LLMs and Their Applications in Medical Artificial Intelligence

WENJI MAO, Institute of Automation, Chinese Academy of Sciences, Beijing, China
XIPENG QIU, Fudan University, Shanghai, China
AHMED ABBASI, Human-centered Analytics Lab, University of Notre Dame, Notre Dame, United States

Medical artificial intelligence (AI) is a cross-disciplinary field focused on developing advanced computing and AI technologies to benefit medicine and healthcare. Globally, medical AI has tremendous potential to support the United Nations' sustainable development goals pertaining to health and well-being. In particular, large language models (LLMs) afford opportunities for positively disrupting medical AI-related research and practice. We present a research framework for LLMs in medical AI. Our framework considers the interplay between health and well-being goals, disease lifecycle stages, and the important emerging role of LLMs in medical AI processes related to various lifecycle stages. As part of our framework, we describe the LLM multiplex—important multimodal, multi-model, multicultural, and multi-responsibility considerations for LLMs in medical AI. We discuss how the five articles in the special issue relate to this framework and are helping us learn about the opportunities and challenges for LLMs in medical AI.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Human-centered computing**;

Additional Key Words and Phrases: Large language models, LLMs, medical AI, health, artificial intelligence

ACM Reference Format:
Wenji Mao, Xipeng Qiu, and Ahmed Abbasi. 2025. LLMs and Their Applications in Medical Artificial Intelligence. *ACM Trans. Manag. Inform. Syst.* 16, 2, Article 10 (March 2025), 7 pages. https://doi.org/10.1145/3711837

1 Introduction

The **United Nations (UN) sustainable development goals (SDGs)** include a set of goals related to health and well-being. Globally, these health-related SDGs have motivated a number of public health initiatives and nation-specific long-term targets. For instance, in the United States, Public Health 3.0 calls for timely, actionable, and granular intelligence to improve health and health equity of communities [8, 15]. In Europe, there is discussion on how to advance the UN's health and well-being SDGs across the European Union [11]. "Japan 2035" details the nation's public health targets and aspirational state for their national health system [26]. Similarly, "Healthy China 2030" outlines the country's vision for public health services, the medical industry, as well as food and drug safety [29]. In all such public health initiatives, medical **artificial intelligence (AI)** has an important role to play.

Authors' Contact Information: Wenji Mao, Institute of Automation, Chinese Academy of Sciences, Beijing, China; e-mail: wenji.mao@ia.ac.cn; Xipeng Qiu, Fudan University, Shanghai, China; e-mail: xpqiu@fudan.edu.cn; Ahmed Abbasi, Human-centered Analytics Lab, University of Notre Dame, Notre Dame, Indiana, United States; e-mail: aabbasi@nd.edu.

ACM 2158-656X/2025/03-ART10
https://doi.org/10.1145/3711837

Within the medical AI space, **large language models (LLMs)** present tremendous opportunities for research and practice that enhances health and well-being outcomes at scale. Accordingly, the purpose of this article and accompanying special issue is to foster a dialogue on a research agenda for LLMs in medical AI. We present a research framework for LLMs in medical AI that considers the interplay between health and well-being goals, disease lifecycle stages, and the important emerging role of LLMs in medical AI processes related to various lifecycle stages. As part of our framework, we describe the LLM multiplex—important multimodal, multi-model, multicultural, and multi-responsibility considerations for LLMs in medical AI. We also describe how the five articles in the special issue relate to this framework and are elucidating opportunities and challenges for LLMs in medical AI.

2 A Research Framework for LLMs in Medical AI

Figure 1 presents our research framework for LLMs in medical AI. The framework is composed of four components: (1) relevant health and well-being goals, (2) the four disease lifecycle stages, (3) different types of medical AI processes, and (4) the LLM multiplex. Details are as follows.

2.1 Health and Well-being Goals and Disease Lifecycles

As depicted at the top of Figure 1, the UN's health and well-being SDG includes reducing infant and child mortality rates, maternal mortality, diseases, enhancing mental health and wellness, reducing substance abuse, as well as other important health considerations [22]. A classic mantra in medical research and practice is "prevention is better than cure." This saying refers to the four stages of the disease lifecycle (depicted in the second part of Figure 1): prevention, diagnosis, treatment, and management [27]. Prevention involves periodic checkups, health lifestyle and diet practices, as well as routine screenings [8]. Diagnosis entails identifying illnesses and diseases as well as providing accurate patient prognoses. Treatment relates to alleviation of illness and disease through interventions such as medication, therapy, surgery, and so on [27]. The management stage entails recovery, rehabilitation, and related post-treatment activities. The lifecycle stages are the states and progressions by which the illnesses and diseases associated with the aforementioned well-being goals manifest and progress in individual patients. In the ensuing subsection, we describe how medical AI processes informed by the LLM multiplex can enhance lifecycle outcomes (bottom part of Figure 1).

2.2 Medical AI Processes and the LLM Multiplex

Two common ways in which AI can enhance existing processes is via AI-enabled automation and AI-embedded augmentation [1]. The former entails replacing manual human processes such as diagnosis, question answering, or chat help/support with AI-enabled diagnoses (i.e., machine learning predictions), question answering (LLMs for Q&A), or chat bots. Additionally, with the rise of generative AI, machine learning models such as LLMs can perform an increasing array of zero or few-shot assessment/scoring use cases, as well as generative AI use cases such as generating multi-media responses, having conversations, producing knowledge graphs, and so on [2]. For medical AI processes, this creates a 2×2 of AI-in-the-loop combinations in terms of automation/augmentation and assessment/generation (depicted in the third section of Figure 1).

The state-of-the-art for natural language processing has seen a few major shifts in recent years in text representation learning. Namely, from static word embeddings to contextual embeddings to transformer-based language models culminating with the rise of LLM foundation models [2, 4, 5, 20]. When envisioning the role of LLMs in medical AI, particularly as part of a global public health narrative, a key consideration is the LLM multiplex. The LLM multiplex is a set of multi-dimensional LLM-related factors that are salient in an array of contexts, including medical AI

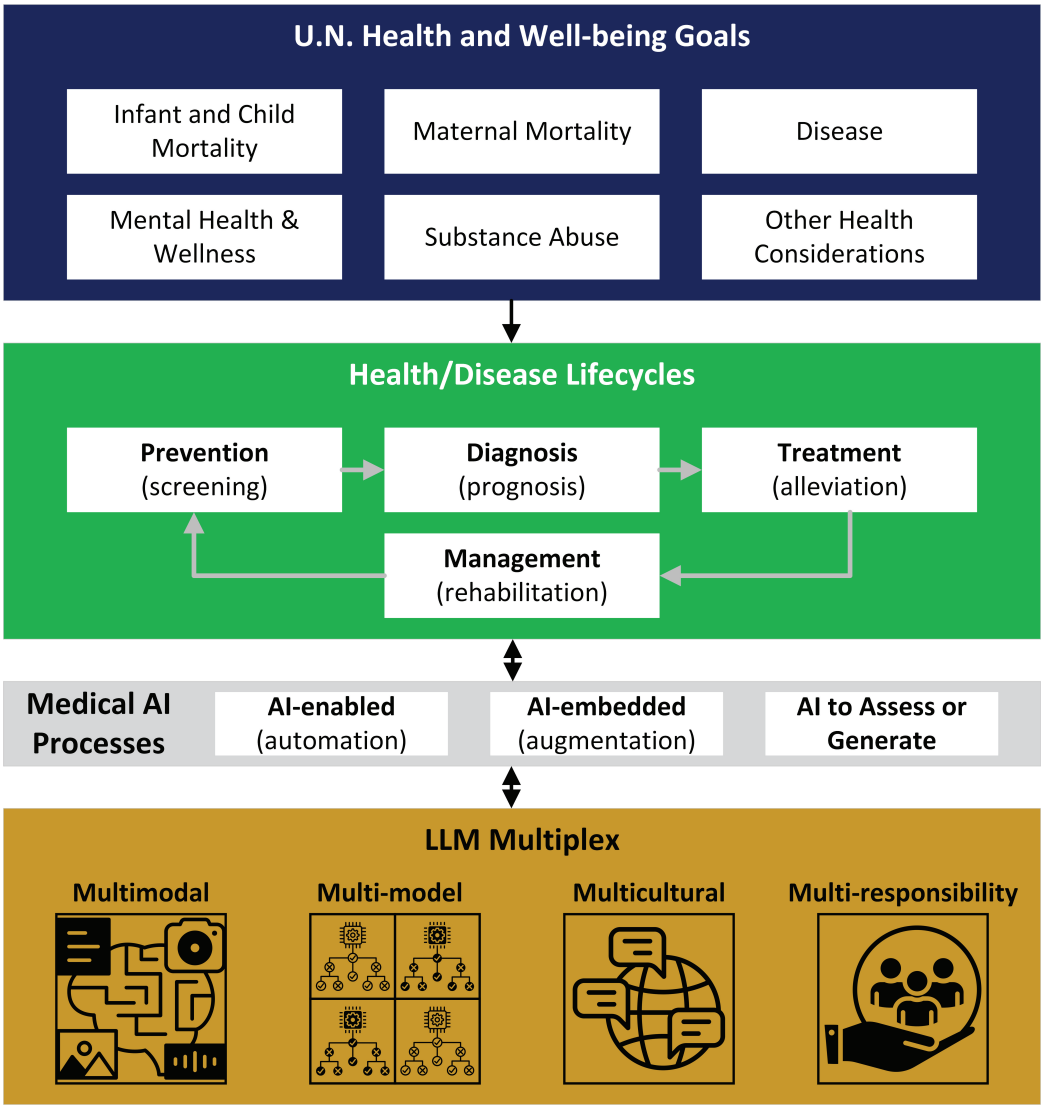

Fig. 1. A framework for LLMs in medical AI.

processes for any and all disease lifecycle stages. Depicted in the bottom part of Figure 1, these factors include multimodal, multi-model, multicultural, and multi-responsibility.

Multimodal refers to the fact that for any medical AI processes encompassing LLMs, patient data is inherently multimodal, potentially spanning speech text, audio, visual, and spatio-temporal input signals [12, 25]. *Multi-model* pertains to the use of modeling pipelines, architectures, and medical AI systems involving multiple models running in parallel, sequentially, and/or arranged in elaborate ensembles [12]. *Multi-cultural* relates to the importance of considering cultural artifacts such as language, customs, beliefs, values, and norms [16, 32]. Recent studies have noted a paucity of research on cultural alignment in LLMs [3]. *Multi-responsibility* relates to the importance of the various tenets of responsible AI, including fairness, privacy, transparency (i.e., explainability and

Paper	Health & Well-being Goals	Lifecycle Stage(s)	Medical AI Processes	LLM Multiplex
Differentially Private Low-Rank Adaptation of Large Language Model Using Federated Learning [21]	Question answering for various health & well-being goals	Prevention, Diagnosis	**AI-enabled generation:** Automatic patient question answering	**Multi-responsibility:** Ensure privacy through federated learning
Language Models for Online Depression Detection: A Review and Benchmark Analysis on Remote Interviews [23]	Mental health and wellness	Diagnosis	**AI-enabled assessment:** Automated depression detection	**Multimodal and Multi-model:** Assess patients' speech to text transcriptions for multiple ASR-LLM combinations
MOSS-MED: Medical Multimodal Model Serving Medical Image Analysis [7]	Visual question answering for various health & well-being goals	Prevention, Diagnosis	**AI-enabled assessment & generation:** Automatic user question answering for medical image analysis	**Multimodal:** Support biomedical vision-language understanding
Zero-Shot Construction of Chinese Medical Knowledge Graph with GPT-3.5-turbo and GPT-4 [30]	Knowledge graph construction for various health & well-being goals	Prevention, Diagnosis, Treatment	**AI-embedded generation:** Human expert annotation coupled with LLMs for graph construction	**Multi-cultural:** Build a knowledge graph for the Chinese medical domain using GPT
ShennongMGS: An LLM-based Chinese Medication Guidance System [9]	Medication guidance for various health & well-being goals	Treatment	**AI-embedded & AI-enabled generation:** Human experts/physicians evaluate the LLM outputs in regards to advice validity, completeness, etc.	**Multi-cultural:** Build a Chinese medication guidance system using bilingual conversant LLMs

Fig. 2. How articles in the special issue relate to the framework.

interpretability), security, and the role of the human in the loop [2, 14]. The fairness considerations in medical AI include disparate impact and lack of equal opportunities [1, 13, 24, 31]. LLMs can potentially exacerbate these concerns across protected attributes such as various demographic dimensions [10, 18], both for upstream representational harm and downstream allocational harm [17]. Similarly, transparency considerations such as interpretability and explainability of diagnosis, prognosis predictions, and optimal treatment strategies are crucial for garnering physician and patient buy-in [28]. Moreover, whereas LLM privacy and security is a major concern in various contexts [6, 19], the sensitive nature of medical AI makes them an issue of paramount importance. We believe this LLM multiplex—the multimodal, multi-model, multi-cultural, multi-responsiblity environments where LLMs have great potential to inform and enhance global public health—are a crucial aspect of the research framework for LLMs in medical AI. In the following section, we describe how the five articles in the special issue shed light on important facets of this framework.

2.3 Articles in the Special Issue

Figure 2 depicts the five articles in the special issue (rows in the figure) and how they relate to the research framework (columns in the figure). Liu et al. [21] examine federated learning strategies for preserving privacy collaboratively fine-tuning LLMs. They focus on the important LLM tasks such as question answering and analyzing patient-physician medical dialogue using two medical testbeds (as well as datasets from other domains). Their work has implications for various health and well-being goals related to the prevention and diagnosis lifecycle stages, with implications for AI-enabled generation. In regards to the LLM multiplex, their work aligns

with multi-responsibility—how to automatically generate answers to patient questions while preserving privacy of those on which the LLM was fine-tuned.

Qin and Cook et al. [23] review and benchmark LLMs for detecting depression from speech to text transcriptions of remote patient-physician interviews. Their work is closely aligned with the mental health global crisis that is overburdening psychiatrists and related mental health professionals. They focus on diagnosis—how well LLMs can detect depression from remote patient interviews. Their research relates to AI-enabled assessment—using LLMs for fine-tuned classification. They connect with the LLM multiplex facet of multimodal (by focusing on speech to text transcriptions). They also relate to multi-model by benchmarking the interaction effects between multiple automated speech recognition models and various downstream classification LLMs.

Dai et al. [7] tackle visual question and answering by proposing a multimodal LLM that can answer questions involving analysis of text and images. Their research relates to various health and well-being goals mostly at the prevention and diagnosis stages, focusing on AI-enabled assessment and generation. They relate to the LLM multiplex aspect of multimodal LLMs that support biomedical vision-language understanding for automatic answering of user questions.

Wu et al. [30] evaluate the efficacy of building Chinese medical knowledge graphs using LLMs such as GPT-3.5 and GPT-4. Beginning with expert annotated knowledge, they use LLMs to construct the graphs. Their work has implications for various health and well-being goals across the prevention, diagnosis, and treatment stages. The work is a nice example of AI-embedded generation where the LLM augments the human experts' annotation efforts with graph construction support. It relates to the multi-cultural aspect of the LLM multiplex as multi-lingual knowledge of Chinese and English must be employed to effectively construct the graphs.

Dou et al. [9] design and develop a Chinese medication guidance system that can automate or augment medication guidance during the treatment phase. They employ a robust evaluation that includes domain experts and physicians, to ensure the LLM guidance is valid, complete, and contains appropriate warnings about adverse drug reactions. Their work relates to the LLM multiplex dimension of multi-cultural LLMs to capture expert knowledge from local physicians and pharmacists, their work employs bilingual conversational LLMs.

3 Conclusions

We conclude with a few possible research directions we anticipate will become more pervasive in the coming years as the research on LLMs in medical AI becomes increasingly robust:

- *From AI Processes to Full-stack AI Companions* - Presently, most LLM artifacts proposed for medical AI focus on one or two stages of the disease lifecycle. As personalized LLMs "meet" precision medicine, research that explores the use of LLMs across all stages of the lifecycle, acting as personalized AI companions to patients, will be at a premium.
- *A Multiplicity in the LLM Multiplex* - The current LLM medical AI research landscape is exploring one or two facets of the LLM multiplex. As LLMs get further integrated into medical AI processes, research is needed to design fully fleshed out multiplex solutions encompassing multiple modalities, models, cultures, and responsible AI tenets.
- *Complex Medical AI Processes Spanning Automation/Augmentation and Assessment/Generation* - Most existing research is focusing on medical AI processes with a single automation or augmentation and assessment or generation characteristic. Consistent with the shift toward omni-lifecycle, LLM multiplex intensive research and practice, we anticipate a need for research on increasingly sophisticated medical AI processes that integrate various combinations of the medical AI process 2×2.

Acknowledgments

We thank the TMIS editor-in-chief, reviewers, and authors for their important contributions to the special issue. Their efforts also informed our ideas around the research framework for LLMs in medical AI.

References

[1] Ahmed Abbasi, Jingjing Li, Gari Clifford, and Herman Taylor. 2018. Make "fairness by design" part of machine learning. *Harv. Bus. Rev.* (2018). Retrieved August 1, 2018 from https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning

[2] Ahmed Abbasi, Jeffrey Parsons, Gautam Pant, Olivia R Liu Sheng, and Suprateek Sarker. 2024. Pathways for design research on artificial intelligence. *Inf. Syst. Res.* 35, 2 (2024), 441–459.

[3] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling culture in llms: A survey. arXiv:2403.15412. Retrieved from https://arxiv.org/abs/2403.15412

[4] Benjamin Ampel, Chi-Heng Yang, James Hu, and Hsinchun Chen. 2024. Large language models for conducting advanced text analytics information systems research. (unpublished). https://dl.acm.org/doi/10.1145/3682069

[5] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv:2108.07258. Retrieved from https://arxiv.org/abs/2108.07258

[6] Yijun Chen and Reuben Kirkham. 2024. Exploring how UK public authorities use redaction to protect personal information. *ACM Trans. Manage. Inf. Syst.* 15, 3, Article 11 (Sept. 2024), 23 pages. https://doi.org/10.1145/3651989

[7] Junqi Dai, Qin Zhu, Jun Zhan, Bo Wang, and Xipeng Qiu. 2024. MOSS-MED: Medical multimodal model serving medical image analysis. (unpublished). https://dl.acm.org/doi/10.1145/3688005

[8] Karen B. DeSalvo, Y. Claire Wang, Andrea Harris, John Auerbach, Denise Koo, and Patrick O'Carroll. 2017. Peer reviewed: Public health 3.0: A call to action for public health to meet the challenges of the 21st century. *Prevent. Chron. Dis.* 14 (2017).

[9] Yutao Dou, Yuwei Huang, Xiongjun Zhao, Haitao Zou, Jiandong Shang, Ying Lu, Xiaolin Yang, Jian Xiao, and Shaoliang Peng. 2024. ShennongMGS: An LLM-based chinese medication guidance system. (unpublished). https://dl.acm.org/doi/10.1145/3658451

[10] Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1012–1023.

[11] George H. Ionescu, Daniela Firoiu, Anca Tănasie, Tudor Sorin, Ramona Pîrvu, and Alina Manta. 2020. Assessing the achievement of the SDG targets for health and well-being at EU level by 2030. *Sustainability* 12, 14 (2020), 5829.

[12] Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O. Cotes, and Gari D. Clifford. 2024. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *IEEE J. Biomed. Health Inf.* (2024).

[13] Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O. Cotes, and Gari D. Clifford. 2024. Evaluating and mitigating unfairness in multimodal remote mental health assessments. *PLOS Digit. Health* 3, 7 (2024), e0000413.

[14] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: A review. *ACM Comput. Surv.* 55, 2, Article 39 (Jan. 2022), 38 pages. https://doi.org/10.1145/3491209

[15] Brent Kitchens, Jennifer L. Claggett, and Ahmed Abbasi. 2024. Timely, granular, and actionable: Designing a social listening platform for public health 3.0. *MIS Quart.* 48, 3 (2024).

[16] Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* 84, 5 (2019), 905–949.

[17] John P. Lalor, Ahmed Abbasi, Kezia Oketch, Yi Yang, and Nicole Forsgren. 2024. Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines. *ACM Trans. Inf. Syst.* 42, 4 (2024), 1–41.

[18] John P. Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3598–3609.

[19] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. arXiv:2004.09984. Retrieved from https://arxiv.org/abs/2004.09984

[20] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open* 3 (2022), 111–132.

[21] Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. 2024. Differentially private low-rank adaptation of large language model using federated learning. (unpublished). https://dl.acm.org/doi/10.1145/3682068

[22] Ana Raquel Nunes, Kelley Lee, and Tim O'Riordan. 2016. The importance of an integrating framework for achieving the sustainable development goals: The example of health and well-being. *BMJ Global Health* 1, 3 (2016), e000068.

[23] Ruiyang Qin, Ryan Cook, Kai Yang, Ahmed Abbasi, David Dobolyi, Salman Seyedi, Emily Griner, Hyeokhyen Kwon, Robert Cotes, Zifan Jiang, and Gari Clifford. 2024. Language models for online depression detection: A review and benchmark analysis on remote interviews. (unpublished). https://dl.acm.org/doi/10.1145/3673906

[24] Ruiyang Qin, Yuting Hu, Zheyu Yan, Jinjun Xiong, Ahmed Abbasi, and Yiyu Shi. 2024. Fl-nas: Towards fairness of nas for resource constrained devices via large language models. In *Proceedings of the 29th Asia and South Pacific Design Automation Conference (ASP-DAC'24)*. IEEE, 429–434.

[25] Salman Seyedi, Emily Griner, Lisette Corbin, Zifan Jiang, Kailey Roberts, Luca Iacobelli, Aaron Milloy, Mina Boazak, Ali Bahrami Rad, Ahmed Abbasi, et al. 2023. Using HIPAA (health insurance portability and accountability act)–compliant transcription services for virtual psychiatric interviews: Pilot comparison study. *JMIR Mental Health* 10 (2023), e48517.

[26] Yasuhisa Shiozaki. 2016. A leadership vision for the future of japan's health system. *Health Syst. Reform* 2, 3 (2016), 179–181.

[27] Richard P. Strong. 1942. *Stitt's Diagnosis, Prevention and Treatment of Tropical Diseases, Vol. 1*. The Blakiston Company.

[28] Thiti Suttaket and Stanley Kok. 2024. Interpretable predictive models for healthcare via rational multi-layer perceptrons. *ACM Trans. Manage. Inf. Syst.* 15, 3, Article 12 (Sept. 2024), 43 pages. https://doi.org/10.1145/3671150

[29] Xiaodong Tan, Xiangxiang Liu, and Haiyan Shao. 2017. Healthy China 2030: A vision for health care. *Value Health Region. Issues* 12 (2017), 112–114.

[30] Ling-I Wu, Yuxin Su, and Guoqiang Li. 2024. Zero-shot construction of chinese medical knowledge graph with GPT-3.5-turbo and GPT-4. (unpublished). https://dl.acm.org/doi/10.1145/3657305

[31] Chenglong Zhang, Varghese S. Jacob, and Young U. Ryu. 2024. Modeling individual fairness beliefs and its applications. *ACM Trans. Manage. Inf. Syst.* 15, 3, Article 14 (Sept. 2024), 26 pages. https://doi.org/10.1145/3682070

[32] Ruike Zhang, Yuan Tian, Penghui Wei, Daniel Zeng, and Wenji Mao. 2024. An LLM-enabled knowledge elicitation and retrieval framework for zero-shot cross-lingual stance identification. In *Findings of the Association for Computational Linguistics (EMNLP'24)*. 12253–12266.

Received 5 December 2024; revised 5 December 2024; accepted 20 December 2024