J Biomed Eng = Shengwu Yixue Gongchengxue Zazhi 2025 Apr 25;42(2):326–333. doi: 10.7507/1001-5515.202408001

11. Noar SM, Harrington NG, eds. *eHealth applications*. 1st ed. New York, NY: Routledge; 2012.

12. Petty RE, Cacioppo JT. The elaboration likelihood model of persuasion Adv Exp Soc Psychol. 1986;19:124–205. doi:10.1016/S0065-2601(08)60214-2

13. Hao L, Goetze S, Alessa T, Hawley MS. Effectiveness of computer-tailored health communication in increasing physical activity in people with or at risk of long-term conditions: systematic review and meta-analysis. J Med Internet Res. 2023;25:e46622. doi:10.2196/46622

14. Chou WY, Prestin A, Lyons C, Wen KY. Web 2.0 for health promotion: reviewing the current evidence. Am J Public Health. 2013;103(1):e9-e18. doi:10.2105/AJPH.2012.301071

15. Li X, Zhang H, Zhou XH. Chinese clinical named entity recognition with variant neural

structures based on BERT methods. J Biomed Inform. 2020;107:103422. doi:10.1016/j.jbi.2020.103422

16. Goodrum H, Roberts K, Bernstam EV. Automatic classification of scanned electronic health record documents. Int J Med Inform. 2020;144:104302. doi:10.1016/j.ijmedinf.2020.104302

17. Le L, Zuccon G, Demartini G, Zhao G, Zhang X. Leveraging semantic type dependencies for clinical named entity recognition. AMIA Annu Symp Proc, AMIA Symp 2022;2022:662–671

18. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc. 2024;31(9):1812-1820. doi:10.1093/jamia/ocad259

19. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. arXiv [preprint]. June 17, 2021. doi:10.48550/arXiv.2106.09685. [accessed 2025-01-27]

20. Lau ECH, Rajput VK, Hunter I, et al. Telehealth and precision prevention: bridging the gap for individualised health strategies. Yearb Med Inform. 2024;33(1):64-69. doi:10.1055/s-0044-1800720

21. Wilson AJ, Staley C, Davis B, Anton B. Libraries advancing health equity: a literature review. Ref Serv Ref; 2023; 51(1):65–762023. doi:10.1108/RSR-09-2022-0037

22. Bui N, Nguyen G, Nguyen N, et al. Fine-tuning large language models for improved health communication in low-resource languages. Computer Methods and Programs in Biomedicine. 2025;263(4):108655. doi:10.1016/j.cmpb.2025.108655

23. Clark EC, Neumann S, Hopkins S, Kostopoulos A, Hagerman L, Dobbins M. Changes to public health surveillance methods due to the COVID-19 pandemic: scoping review. JMIR Public Health Surveill. 2024;10:e49185. doi:10.2196/49185

24. Chiolero A, Buckeridge D. Glossary for public health surveillance in the age of data science. J Epidemiol Community Health. 2020;74(7):612-616. doi:10.1136/jech-2018-211654

25. Luo R, Sun L, Xia Y, et al. BioGPT: Generative pre-trained transformer for biomedical text

generation and mining. 2022; Brief. Bioinform. 23(8) 10.1093/bib/bbac409.

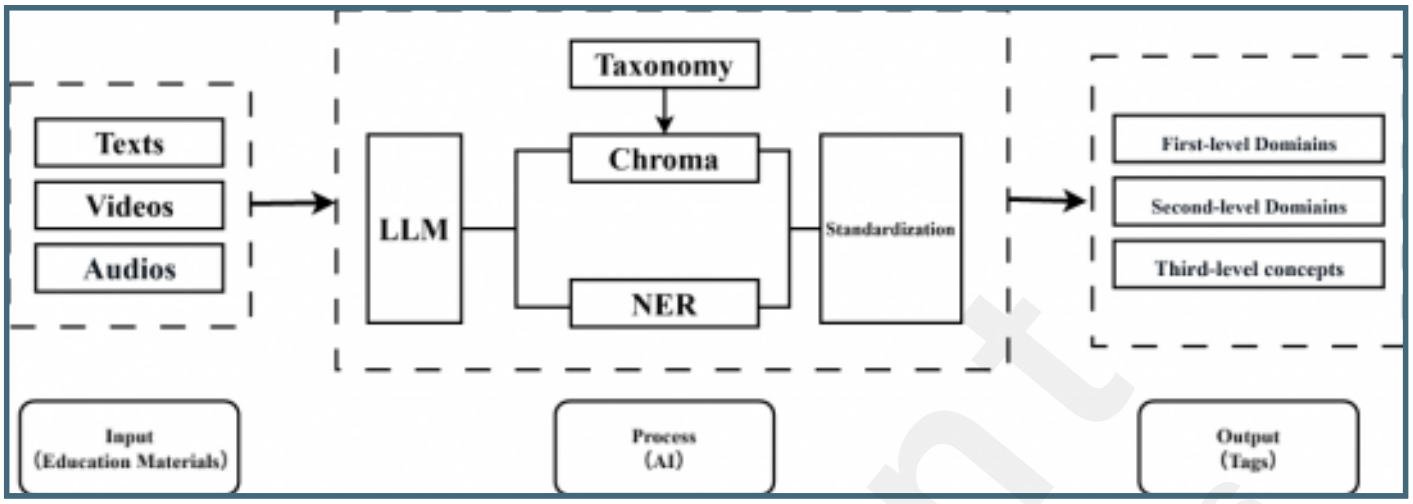
26. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682

27. Sun A, Zhou XH. Estimation of diagnostic test accuracy without gold standards. Stat Med. 2025;44(3-4):e10315. doi:10.1002/sim.10315

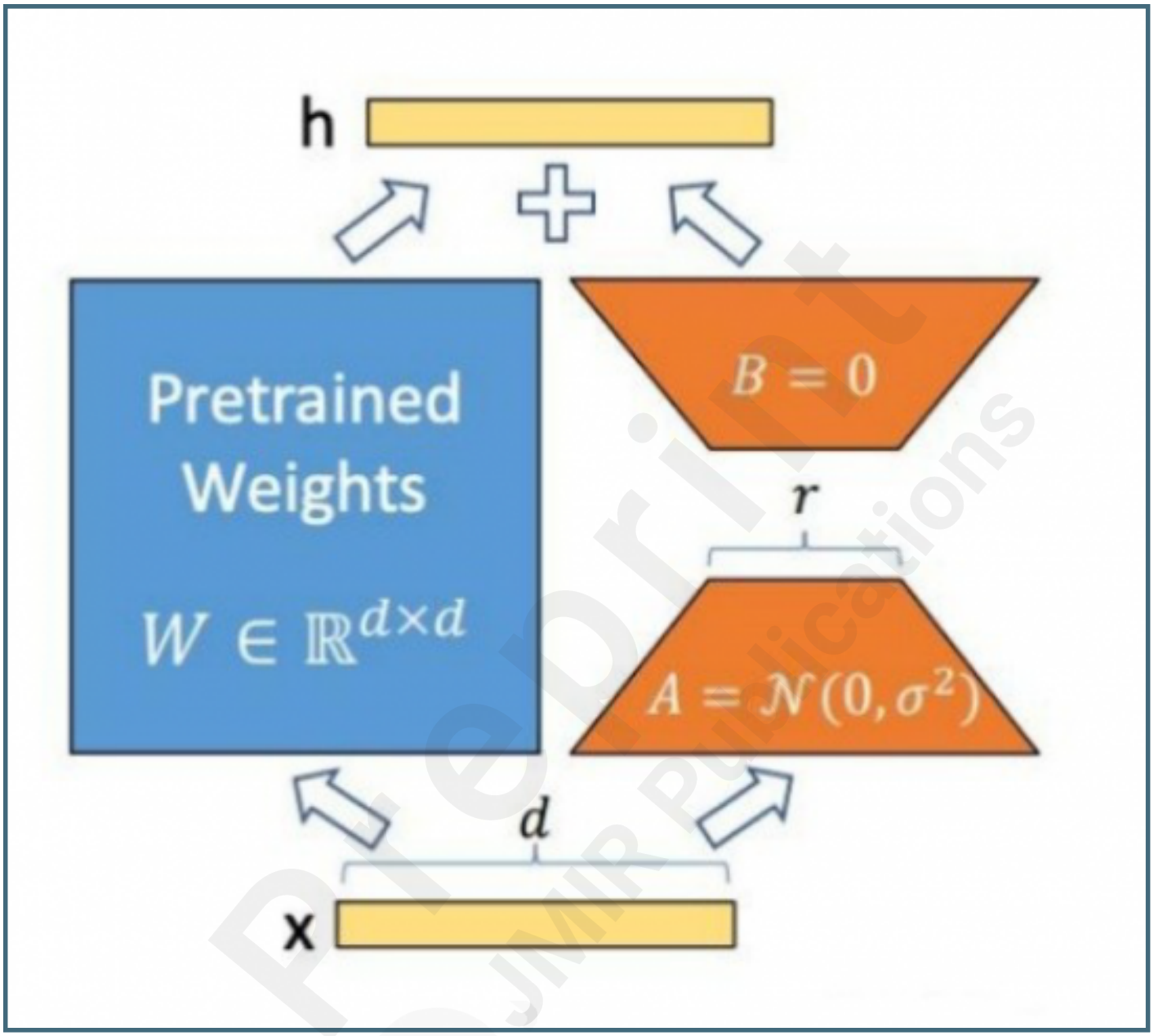
Supplementary Files

Figures

The workflow of the automated tagging system architecture.



LoRA fine-tuning algorithm principles.



Multimedia Appendixes

Full delphi protocol and process for taxonomy.
URL: http://asset.jmir.pub/assets/997b053d3195c748aa8b2a1d263f1939.docx

Model development design and parameter settings.
URL: http://asset.jmir.pub/assets/a5a58bf6c1084b6e420d9807ff3637d0.docx

Chatgpt coversation.
URL: http://asset.jmir.pub/assets/e99cfc38ba65ab023ac7760648aaf862.docx