Original Paper

Evaluating Large Language Model–Generated Virtual Hypertension Patients as Proxies for Real Patient Perspectives: Comparative Study

Yijun Wang[1,2*], MD; Hengyang Liu[3*], MD; Du Tang[4*], MD; Siyi Cheng[5*], MD; Peng Zhao[1], MD; Wenqiang Fang[1], MD; Zeyan Li[6], MD; Bing Wu[7], MD; Long Tang[8], MD; Wenyuan Yin[9], MD; Bowen Zhou[10], MD; Shoupeng Duan[11*], MD; Chen Peng[3*], MD; Wuping Tan[12*], MD; Jun Wang[1,13*], MD

[1]Department of Cardiology, The First Affiliated Hospital of Bengbu Medical University, Bengbu, China
[2]West China School of Medicine, Sichuan University, Chengdu, China
[3]Department of Cardiology, Renmin Hospital of Wuhan University, Wuhan, China
[4]Division of Cardiology, Department of Internal Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[5]Department of Cardiology, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
[6]Department of Cardiology, Guiqian International General Hospital, Guiyang, China
[7]Institute of Clinical Medicine and Department of Cardiology, Renmin Hospital, Hubei University of Medicine, Shiyan, China
[8]Department of Cardiology, The Affiliated Xuancheng Hospital of Wannan Medical University, Xuancheng, China
[9]Department of Electrocardiology, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, China
[10]Department of Cardiology, Suzhou First People's Hospital, Suzhou, China
[11]Department of Cardiology, Wuhan No.1 Hospital, Wuhan, China
[12]Department of Cardiology, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China
[13]National Comprehensive Utilization of Science and Technology Information Resources and Public Service Center, Scientific and Technical Information (STI)-Zhilian Research Institute for Innovation and Digital Health, Beijing, China
*these authors contributed equally

Corresponding Author:

Jun Wang, MD
Department of Cardiology, The First Affiliated Hospital of Bengbu Medical University
No. 287, Changhuai Road, Longzihu District
Bengbu 233000
China
Phone: +86 05523086107
Email: junwang0607@163.com

Abstract

Background: Large language models (LLMs) have shown promise in generating virtual patients for health care applications. However, whether LLM-generated virtual patients can reproduce patient-reported priority rankings remains uncertain.

Objective: We evaluated whether LLM-generated virtual hypertension patients could approximate real patient-reported perspectives, and whether retrieval-augmented generation (RAG) improved alignment with patient-reported priority rankings.

Methods: A total of 20 patients from multiple regions completed structured questionnaires ranking priorities across 7 domains related to AI-assisted hypertension management. A total of 8 LLMs were evaluated under 3 simulation conditions: unconstrained generation, patient-specific simulation, and RAG-enhanced simulation. For unconstrained generation, each model produced 20 virtual patient profiles, each constructed by grouping baseline variables into the study categories and selecting the most frequent category across 10 repeated generations. For patient-specific baseline and RAG-enhanced simulations, each model generated 10 complete questionnaire responses per participant and condition; the modal complete 4-item ranking within each domain was selected as the final ranking. Model-generated rankings were compared with patient-reported rankings using total accuracy, top-1 accuracy, and full accuracy. In addition, 5 hypertension clinicians independently inferred patient-priority rankings while blinded to both real patient responses and LLM outputs, providing a human reference benchmark.

Results: Aggregated unconstrained virtual patient profiles varied across models, with 45%-55% having college-level education, and 75%-95% using antihypertensive medication. In patient-specific baseline simulations, total accuracy ranged

from 46.3% to 71.4%. RAG improved all 3 accuracy metrics across the eight models, increasing mean total accuracy from 62.8% to 76.6%, top-1 accuracy from 51.6% to 72.9%, and full accuracy from 22.2% to 49.4% (all $P<.001$). The largest mean domain-level gains occurred in safety escalation awareness and emotional support and reassurance. Qwen-RAG achieved the highest total accuracy (86.1%) and full accuracy (64.3%). In descriptive comparisons, its corresponding values exceeded those of the clinician consensus (72.5% and 45.0%), while top-1 accuracy was numerically similar (80.7% vs 80.0%).

Conclusions: LLM-generated virtual hypertension patients can moderately approximate real patient perspectives, but unconstrained aggregated profiles differed demographically, behaviorally, and clinically from the enrolled reference cohort under the study conditions. RAG substantially improves concordance between virtual patient rankings and patient-reported priorities and may provide a scalable methodological approach for incorporating patient-perspective proxies into digital health research contexts.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500102798; https://tinyurl.com/2vahxsed

J Med Internet Res 2026;28:e107865; doi: 10.2196/107865

Keywords: large language models; retrieval-augmented generation; hypertension; virtual patients; patient perspectives; patient interviews

Introduction

Hypertension is a major modifiable contributor to cardiovascular morbidity and mortality, underscoring the need for effective prevention, cardiovascular risk prediction, and individualized long-term management [1-4]. Although evidence-based pharmacological treatment and guideline-directed management have substantially improved cardiovascular prevention, long-term blood pressure control remains suboptimal in many settings [1]. Effective hypertension management depends not only on clinical decisions, but also on patients' understanding of disease risk, medication adherence, self-monitoring behaviors, health literacy, emotional responses, and interactions with health care systems [5]. These patient-level factors are particularly important for digital health interventions, where usability, trust, perceived relevance, and alignment with real-world patient needs strongly influence adoption and sustained engagement [6,7]. Despite this progress, the extent to which digital health technologies align with patient-perceived needs-rather than clinically assumed priorities-remains poorly characterized.

Patient perspectives are therefore central to patient-centered cardiovascular care and digital medicine research [6,8]. Interviews, questionnaires, focus groups, and patient-reported outcome studies can identify unmet needs and guide the design of clinically relevant digital tools [6,7]. However, obtaining high-quality real-world patient perspectives is often time-consuming and resource-intensive, especially when studies aim to capture diverse experiences across socioeconomic backgrounds, health literacy levels, disease stages, and health care settings [9-11]. As digital medicine expands, scalable and transparent approaches for incorporating patient perspectives into technology development and evaluation are increasingly needed [12,13].

Large language models (LLMs) offer new opportunities to generate virtual patients capable of simulating human-like communication, preferences, and decision-making [14-16]. In this context, virtual patients refer to LLM-generated responses produced from specified patient profiles to simulate how patients might prioritize support needs within a defined task. They are not intended to function as longitudinally updated digital twins or as synthetic cohorts for population-level simulation [14-19]. In chronic diseases such as hypertension, LLM-generated virtual patients could potentially support early-stage evaluation of patient education materials, digital adherence tools, AI-assisted self-management systems, and clinician-patient communication strategies. However, this potential relies on a critical assumption: that model-generated virtual patients can meaningfully reflect the priorities and concerns of real patients.

This assumption remains insufficiently validated. Most evaluations of LLMs in medicine have focused on clinical knowledge, diagnostic reasoning, examination performance, guideline concordance, or physician-facing decision support [20-27], whereas comparatively few have examined whether LLM-generated patients can reproduce authentic patient perspectives. This distinction is important because generating a clinically plausible patient profile is not equivalent to representing how a real patient perceives risk, prioritizes care needs, or responds to AI-assisted health management. LLMs may encode demographic, socioeconomic, cultural, and behavioral priors from pretraining and instruction-tuning data [28], such that virtual patients appear fluent and clinically coherent while still misrepresenting real-world populations-particularly individuals with lower health literacy, limited digital access, multimorbidity, or less guideline-concordant health behaviors. However, whether LLM-generated virtual patients can reproduce priorities independently reported by real patients remains insufficiently established, particularly in structured cardiovascular chronic-disease management tasks.

Retrieval-augmented generation (RAG) has been used to improve LLM performance in healthcare tasks and has also been incorporated into virtual patient systems for clinical dialogue and medical education [28,29]. However, whether RAG improves the fidelity with which LLM-generated virtual patients reconstruct priorities independently reported by real patients remains unclear.

In this study, we assessed whether LLM-generated virtual hypertension patients could represent real patient perspectives

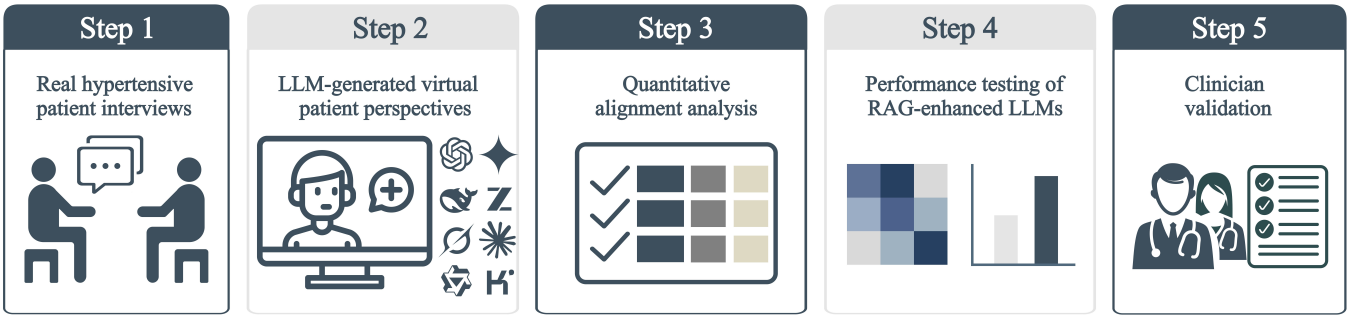
in a structured AI-assisted hypertension management task. We evaluated unconstrained virtual patient profiles generated by 8 LLMs, compared patient-specific model-generated priority rankings with questionnaire responses from real patients, examined the effect of RAG on simulation accuracy, and benchmarked model outputs against clinician-inferred patient-perspective rankings. This study aimed to provide an empirical validation framework for assessing LLM-generated virtual patients as scalable adjunctive proxies for patient-perspective research in cardiovascular digital medicine.

Methods

Study Design and Overview

This study consisted of 5 sequential phases (Figure 1). First, a structured questionnaire was developed to quantitatively capture patient-reported priorities across predefined domains of AI-assisted hypertension management, and structured interviews were conducted with patients to obtain real-world reference rankings. Second, unconstrained virtual patients were generated by 8 LLMs under standardized prompting conditions to describe generation patterns relative to the enrolled reference cohort. Third, patient-specific virtual patients were created by providing individual baseline profiles to each LLM, enabling direct comparison between model-generated and real patient rankings across all domains. Fourth, an RAG framework was implemented and evaluated for its effect on concordance with patient-reported priority rankings. Fifth, 5 hypertension clinicians independently completed an identical ranking task based on standardized patient information packages and generated consensus rankings, which served as a human reference benchmark alongside the best-performing base and RAG-enhanced models.

Figure 1. Overview of the study workflow. Step 1 collected real patient-reported priority rankings using structured interviews and a predefined questionnaire. Step 2 generated virtual patients using 8 large language models. Step 3 evaluated agreement with patient-reported rankings across accuracy metrics and patient subgroups. Step 4 assessed retrieval-augmented generation, using a domain-specific knowledge base, parent-child chunking, and embedding-based retrieval followed by reranking. Step 5 descriptively compared model-generated and clinician-inferred rankings against the patient-reported reference standard. LLMs: large language models; RAG: retrieval-augmented generation.



Ethical Considerations

The study was approved by the Ethics Committee of Suzhou First People's Hospital (approval number SZYYLLky2025011) and the Xuancheng Affiliated Hospital of Wannan Medical University (ethics approval number 2026-1 w005-01), and People's Hospital of Xinjiang Uygur Autonomous Region (ethics approval number KY2026061702), and registered with the Chinese Clinical Trial Registry (ChiCTR2500102798; March 21, 2025). The study was conducted in accordance with the Declaration of Helsinki, Good Clinical Practice principles, and all applicable laws and regulations in China. All participating patients provided written informed consent prior to enrollment. In addition, the cardiologists involved in the blinded evaluation provided written informed consent to participate in the study. Reporting was guided by the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–LLM (TRIPOD-LLM) guideline [30].

Questionnaire Development

A structured questionnaire was developed de novo to enable standardized and quantitative assessment of patient-reported support priorities in the context of AI-assisted hypertension management. Development followed a structured process comprising a review of prior literature and related work, followed by iterative expert review involving cardiologists and digital health researchers.

The final questionnaire comprised 7 predefined domains (Table S1 in Multimedia Appendix 1): (1) medication adherence, (2) blood pressure monitoring and interpretation, (3) lifestyle modification, (4) emotional support and reassurance, (5) health care system interaction, (6) AI trust and acceptability, and (7) safety escalation awareness. Each domain contained 4 candidate support functions. Participants ranked the 4 functions within each domain in order of perceived importance, with rank 1 indicating the highest priority. The same questionnaire was administered to both real patients and LLM-generated virtual patients to ensure direct comparability of responses.

Sample Size Calculation

For the paired comparison between baseline and RAG-enhanced simulations, a power calculation was performed assuming a standardized paired effect size of dz=0.70, a 2-sided α level of .05, and 80% power. Under these assumptions, the minimum required sample size was 19 participants; therefore, enrollment of 20 participants provided approximately 84.4% power to detect an effect of this magnitude.

The unit of analysis was the participant-level paired outcome. This calculation was intended to assess power for the paired comparison and was not designed to establish population representativeness.

Study Population

Participants were recruited between April 6 and April 30, 2026, through hypertension-related WeChat (Tencent Holdings Limited) communities and patient follow-up groups associated with 3 hospitals. Standardized recruitment notices described the study purpose, eligibility requirements, procedures, privacy protections, voluntary nature of participation, and contact information. Interested individuals voluntarily contacted the research team and were screened by trained investigators using a standardized checklist based on the prespecified inclusion and exclusion criteria (Table S2 in Multimedia Appendix 1).

Recruitment used nonprobability convenience sampling, followed by purposive selection among eligible volunteers to enhance geographic, demographic, socioeconomic, clinical, and digital-health diversity within the prespecified sample of 20. Of the 36 individuals screened, seven did not meet the eligibility criteria, and 4 eligible individuals could not be recontacted. Among the remaining 25 eligible candidates, 20 were selected and enrolled. The other 5 eligible candidates were not selected because enrollment was limited to the prespecified sample of 20 and purposive selection prioritized diversity across these characteristics. Eligibility assessment and selection decisions were completed before the patient-priority interviews, without reference to questionnaire rankings or subsequent simulation results. The detailed recruitment procedures and participant flow are presented in Table S3 and Figure S1 in Multimedia Appendix 1, respectively. Before baseline assessment and the structured interview, enrolled participants received standardized study information and provided written informed consent. Participation was voluntary, and deidentified participant information was used in the LLM simulations.

Baseline demographic, socioeconomic, lifestyle, clinical, and digital-health characteristics were collected using a standardized assessment instrument. These variables included age, sex, educational attainment, occupational status, marital status, living environment, household income, hypertension-related characteristics, cardiovascular risk factors, smartphone use, prior experience with LLMs, and previous use of mobile applications for hypertension management. For the clinician benchmark, 5 cardiologists experienced in hypertension management were separately recruited as independent evaluators. All clinician evaluators provided written informed consent before participation.

Model Setup and RAG Framework

A total of 8 LLMs were evaluated: Kimi-K2.6 (Moonshot AI), Qwen-3.6-Max-Preview (Alibaba Cloud), GLM-4.6 (Zhipu AI/Z.ai), DeepSeek-V4-Flash (DeepSeek AI), Claude-Opus-4.7-Fast (Anthropic), Grok-4.20 (xAI), Gemini-3.1-Flash-Lite (Google), and ChatGPT-5.3-Chat (OpenAI). All models were accessed through the OpenRouter platform using standardized API configurations. Detailed model identifiers, version information, and access URLs are provided in Table S4 in Multimedia Appendix 1.

To ensure fair cross-model comparison, all LLMs were evaluated using identical system instructions, patient profile structures, questionnaire content, and response-format requirements. No model-specific prompt optimization or manual adaptation was performed. Prompt templates for unconstrained generation, patient-specific simulation, and RAG-enhanced simulation are provided in Tables S5 and S6 in Multimedia Appendix 1.

An RAG framework was developed to investigate whether external knowledge grounding could improve the fidelity of alignment of LLM-generated priority rankings with patient-reported rankings. A domain-specific knowledge base was constructed from 2 complementary sources: (1) evidence-based clinical guidelines for hypertension management and (2) peer-reviewed literature describing patient interviews, patient-reported experiences, and perspectives related to hypertension care. Detailed parameters of RAG are provided in Table S7 in Multimedia Appendix 1.

For each simulation task, the retrieved contextual information was kept identical across all evaluated LLMs to ensure comparability. The RAG framework modified only the external knowledge context provided to the models, while all other components, including patient characteristics, questionnaire items, system prompts, generation parameters, and output-format constraints, remained unchanged. To avoid information leakage, real patient-reported priority rankings were strictly excluded from retrieval queries, the knowledge base, and all model prompts throughout the study.

For each model-participant combination in the patient-specific baseline and RAG-enhanced conditions, 10 separate simulation runs were conducted using the same input. Within each domain, the 10 outputs were grouped by identical complete 4-item rankings. The frequency of each ranking category was counted, and the category with the highest frequency was selected as the final ranking. We assessed whether the selected complete ranking occurred in at least 6 of the 10 runs, corresponding to a strict majority (>50%), and whether ranking categories were tied for the highest frequency within each model-participant-condition-domain aggregation.

Study Procedures

Real Patient Perspective Acquisition

Following enrollment, participants completed a standardized baseline assessment. Structured interviews were then conducted by trained investigators using the predefined questionnaire. For each of the 7 domains, interviewers presented the 4 candidate support functions and asked participants to rank them according to perceived importance in the context of AI-assisted hypertension management. To minimize interviewer influence, questions were presented verbatim according to the questionnaire wording, and no feedback, additional clarification, or guidance regarding

ranking decisions was provided. Rankings were recorded on standardized case report forms and subsequently entered into a secure study database. These rankings constituted the reference standard against which all LLM-generated and clinician-inferred responses were evaluated.

Unconstrained Virtual Patient Generation

To characterize the intrinsic patient-generation tendencies of each LLM, an unconstrained generation experiment was conducted. Each model was instructed via a standardized prompt to generate one virtual patient profile per invocation, without input from real patient profiles or a predefined target distribution of demographic, socioeconomic, lifestyle, or clinical characteristics. Generated profiles corresponded to the same baseline variables collected from real participants. No questionnaire responses were generated in this phase.

For each model, 20 virtual patients were constructed, each based on 10 repeated profile generations under identical prompting conditions. Before aggregation, baseline variables were grouped into the categories used in the study. Numerical characteristics were represented by the analysis categories shown in Figure 2 and Table S9 in Multimedia Appendix 1, including age group and BMI group derived from height and weight, rather than aggregated as exact numerical values. For each baseline characteristic, the category occurring most frequently across the 10 generations was selected for the final categorical profile used in the descriptive analysis. These profiles were used to describe model-specific patterns and compare their characteristics with those of the enrolled participant cohort.

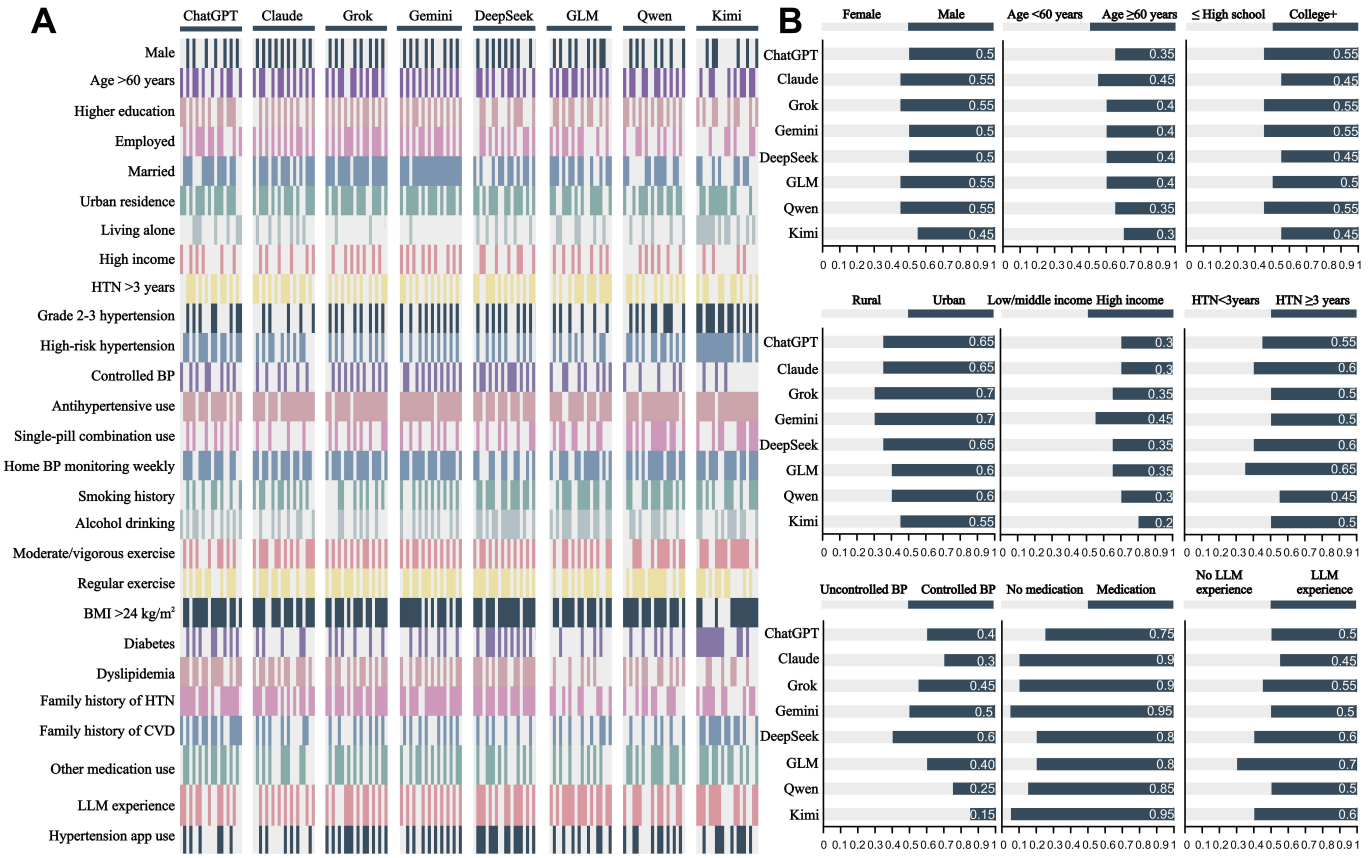


Figure 2. Distributional characteristics and systematic tendencies of unconstrained large language model–generated virtual patient profiles across 8 models. BP: blood pressure; LLM: large language model; CVD: cardiovascular disease; GLM: general language model; HTN: hypertension.

Patient-Specific Virtual Patient Simulation

To evaluate LLM performance in reconstructing individual patient-reported priority rankings, patient-specific simulations were conducted using the enrolled participants' baseline profiles. For each participant, all available baseline characteristics were incorporated into a standardized simulation prompt. Models were instructed to respond to the structured questionnaire from the perspective of the described patient.

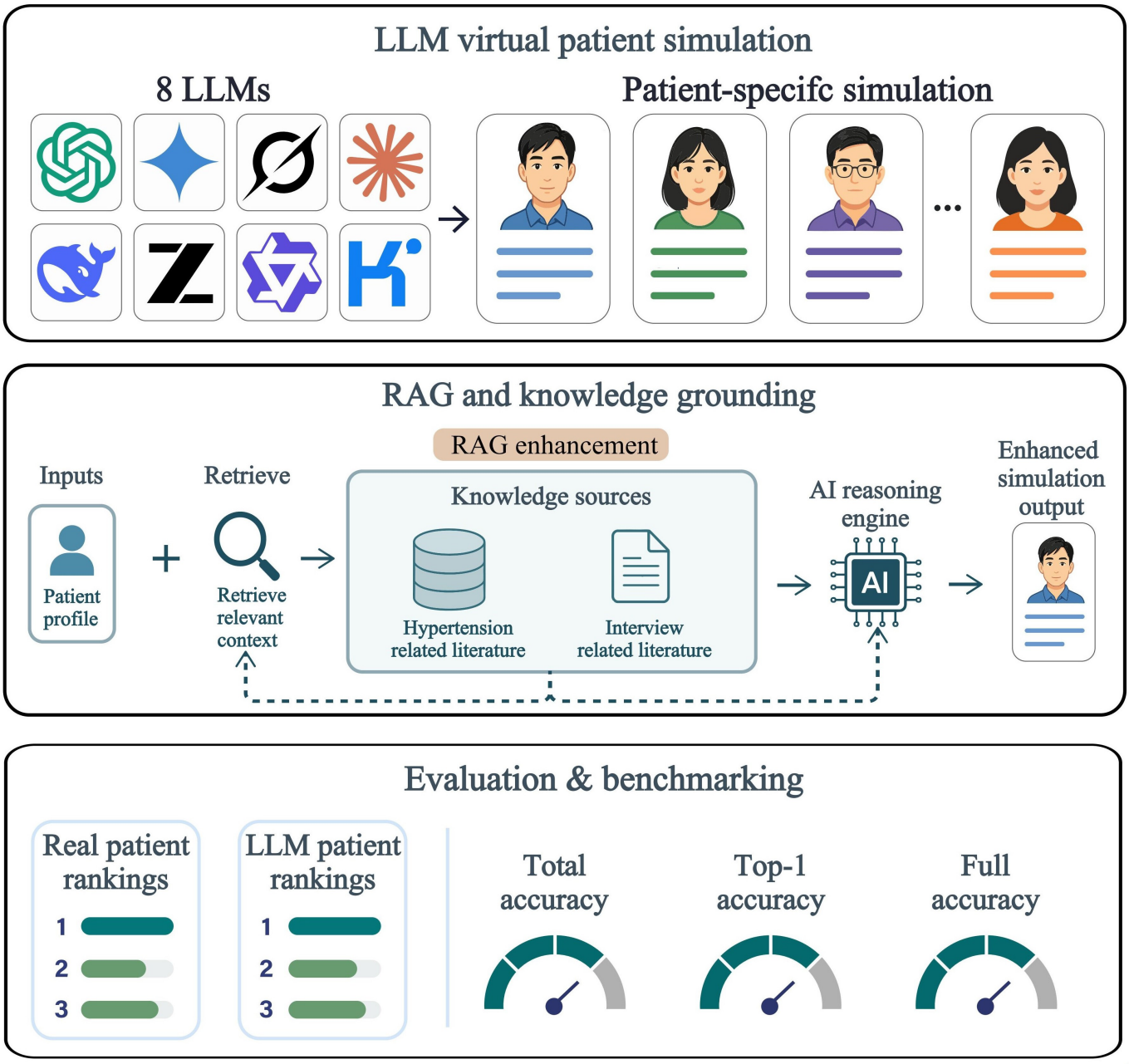
Each of the 8 LLMs generated 10 complete questionnaire responses per enrolled participant under identical prompting conditions. Within each domain, the complete 4-item ranking occurring most frequently across the 10 responses was selected as the final ranking. Each simulation produced rankings across all 7 domains, enabling direct comparison with the corresponding real patient responses.

RAG-Enhanced Simulation

Following the baseline simulation phase, the patient-specific simulation procedure was repeated in full using the RAG-enhanced versions of all 8 models under identical experimental conditions. Baseline profiles were provided

to each RAG-enhanced model, and the same questionnaire was administered. The RAG framework retrieved relevant documents from the domain-specific knowledge base for each patient simulation and appended them to the prompt as contextual augmentation (Figure 3).

Figure 3. Overview of retrieval-augmented generation–enhanced virtual patient simulation and evaluation against real patient-reported priorities. LLMs: large language models; RAG: retrieval-augmented generation.



Clinician Blinded Evaluation

To provide a human comparator for the patient-priority inference task, 5 hypertension specialists independently completed the same 7-domain ranking questionnaire. Each clinician reviewed a standardized patient information package corresponding to that used in the LLM simulations and was blinded to the real patient-reported and LLM-generated rankings. Clinicians were instructed to rank the support functions from the presumed perspective of the corresponding patient. After completing the task independently, they jointly reviewed discordant rankings and resolved disagreements through structured consensus discussion to produce one final clinician ranking for each patient and domain. If consensus could not be reached, the prespecified adjudication procedure assigned the final decision to the clinician with the longest relevant clinical experience. The clinician consensus served as the human comparator; the corresponding real patient-reported rankings remained the reference standard.

Data Processing and Response Coding

Responses obtained from both real patients and LLM-generated virtual patients were transformed into a structured dataset for subsequent analysis. For each of the 7 questionnaire

domains, participants and virtual patients were required to rank 4 candidate support functions according to perceived importance.

Rankings were converted into ordinal values ranging from 1 to 4, with lower values indicating higher priority. For each participant and virtual patient simulation, a complete ranking profile was generated across all 7 domains. These structured rankings were used to derive domain-level and participant-level outcome measures for subsequent analyses.

For patient-specific simulations, rankings generated by virtual patients were directly matched to the corresponding rankings provided by the real participant from whom the baseline profile had been derived. Domain-specific rankings and overall ranking patterns were subsequently used to evaluate accuracy between real and virtual patient perspectives.

Data extraction and coding were performed using standardized procedures prior to statistical analysis. All coding and verification processes were completed, with discrepancies resolved through discussion and consensus.

Study Outcomes

The primary outcome was total accuracy, a normalized score measuring agreement between LLM-generated and patient-reported priority rankings. For each patient and domain, each of the four support functions received an item-level score of 4 minus the absolute difference between its generated and patient-reported rank positions. The 4 scores were summed to yield a domain-level score ranging from 8 to 16, with 16 indicating an exact match. Scores across the 7 domains were then summed and linearly rescaled from their theoretical range of 56-112 to 0%-100%. Higher values indicate greater ranking agreement. Detailed calculations are provided in Table S8 in Multimedia Appendix 1.

Secondary outcomes included top-1 accuracy, defined as the proportion of domains in which the LLM-predicted highest-priority function exactly matched the patient-reported highest-priority function, and full accuracy, defined as the proportion of domains in which the complete 4-item ranking generated by the LLM exactly matched the patient-reported ranking. Performance was additionally compared across LLMs, simulation strategies, questionnaire domains, and patient subgroups. Exploratory subgroup analyses were conducted according to sex, educational attainment, household income, residential setting, prior LLM experience, and blood pressure management application use.

Statistical Analysis

Rankings were analyzed using Python. Continuous variables were summarized as mean (SD) or median (IQR), and categorical variables as counts and percentages. Total accuracy was treated as a continuous composite rank-agreement score and reported as point estimates. Inferential analyses used participant-level outcomes aggregated across seven domains. Baseline-RAG comparisons used paired Wilcoxon tests with exact sign-permutation P values, without multiplicity adjustment; overall comparisons also averaged outcomes across models. All tests were 2-sided, with $P<0.05$ considered significant after adjustment where applicable.

Data Security

All patient-level data were deidentified prior to analysis and LLM simulation. Direct identifiers—including names, contact details, medical record numbers, and exact addresses—were removed and were not incorporated into any model prompt. Simulation inputs contained only study-relevant demographic, clinical, and digital health variables. Real patient-reported priority rankings were withheld from all prompts, retrieval queries, and the RAG knowledge base throughout the study to prevent data leakage and preserve evaluation validity. The RAG knowledge base comprised exclusively external clinical guidelines, expert consensus documents, and peer-reviewed literature, with no patient-level data or reference rankings. LLM outputs were stored as generated, without post hoc modification. All study data were held on secure institutional servers with access controls, routine backups, and institutional data governance oversight.

Results

Study Population and Generated Simulation Dataset

A total of 20 patients completed the structured patient-priority ranking questionnaire across seven hypertension-management domains, generating 140 patient-reported domain-level rankings as the reference standard. Baseline characteristics of the 20 patients are presented in Table 1. A total of 8 LLMs were evaluated under 3 simulation conditions: unconstrained generation, patient-specific baseline simulation, and patient-specific RAG-enhanced simulation. For the patient-specific baseline and RAG-enhanced conditions, the final model-generated ranking within each domain was the complete 4-item ranking occurring most frequently across the 10 repeated questionnaire responses (mode-based aggregation). Across the patient-specific baseline and RAG-enhanced conditions, the selected complete ranking occurred in at least 6 of the 10 runs in every model-participant-condition-domain aggregation. Thus, 100% of the selected rankings met the strict-majority threshold (>50%). No ranking categories were tied for the highest frequency, and no tie-breaking procedure was required.

Table 1. Baseline characteristics of patients (N=20).

Characteristics	Values
Male, n (%)	13 (65)
Age (years), mean (SD)	58.5 (16.24)
Education, n (%)	
Middle school and below	8 (40)
High school	5 (25)
College diploma or above	7 (35)
Occupation, n (%)	
Student	2 (10)
Work	7 (35)
Retirement	11 (55)
Marital status, n (%)	
Unmarried	2 (10)
Married	14 (70)
Widowed	4 (20)
Urban, n (%)	16 (80)
Living alone, n (%)	2 (10)
Family income, n (%)	
<5000 Yuan, n (%)	3 (15)
5000-10000 Yuan, n (%)	7 (35)
>10,000 Yuan, n (%)	10 (50)
Diagnosed with hypertension for ≥3 years, n (%)	13 (65)
Blood pressure classification, n (%)	
Prehypertension	3 (15)
Stage 1 hypertension	6 (30)
Stage 2 hypertension	11 (55)
High cardiovascular risk, n (%)	15 (75)
Blood pressure control, n (%)	16 (80)
Number of antihypertensive medications taken, n (%)	
0	5 (25)
1	11 (55)
≥2	4 (20)
Combination antihypertensive drugs, n (%)	6 (30)
Blood pressure monitored at least once per week, n (%)	5 (25)
Smoking, n (%)	5 (25)
Drinking, n (%)	10 (50)
Moderate to high exercise intensity, n (%)	6 (30)
Diabetes, n (%)	4 (20)
Hyperlipidemia, n (%)	12 (60)
Family history of hypertension, n (%)	8 (40)
Family history of cardiovascular disease, n (%)	9 (45)
Using a blood pressure management app, n (%)	5 (25)
Experiences with large language models, n (%)	
I don’t know what a large language model is	9 (45)
I’ve heard of it, but I’ve never used it.	5 (25)
I’ve used it	6 (30)

Baseline Characteristics of LLM-Generated Virtual Hypertension Patients

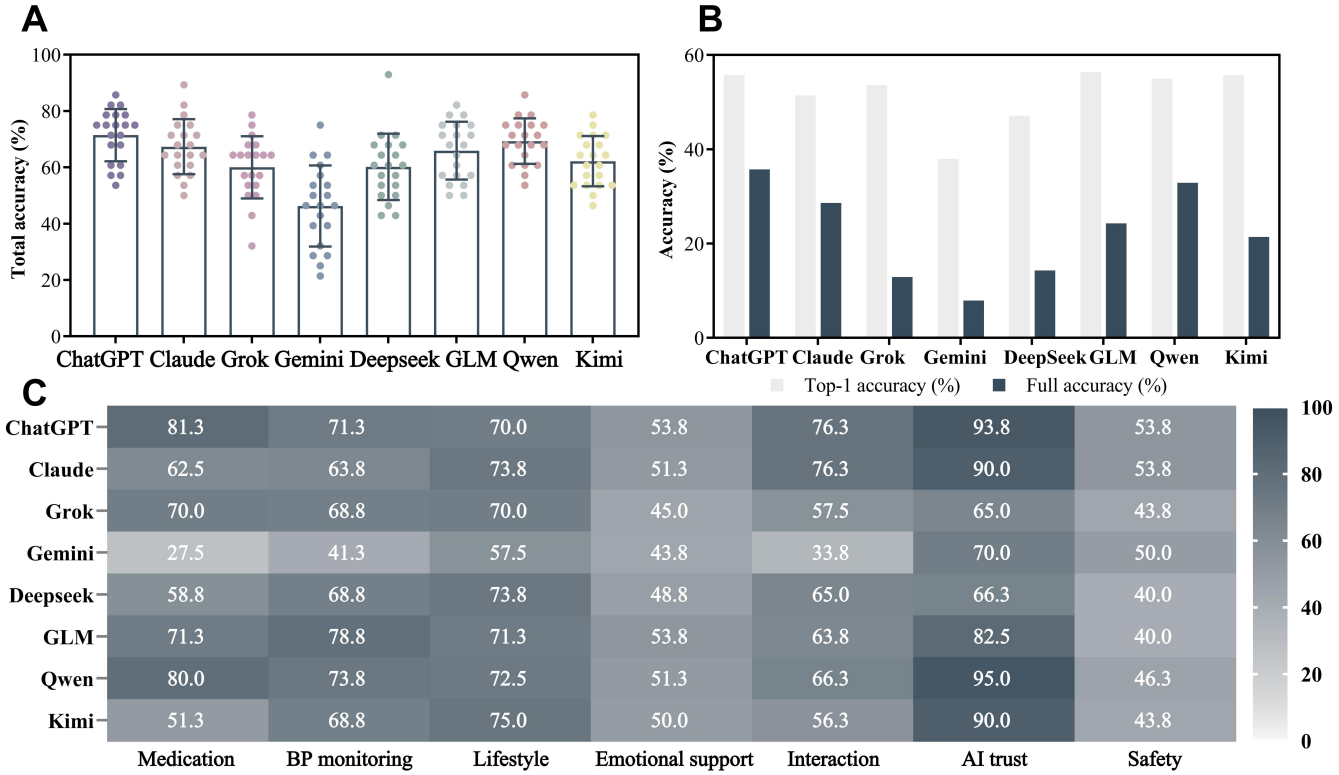
Among the 20 aggregated unconstrained virtual patient profiles generated by each evaluated LLM, the proportion of male profiles ranged from 45% to 55%, and the proportion aged 60 years or older ranged from 30% to 45% (Figure 2 and Table S9 in Multimedia Appendix 1). These descriptive distributions characterize the generated profiles under the study-specific prompting and category-based aggregation procedure; they do not establish agreement with population-level epidemiological distributions.

The aggregated virtual profiles differed from the enrolled reference cohort across several clinical and socioeconomic characteristics (Figure 2 and Table S9 in Multimedia Appendix 1). Across the 8 models, 60% to 85% of profiles had BMI 24 kg/m² or greater, 75% to 95% used antihypertensive medication, and 45% to 55% had college-level education. Physical activity and other characteristics also varied across models. These findings describe differences between the generated profiles and the 20 enrolled participants under the study conditions. They should not be interpreted as estimates of overrepresentation or underrepresentation relative to the wider hypertension population.

Patient-Specific Simulation Performance of Base LLMs

Across patient-specific simulations, LLMs demonstrated variable performance in reproducing real patient questionnaire responses across predefined domains (Figure 4).

Figure 4. Patient-specific simulation performance of large language models in reproducing real patient questionnaire responses across domains and accuracy metrics. BP: blood pressure.



Overall alignment scores varied substantially across the evaluated configurations, with the ChatGPT configuration achieving the highest total accuracy (71.4%), followed by Qwen (69.3%) and Claude (67.3%). The evaluated Gemini configuration showed the lowest total accuracy (46.3%), with wider variability indicated by the largest standard deviation (14.4%) across patient profiles (Figure 4A).

Top-1 accuracy-based evaluation revealed further differences in model performance (Figure 4B). Top-1 accuracy (Table S11 in Multimedia Appendix 1) ranged from 37.9% (Gemini) to 56.4% (GLM), while Full Accuracy (Table S12 in Multimedia Appendix 1) was generally lower across all models, ranging from 7.9% (Gemini) to 35.7% (ChatGPT). The substantial gap between top-1 accuracy and full accuracy across all models indicates that base LLMs were more capable of identifying a patient's highest-priority support function than reproducing the complete preference hierarchy.

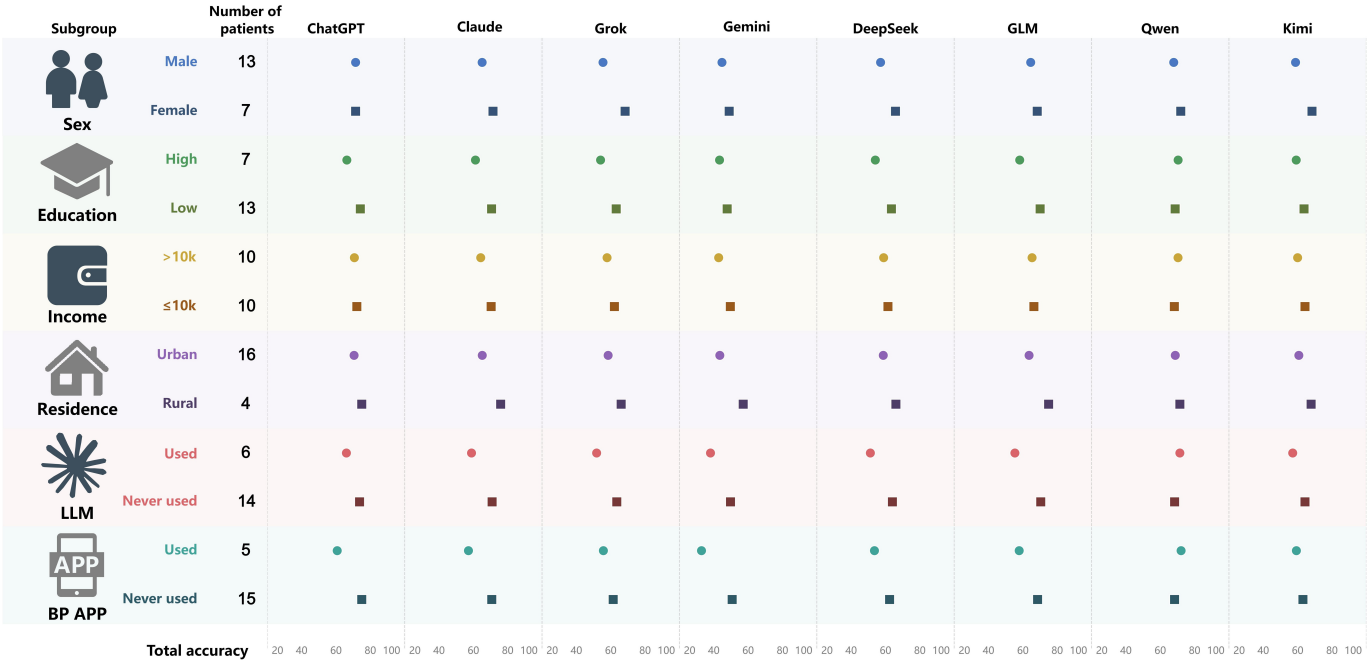
Domain-level performance analysis demonstrated heterogeneous accuracy across questionnaire domains (Figure 4C). Higher accuracy was observed in medication adherence, blood pressure monitoring and interpretation, and AI trust and acceptability, whereas consistently lower performance was observed in emotional support and reassurance and safety escalation awareness. Across the 8 models, mean baseline total accuracy was 62.8% (range: 46.3%-71.4%), mean top-1 accuracy was 51.6%, and mean full accuracy was 22.2%. Overall, patient-specific simulation performance under baseline conditions varied considerably across both models and domains, with higher concordance

with patient-reported priority rankings in structured clinical domains and lower concordance in psychosocial and safety-related domains.

Exploratory Subgroup Analysis of Baseline Simulation Performance

Exploratory subgroup analyses assessed baseline total accuracy across sex, education, income, residence, prior LLM use, and blood-pressure management application use (Figure 5). Fourteen of the 48 within-model comparisons showed nominal significance (P<.05): sex for Grok and Kimi; education for Claude and GLM; residence for Claude, Grok, and GLM; prior LLM use for Claude, Grok, DeepSeek, and GLM; and application use for ChatGPT, Claude, and Gemini. However, none remained significant after Holm adjustment. Similarly, no subgroup by model interaction remained significant after Greenhouse-Geisser correction and Holm adjustment (sex, adjusted P>.99; education, adjusted P>.99; income, adjusted P>.99; residence, adjusted P>.99; prior LLM use, adjusted P=.54; application use, adjusted P=.20). The exact permutation sensitivity analysis likewise identified no significant subgroup-by-model interactions after Holm adjustment. Given the small subgroup sizes, these exploratory findings should be interpreted cautiously; nonsignificant results do not establish equivalent performance across patient groups.

Figure 5. Baseline Total accuracy across patient subgroups and 8 large language models. Filled circles and squares represent the first and second subgroup listed within each dimension, respectively. Note: Higher education denotes a college diploma or above; income is monthly household income in ren min bi. BP: blood pressure; LLM: large language model.



RAG-Enhanced Simulation Improves Alignment With Real Patient-Reported Priorities

To evaluate whether external knowledge augmentation could improve the alignment with patient-reported priority rankings, all 8 LLMs were reevaluated after integration with the RAG framework under identical experimental conditions (Figure 6A).

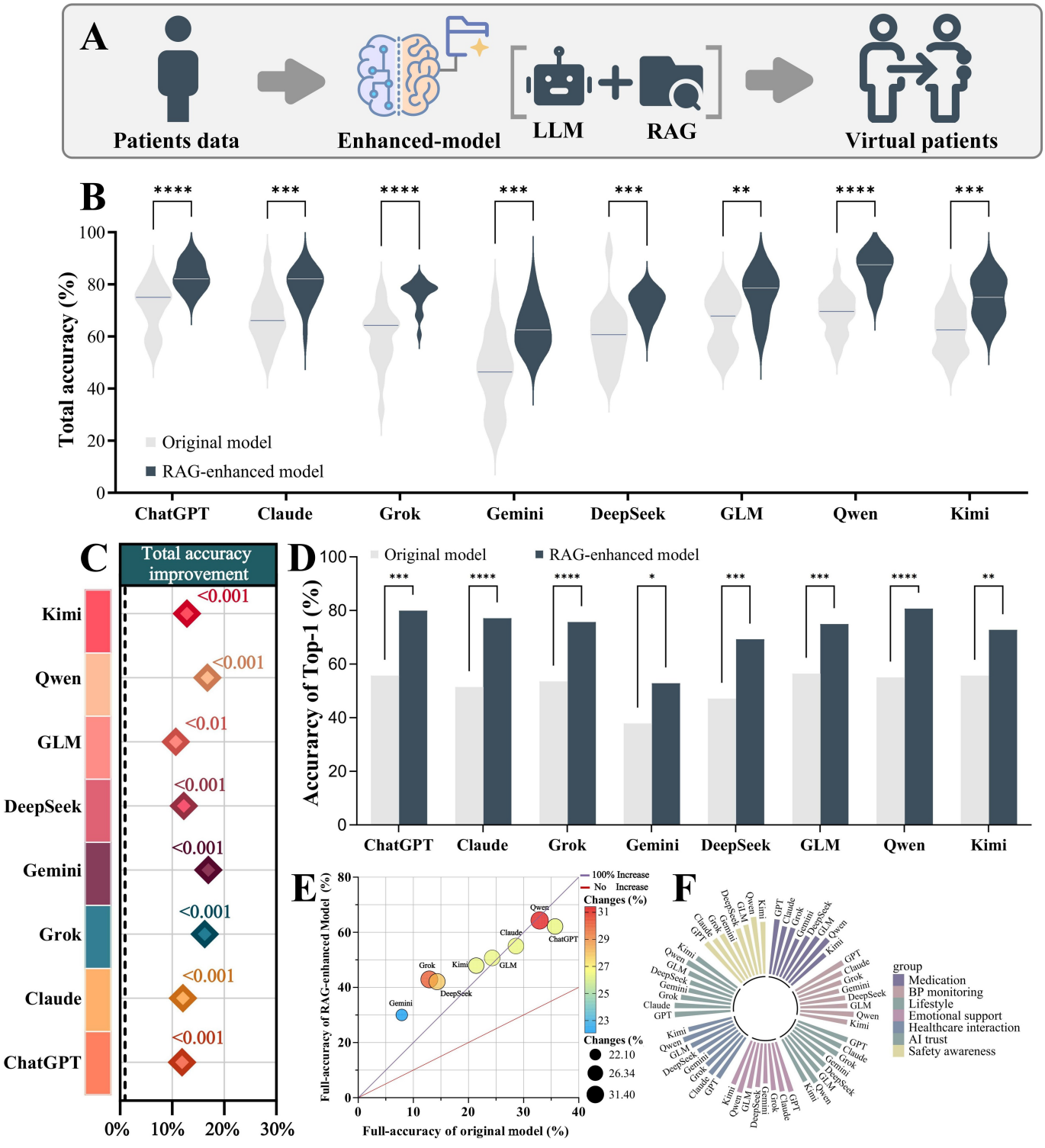
Across the 8 evaluated LLMs, RAG enhancement consistently improved alignment with real patient-reported rankings across all 3 evaluation metrics (Figures 6B-6E). Mean total accuracy increased from 62.8% under baseline simulation to 76.6% after RAG enhancement (P<.001). Top-1 accuracy increased from 51.6% to 72.9% (P<.001), and full accuracy increased from 22.2% to 49.4% (P<.001). The largest relative improvement was observed for full accuracy, indicating that retrieval augmentation improved not only identification of the highest-priority support function, but also reconstruction of the complete patient-reported priority hierarchy.

Model-level analysis showed that RAG-related improvements were observed across all eight evaluated LLMs, although the magnitude of improvement varied by model (Figures 6B-6C). Among the evaluated configurations, Qwen achieved the highest total accuracy after RAG enhancement (86.1%), followed by ChatGPT (83.4%), Claude (79.5%), GLM (76.6%), Grok (76.3%), Kimi (75.0%), DeepSeek (72.5%), and Gemini (63.2%). The largest absolute improvements in total accuracy were observed for Gemini, increasing from 46.3% to 63.2%; Qwen, from 69.3% to 86.1% (16.8 percentage points); and Grok, from 60.0% to 76.3%, indicating that retrieval augmentation improved performance across models with both higher and lower baseline accuracy.

Accuracy-based metrics further confirmed the improvement in ranking concordance after RAG enhancement (Figures 6D-6E). Among the evaluated configurations, Qwen showed the highest post-RAG top-1 accuracy (80.7%),

followed by ChatGPT (80.0%) and Claude (77.1%). For full accuracy, Qwen likewise showed the highest value among the evaluated configurations (64.3%), followed by ChatGPT (62.1%) and Claude (55.0%). The larger gain observed in full accuracy relative to top-1 accuracy indicates that RAG enhancement improved the evaluated configurations' ability to reproduce not only the top-ranked patient priority, but also the full hierarchy of patient preferences.

Figure 6. Retrieval-augmented generation improves alignment between large language model (LLM)-generated virtual patients and real patient-reported priorities. (A) Overview of the retrieval-augmented generation (RAG)-enhanced simulation workflow. (B) Participant-level total accuracy before and after RAG enhancement. (C) Mean improvement in total accuracy, expressed in percentage points. (D) Top-1 accuracy before and after RAG enhancement. (E) Baseline versus RAG-enhanced full accuracy; bubble size and color indicate improvement in percentage points. (F) Post-RAG domain-level total accuracy across 8 LLMs. *P*<.05 (*), *P*<.01 (**), *P*<.001 (***), and *P*<.0001 (****). LLM: large language model; RAG: retrieval-augmented generation.



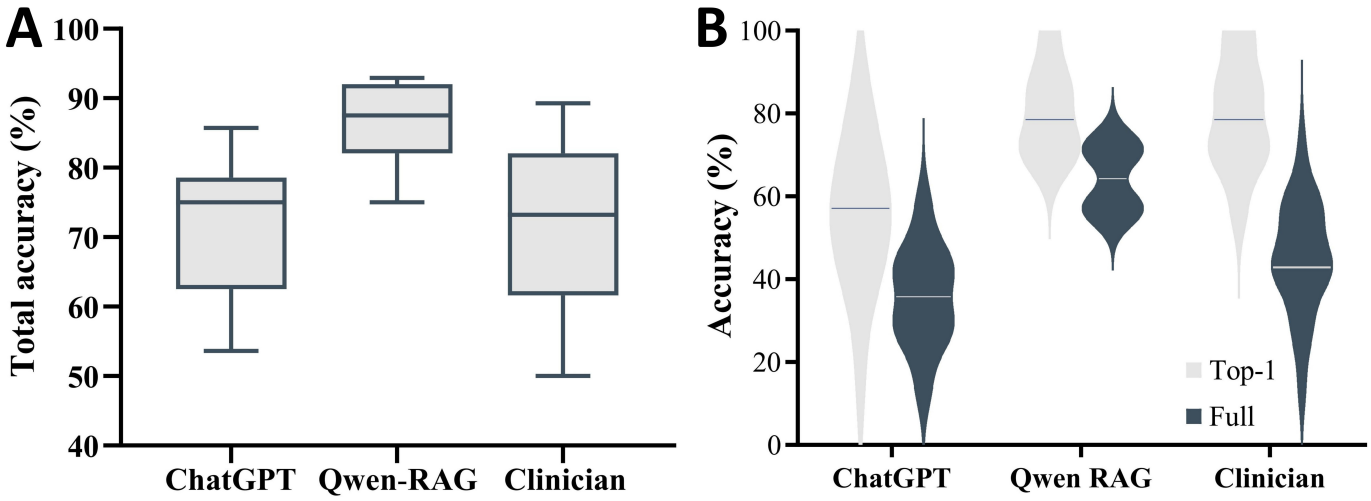
Domain-Level Heterogeneity in RAG-Related Gains

Post-RAG domain-level total accuracy across the 8 evaluated LLMs is presented in Figure 6F and Table S13 in Multimedia Appendix 1. Mean domain-level gains were calculated by averaging model-specific changes in total accuracy from baseline to RAG enhancement. The magnitude of improvement varied across questionnaire domains. The largest mean gains were observed for safety escalation awareness (+32.7 percentage points) and emotional support and reassurance (+24.4 percentage points). More moderate improvements were observed for medication adherence (+15.9 percentage points), health care system interaction (+11.3 percentage points), blood pressure monitoring and interpretation (+8.4 percentage points), and lifestyle modification (+5.6 percentage points). In contrast, AI trust and acceptability showed a mean change of –2.0 percentage points across the 8 models. Thus, RAG enhancement improved average ranking agreement in 6 of the 7 domains, with the magnitude of benefit varying by domain.

Comparison With Clinician-Inferred Patient-Perspective Rankings

To compare model-generated and clinician-inferred rankings within the same patient-priority inference task, we evaluated the best-performing base model (ChatGPT), the best-performing RAG-enhanced model (Qwen-RAG), and the clinician consensus against the corresponding patient-reported rankings (Figure 7). All 3 comparators were assessed using total accuracy, top-1 accuracy, and full accuracy (Table S14 in Multimedia Appendix 1). The clinicians received standardized patient information packages and were blinded to the patients' reported rankings and the LLM-generated rankings when completing their independent assessments.

Figure 7. Agreement of large language model (LLM)-generated and clinician-inferred rankings with patient-reported priorities. The best-performing base model (ChatGPT), the best-performing retrieval-augmented generation–enhanced model (Qwen-RAG), and the clinician consensus were evaluated against the same patient-reported rankings using total accuracy, top-1 accuracy, and full accuracy. RAG: retrieval-augmented generation.



For total accuracy, Qwen-RAG achieved 86.1%, compared with 72.5% for the clinician consensus and 71.4% for the ChatGPT baseline. The corresponding top-1 accuracy values were 80.7%, 80.0%, and 55.7%. Full accuracy was 64.3% for Qwen-RAG, 45.0% for the clinician consensus, and 35.7% for the ChatGPT baseline. Thus, Qwen-RAG had numerically higher total accuracy and full accuracy than the clinician consensus, while their top-1 accuracy values differed by 0.7 percentage points. Full accuracy was lower than top-1 accuracy for each of the 3 comparators.

Discussion

Principal Findings

In this comparative study, under baseline conditions, LLM-generated virtual hypertension patients showed moderate ability to approximate real patient priorities in an AI-assisted hypertension management task; this ranking concordance was substantially improved by RAG enhancement, with the best-performing model achieving a total accuracy of 86.1% and a full accuracy of 64.3%. In the unconstrained profile-generation phase, the aggregated virtual profiles showed demographic, clinical, and behavioral differences from the enrolled reference cohort, including a higher proportion with college-level education. These descriptive differences are conditional on the study sample, prompts, and category-based aggregation procedure and do not establish population-level representational bias. After RAG enhancement, mean total accuracy increased from 62.8% to 76.6%, top-1 accuracy from 51.6% to 72.9%, and full accuracy from 22.2% to 49.4%. The best-performing RAG-enhanced model achieved a total accuracy of 86.1% and a full accuracy of 64.3%, showing higher total accuracy than the clinician-inferred consensus benchmark under these standardized experimental conditions while achieving comparable top-1 accuracy. These findings suggest that LLM-generated virtual patients can provide useful but imperfect approximations of real patient perspectives, and that their use in digital medicine requires explicit grounding, empirical benchmarking, and transparent reporting.

Previous studies have evaluated LLM-based virtual patients for clinical decision-making training with structured feedback [31], clinical reasoning through conversational and robotic simulations [32], and communication skills training [33]. Related work has explored digital twins for modeling patient health trajectories [14,17]. Our study complements these approaches by evaluating whether patient-specific simulations reproduce priorities independently reported by real patients, rather than assessing educational performance or predicting clinical trajectories. The central contribution of this study is not the generation of virtual patients itself, but the empirical validation of LLM-generated patient-perspective simulations against real patient data. Whereas prior evaluations of LLMs in medicine have largely emphasized diagnostic accuracy, guideline concordance, and physician-facing decision support, comparatively few have asked whether models can represent how patients themselves prioritize care [34,35]. Our study reframes the evaluation target from clinical correctness to agreement with patient-reported priority rankings. As digital health technologies are increasingly incorporated into chronic disease management, systematically integrating patient perspectives into technology development and evaluation is essential for improving clinical relevance, usability, trust, and ethical acceptability [36,37]. This is particularly important in hypertension, where long-term disease control depends heavily on patient understanding, self-monitoring behaviors, medication adherence, lifestyle modification, emotional responses, and sustained engagement with healthcare systems [3,38]. Although interviews, questionnaires, focus groups, patient-reported outcome measures, and co-design frameworks remain the methodological standard for capturing patient perspectives, these approaches are often resource-intensive and difficult to scale across diverse populations and iterative stages of technology development. Our findings suggest that empirically benchmarked virtual patients can serve as scalable adjunctive tools for early-stage patient-perspective assessment, particularly when direct patient engagement is not yet feasible or when rapid formative evaluation is needed.

RAG may improve alignment by providing task-relevant clinical and patient-perspective information. One possible mechanism is that retrieved context helps models interpret the supplied patient characteristics within the specific hypertension-management task, reducing reliance on generalized assumptions about patient needs. This explanation remains hypothetical, as the study did not isolate the mechanisms underlying the observed improvement. Importantly, improved ranking agreement should be distinguished from reduced representational bias: this study did not directly test whether RAG reduced the demographic or behavioral discrepancies between the unconstrained profiles and the enrolled cohort. The latter remains a separate concern, as LLM-generated patient profiles without explicit population grounding may fail to reflect the diversity of the intended patient population [39].

In this study, the aggregated unconstrained profiles differed from the enrolled reference cohort in education, medication use, and lifestyle characteristics. Because the cohort was small and not population-representative, and the generated profiles underwent category-based mode aggregation, these comparisons cannot establish the extent of representational bias in the wider hypertension population. Zack et al [40] reported that GPT-4 generated clinical vignettes with stereotyped racial and gender distributions. The risk of representational bias remains relevant to patient-centered digital medicine, because technologies optimized against idealized virtual patients may fail to address the needs of patients who face the greatest barriers to long-term disease management. By grounding responses in task-specific clinical and patient-facing information, RAG improved alignment with real patient-reported priority rankings [41-44]. However, RAG should not be viewed as a complete solution to representational bias, because its performance remains dependent on the quality, scope, currency, and population representativeness of the retrieval corpus.

The comparison with clinician-inferred rankings provides an additional perspective on the challenge of representing patient priorities. Clinicians bring essential disease-specific expertise and practical knowledge of hypertension care, but in this study they were asked to infer patients' preferences, not to make diagnoses or treatment decisions. They may not always accurately infer subjective patient concerns, particularly those related to emotional burden, usability, trust, privacy, communication preferences, and the perceived relevance of digital tools. The finding that the best-performing RAG-enhanced model achieved higher total accuracy than the clinician consensus while showing comparable top-1 accuracy does not imply that LLMs have greater clinical expertise. Rather, grounded virtual patient simulation may provide a complementary source of information when considered alongside real patient responses and expert judgment. Divergence between clinician-inferred and patient-reported rankings may reflect differences between clinically informed assumptions and patients' perceived needs. Identifying such potential differences early could be useful in digital health design, when researchers assess whether proposed functions align with patient priorities before large-scale implementation. Virtual patient simulations should therefore complement, not replace, clinical expertise or direct patient engagement.

The most appropriate near-term role of LLM-generated virtual patients is therefore low-risk formative evaluation rather than clinical decision-making or replacement of patient engagement. In early development, virtual patients may help screen whether proposed functions align with patient priorities. During prototype refinement, they may support rapid iteration of patient-facing content, communication strategies, and interaction design. Before deployment, they may help identify representational gaps that require targeted recruitment or additional qualitative inquiry. In this sense, virtual patients can become part of a broader patient-centered digital medicine workflow that also includes co-design, real-world data, patient-reported outcomes, usability testing, and dynamic feedback after implementation. Their value lies in accelerating and structuring early-stage evaluation, not in substituting for real-world validation.

These considerations point to a broader vision in which patient perspectives are embedded systematically across the digital health development lifecycle rather than consulted only at isolated stages. Such integration can be operationalized through scalable and transparent mechanisms: structured co-design that incorporates patient representatives and advisory input early in prototype development and makes design objectives and evaluation criteria explicit; real-world data and patient-reported outcomes that extend clinical validation beyond conventional performance metrics to include usability, satisfaction, and psychological burden; and continuous post-deployment feedback loops that support iterative refinement while safeguarding patients' rights to informed participation and transparency regarding data use. Within this lifecycle, empirically benchmarked LLM-generated virtual patients occupy a specific and bounded role—not as a replacement for any of these components, but as a scalable adjunct for early, low-risk formative assessment that can help prioritize which patient-centered questions warrant fuller investigation through co-design, real-world evidence, and prospective evaluation. Realizing this vision will require transparent reporting of underlying assumptions and limitations at each stage, so that scalability does not come at the cost of representativeness, equity, or accountability.

Importantly, the virtual hypertension patients evaluated in this study should be clearly distinguished from digital twins, synthetic cohorts, and autonomous AI patients. They were not continuously linked to longitudinal individual-level physiological, behavioral, or treatment-response data, nor were they designed to predict patient-specific clinical trajectories, simulate closed-loop disease evolution, or act autonomously in complex care environments. Instead, they represent task-specific, text-based patient-perspective simulations designed to approximate how patients prioritize functions in AI-assisted hypertension management. This distinction is essential because overstating LLM-generated virtual patients as digital twins or autonomous patient agents could lead to inappropriate clinical interpretation and premature deployment.

Limitations

First, the reference cohort was small and recruited through non-probability convenience sampling with purposive selection among eligible volunteers. Although we sought diversity across geographic, demographic, socioeconomic, clinical, and digital-health characteristics, the cohort was not designed to be statistically representative of the broader population of patients with hypertension. Recruitment through online communities and outpatient follow-up networks may also have preferentially included participants who were more digitally engaged or more willing to participate in research. Therefore, the findings should be interpreted as evidence of performance within this structured reference cohort rather than as population-level estimates of patient preferences. Validation in larger, more diverse, and preferably multicenter cohorts is needed to assess generalizability. Second, several subgroup strata were also small, and the study was not powered to detect subgroup interactions. Non-significant interaction tests therefore cannot establish comparable performance or fairness across patient groups. Third, the evaluation measured agreement with patient-reported rankings of predefined support functions. These functions may not capture all patient concerns, and the forced-ranking format did not allow ties. Total accuracy, top-1 accuracy, and full accuracy thus assess preference reconstruction within this task, not the full range of patient experiences. Priorities may also change over time. We did not compare patient-specific simulations with simulations lacking individual baseline information or with a fixed generic ranking. The observed concordance therefore does not isolate the incremental contribution of patient-specific information and may partly reflect priorities shared across participants. Future studies should include such comparators to assess the added value of individual patient information, alongside open-ended responses, qualitative interviews, patient-reported outcome measures, and longitudinal assessment. Fourth, RAG performance depends on the quality and representativeness of the retrieval corpus. This study could not isolate the respective effects of retrieved content, prompting, model architecture, and generation settings. The model comparison involved specific API endpoints, including configurations labeled flash or lite. We did not test alternative flagship endpoints or assess whether their use would change the relative results. Consequently, the observed performance differences, including the lower baseline accuracy of the tested Gemini configuration, should be attributed to the evaluated endpoints under these experimental conditions and should not be interpreted as definitive rankings of model families or providers. Fifth, agreement among the 5 clinicians' independent rankings before consensus formation was not quantified. The final consensus rankings therefore do not indicate the extent of initial agreement or disagreement among the clinicians, limiting assessment of the internal reliability of the human comparator. The clinician consensus should consequently be interpreted as a comparator generated by this specific panel and consensus procedure, rather than as a definitive benchmark of clinician performance. This comparison did not assess diagnosis or treatment decisions and should not be interpreted as evidence that LLMs outperform clinicians in clinical practice. Finally, other aspects of virtual-patient validity, including response consistency, calibration, and robustness to model updates, require further evaluation.

Conclusion

This study suggests that LLM-generated virtual hypertension patients can moderately approximate real patient perspectives and that RAG substantially improves concordance with patient-reported priority rankings. More broadly, the study provides an empirical framework for evaluating virtual patient simulations as scalable adjunctive tools in cardiovascular digital medicine. These systems should not be regarded as substitutes for real patient engagement, co-design, patient-reported outcome research, or longitudinal real-world evaluation. Future studies should validate this approach in larger and more diverse cohorts, examine its performance across different disease areas and digital health tasks, and develop reporting standards for transparent, reproducible, and

ethically responsible virtual patient simulation in medical AI research.

Generative AI Use

During the preparation and revision of this manuscript, the authors used ChatGPT (OpenAI) to assist with English-language editing and revision of manuscript text and to provide preliminary conceptual ideas for Figure 7. All AI-assisted text was reviewed, verified, and edited by the authors. The final version of Figure 7 was created by the authors, and no AI-generated image was directly incorporated into the manuscript. The authors take full responsibility for the final manuscript and its visual content.

Funding

This study was supported by the Clinical and Translational Research Project of Anhui Province (202427b10020086, 202427b10020089, 202427b10020097); the Research Funds of the Joint Research Center for Regional Diseases of IHM (2024bydjk001, 2024bydjk002, 2024bydjk005); the Scientific Research Project of Anhui Provincial Health Commission (AHWJ2024Aa10053); the Science Research Project of Bengbu Medical University (2024byfy008); the National Engineering Research Center of Science and Technology Information (2025STI135); and the Excellent Young Scholars Program of the First Affiliated Hospital of Bengbu Medical University (2025byyfyyq09).

Data Availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Authors' Contributions

YJW, HYL, DT, and SYC conceived and performed the study. PZ, WQF, ZYL, BW, LT, and BZ contributed to methodological optimization, data collection, data processing, and model evaluation. YJW, HYL, DT, WYY, and SYC performed data analysis and manuscript preparation. SPD, CP, WPT, and JW served as co-corresponding authors, were responsible for correspondence and communication with the journal, supervised the study, and critically revised the manuscript for important intellectual content. Their email addresses are as follows: SPD (whuduansp@163.com), CP (2021203020011@whu.edu.cn), WPT (tanwp7@mail2.sysu.edu.cn), and JW (junwang0607@163.com). All authors approved the final manuscript and consented to its submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional tables and figures.
[PDF File (Adobe File), 295 KB-Multimedia Appendix 1]

Checklist 1

Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–large language model checklist.
[PDF File (Adobe File), 155 KB-Checklist 1]

References

1. Writing Committee Members*, Jones DW, Ferdinand KC, et al. 2025 AHA/ACC/AANP/AAPA/ABC/ACCP/ACPM/AGS/AMA/ASPC/NMA/PCNA/SGIM Guideline for the prevention, detection, evaluation and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Hypertension. Oct 2025;82(10):e212-e316. [doi: 10.1161/HYP.0000000000000249] [Medline: 40811516]
2. Wang J, Xu Y, Zhu J, et al. Multimodal data-driven, vertical visualization prediction model for early prediction of atherosclerotic cardiovascular disease in patients with new-onset hypertension. J Hypertens. Oct 1, 2024;42(10):1757-1768. [doi: 10.1097/HJH.0000000000003798] [Medline: 38937962]
3. Wang Y, Guo F, Wang J, et al. Efficacy of a WeChat-based multimodal digital transformation management model in new-onset mild to moderate hypertension: randomized clinical trial. J Med Internet Res. Dec 4, 2023;25:e52464. [doi: 10.2196/52464]
4. Xu Q, Xu Y, Wang Y, et al. Multiple machine-learning-driven metabolic frameworks for long-term prognostic risk assessment in patients with coexisting hypertension and obstructive sleep apnea: insights from a multicenter cohort study. Front Physiol. 2026;17:1739374. [doi: 10.3389/fphys.2026.1739374] [Medline: 41815786]
5. Chinese Society of Cardiology, Chinese Medical Association; Hypertension Committee of Cross-Straits Medicine Exchange Association; Cardiovascular Disease Prevention and Rehabilitation Committee, Chinese Association of

Rehabilitation Medicine. Clinical practice guideline for the management of hypertension in China. Chin Med J. 2024;137(24):2907-2952. [doi: 10.1097/CM9.0000000000003431]
6. Wang Y, Zhu T, Zhou T, et al. Hyper-DREAM, a multimodal digital transformation hypertension management platform integrating large language model and digital phenotyping: multicenter development and initial validation study. J Med Syst. Apr 2, 2025;49(1):42. [doi: 10.1007/s10916-025-02176-1] [Medline: 40172683]
7. Xu Q, Li Y, Zhu M, et al. Precision cardiovascular medicine with big data and AI. NPJ Digit Med. Mar 17, 2026;9(1):339. [doi: 10.1038/s41746-026-02538-0] [Medline: 41844792]
8. Trivedi R, Shaw T, Sheahen B, Chow CK, Laranjo L. Patient perspectives on conversational artificial intelligence for atrial fibrillation self-management: qualitative analysis. J Med Internet Res. Mar 12, 2025;27:e64325. [doi: 10.2196/64325] [Medline: 40073398]
9. Aiyegbusi OL, McMullan C, Hughes SE, et al. Considerations for patient and public involvement and engagement in health research. Nat Med. Aug 2023;29(8):1922-1929. [doi: 10.1038/s41591-023-02445-x] [Medline: 37474660]
10. Aiyegbusi OL, Cruz Rivera S, Roydhouse J, et al. Recommendations to address respondent burden associated with patient-reported outcome assessment. Nat Med. Mar 2024;30(3):650-659. [doi: 10.1038/s41591-024-02827-9] [Medline: 38424214]
11. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med. 2020;3(1):110. [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
12. Marra C, Chico T, Alexandrow A, et al. Addressing the challenges of integrating digital health technologies to measure patient-centred outcomes in clinical registries. Lancet Digit Health. Mar 2025;7(3):e225-e231. [doi: 10.1016/S2589-7500(24)00223-1] [Medline: 39672737]
13. Angus DC, Khera R, Lieu T, et al. AI, health, and health care today and tomorrow: the JAMA summit report on artificial intelligence. JAMA. Nov 11, 2025;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]
14. Makarov N, Bordukova M, Quengdaeng P, et al. Large language models forecast patient health trajectories enabling digital twins. NPJ Digit Med. Oct 1, 2025;8(1):588. [doi: 10.1038/s41746-025-02004-3] [Medline: 41034564]
15. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback. J Med Internet Res. Apr 4, 2025;27:e68486. [doi: 10.2196/68486] [Medline: 39854611]
16. Zeng J, Qi W, Shen S, et al. Embracing the future of medical education with large language model-based virtual patients: scoping review. J Med Internet Res. Nov 13, 2025;27:e79091. [doi: 10.2196/79091] [Medline: 41232097]
17. Sadée C, Testa S, Barba T, et al. Medical digital twins: enabling precision medicine and medical artificial intelligence. Lancet Digit Health. Jul 2025;7(7):100864. [doi: 10.1016/j.landig.2025.02.004] [Medline: 40518342]
18. Zhang C, Wang FX, Tang X, et al. Virtual medicine: medical AI in human health and diseases. Mil Med Res. 2026;13(1):100012. [doi: 10.1016/j.mmr.2026.100012] [Medline: 42063466]
19. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
20. Pariente B, Varennes O, Burgun A, Azizi M, Amar L, Tsopra R. Empowering patients and clinicians: LLMs in hypertension care, a scoping review. Hypertension. Jul 2026;83(7):e27004. [doi: 10.1161/HYPERTENSIONAHA.126.27004] [Medline: 42021742]
21. Varzideh F, Mone P, Kansakar U, Pande S, Jankauskas SS, Santulli G. Artificial intelligence in cardiovascular medicine: focus on hypertension. Hypertension. Jun 2026;83(6):e26094. [doi: 10.1161/HYPERTENSIONAHA.126.26094] [Medline: 41884885]
22. Li Z, Liu H, Tan W, et al. The effects of multitype prompt engineering for large language models in hypertension treatment decisions. NPJ Digit Med. Apr 15, 2026;9(1):526. [doi: 10.1038/s41746-026-02645-y] [Medline: 41986562]
23. Wang Y, Tan W, Cheng S, et al. Large language model agent for managing patients with suspected hypertension. Hypertension. Jan 2026;83(1):212-224. [doi: 10.1161/HYPERTENSIONAHA.125.25305] [Medline: 41064862]
24. Wang Y, Luan Y, Cheng S, et al. A multi-layer retrieval-augmented large language model framework for enhancing hypertension education. Hypertens Res. Apr 2026;49(4):1428-1440. [doi: 10.1038/s41440-025-02481-9] [Medline: 41501362]
25. Li P, Patel A, Vallamchetla SK, et al. Optimizing retrieval-augmented generation (RAG) in clinical medicine: methods and performance evaluation. J Am Med Inform Assoc. Aug 1, 2026;33(8):1436-1445. [doi: 10.1093/jamia/ocag056] [Medline: 42243630]
26. Layton AT. AI, machine learning, and ChatGPT in hypertension. Hypertension. Apr 2024;81(4):709-716. [doi: 10.1161/HYPERTENSIONAHA.124.19468] [Medline: 38380541]

27. Wang Y, Peng C, Hu R, et al. A knowledge-enhanced domain-aware large language model agent for atrial fibrillation management. NPJ Digit Med. Jul 20, 2026. [doi: 10.1038/s41746-026-03038-x]
28. Luo MJ, Bi S, Pang J, et al. A large language model digital patient system enhances ophthalmology history taking skills. NPJ Digit Med. Aug 4, 2025;8(1):502. [doi: 10.1038/s41746-025-01841-6] [Medline: 40760042]
29. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond). Dec 19, 2025;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
30. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. Jan 2025;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
31. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ. Nov 28, 2024;24(1):1391. [doi: 10.1186/s12909-024-06399-7] [Medline: 39609823]
32. Borg A, Georg C, Jobs B, et al. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study. J Med Internet Res. Mar 3, 2025;27:e63312. [doi: 10.2196/63312] [Medline: 40053778]
33. Elhilali A, Ngo ASH, Reichenpfader D, Denecke K. Large language model-based patient simulation to foster communication skills in health care professionals: user-centered development and usability study. JMIR Med Educ. Dec 12, 2025;11:e81271. [doi: 10.2196/81271] [Medline: 41385781]
34. Wang X, Wang W, Li L, et al. Large language models for acute coronary syndrome triage at first medical contact in emergency departments. NPJ Digit Med. 2026;9(1):755. [doi: 10.1038/s41746-026-02904-y]
35. Brodeur PG, Buckley TA, Kanjee Z, et al. Performance of a large language model on the reasoning tasks of a physician. Science. Apr 30, 2026;392(6797):524-527. [doi: 10.1126/science.adz4433] [Medline: 42060751]
36. Kim M, Patrick K, Nebeker C, et al. The digital therapeutics real-world evidence framework: an approach for guiding evidence-based digital therapeutics design, development, testing, and monitoring. J Med Internet Res. Mar 5, 2024;26:e49208. [doi: 10.2196/49208] [Medline: 38441954]
37. Main A, McCartney H, Ibrar M, et al. Patients' experiences of digital health interventions for the self-management of chronic pain: systematic review and thematic synthesis. J Med Internet Res. Mar 18, 2025;27:e69100. [doi: 10.2196/69100] [Medline: 40101209]
38. Kario K, Nomura A, Harada N, et al. Efficacy of a digital therapeutics system in the management of essential hypertension: the HERB-DH1 pivotal trial. Eur Heart J. Oct 21, 2021;42(40):4111-4122. [doi: 10.1093/eurheartj/ehab559] [Medline: 34455443]
39. Veenhuizen M, O'Malley A. Demographic biases in AI-generated simulated patient cohorts: a comparative analysis against census benchmarks. Adv Simul (Lond). Nov 18, 2025;10(1):58. [doi: 10.1186/s41077-025-00385-9] [Medline: 41254805]
40. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. Jan 2024;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
41. Yang R, Ning Y, Keppo E, et al. Retrieval-augmented generation for generative artificial intelligence in health care. NPJ Health Syst. Jan 25, 2025;2(1):2. [doi: 10.1038/s44401-024-00004-1] [Medline: 42527447]
42. Youbi L, Redjdal A, Seroussi B. Exploring synergies between large language models and knowledge models in healthcare: a scoping review. Stud Health Technol Inform. Jun 29, 2026;338:703-707. [doi: 10.3233/SHTI260936] [Medline: 42394098]
43. Zhao W, Yang X, Lyu Z, Xu C, Guan Z. Road of large language model: source, challenge, and future perspectives. Research (Wash D C). 2025;8:0655. [doi: 10.34133/research.0655] [Medline: 40735465]
44. Collaco BG, Srinivasagam P, Gomez-Cabello CA, et al. Integrating fine-tuning and retrieval-augmented generation for healthcare AI systems: a scoping review. Bioengineering (Basel). Feb 14, 2026;13(2):225. [doi: 10.3390/bioengineering13020225] [Medline: 41749764]

Abbreviations

LLM: large language model
RAG: retrieval-augmented generation
TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–large language model

Edited by Matthew Balcarras; peer-reviewed by Hejia Wan, Masab Mansoor; submitted 24.Jul.2026; initial comments to author 28.Aug.2026; final revised version received 04.Oct.2026; accepted 22.Oct.2026; published 24.Nov.2026

Please cite as:
Wang Y, Liu H, Tang D, Cheng S, Zhao P, Fang W, Li Z, Wu B, Tang L, Yin W, Zhou B, Duan S, Peng C, Tan W, Wang J
Evaluating Large Language Model–Generated Virtual Hypertension Patients as Proxies for Real Patient Perspectives: Comparative Study
J Med Internet Res 2026;28:e107865
URL: https://www.jmir.org/2026/1/e107865
doi: 10.2196/107865

© Yijun Wang, Hengyang Liu, Du Tang, Siyi Cheng, Peng Zhao, Wenqiang Fang, Zeyan Li, Bing Wu, Long Tang, Wenyuan Yin, Bowen Zhou, Shoupeng Duan, Chen Peng, Wuping Tan, Jun Wang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 24.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.