Original Paper

# Mobile Medical Education Apps as Digital Health Information Resources in China and the United States: Longitudinal App Store Metadata Analysis and Topic Modeling of User Review Narratives

Junhao Ma[1,2*], MSc; Tian Shen[1,2*], MSc; Peng Xiang[3,4*], MD; Yingjun Li[1*], PhD; Yuanshuai Chen[3,5], MD; Zhengyang Lu[6], BM; Jingkai Zeng[7], BM; Wei Dai[8], BM; Hanwen Pan[9,10,11,12], BM; Liping Yang[13,14], MD; Yunfan He[15], MS; Xingxi Li[16], MS; Joyce Wu[17,18], BSc; Qi Yu[19,20], MD; Jun Liang[1,2,3,6], MS; Jianbo Lei[21,22], MD, PhD

[1]Department of Epidemiology and Health Statistics, School of Public Health, Hangzhou Medical College, Hangzhou, China

[2]National Key Laboratory of Transvascular Implantable Devices, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China

[3]Department of AI and IT, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China

[4]Intelligent Medical Research Center, Zhejiang University Institute of Computer Innovation Technology, Hangzhou, China

[5]Medical Education Research Center, School of Medicine, Zhejiang University, Hangzhou, China

[6]Department of Preventive Medicine, School of Public Health, Zhejiang University, Hangzhou, China

[7]Department of Clinical Medicine, Taizhou University, Taizhou, China

[8]School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou, China

[9]Foshan Fosun Chancheng Hospital, Guangdong Medical University, Foshan, China

[10]Guangdong Provincial Hospital of Chinese Medicine, Guangzhou, China

[11]Foshan Clinical Medical School, Guangzhou University of Chinese Medicine, Foshan, China

[12]Zhonghe Guoyi Beijing International Medical Research Institute, Beijing, China

[13]College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, Xinjiang, China

[14]Institute of Medical Engineering Interdisciplinary Research, Xinjiang Medical University, Urumqi, China

[15]School of International Relations and Public Affairs, Fudan University, Shanghai, China

[16]Department of Industrial Engineering, Tsinghua University, Beijing, China

[17]Academy for Allied Health Sciences, Scotch Plains, NJ, United States

[18]New Jersey Medical School, Rutgers University, Newark, NJ, United States

[19]Shanxi Key Laboratory of AI-Enhanced Clinical Decision Support, Shanxi Medical University, Taiyuan, China

[20]Xinzhou Vocational and Technical College, Xinzhou, China

[21]Center for Medical Informatics, Institute of Advanced Clinical Medicine, Peking University, Beijing, China

[22]Affiliated Hospital of Southwest Medical University, Luzhou, China

[*]these authors contributed equally

**Corresponding Author:**
Jianbo Lei, MD, PhD
Center for Medical Informatics, Institute of Advanced Clinical Medicine
Peking University
38 Xueyuan Rd, Haidian District
Beijing, 100191
China
Phone: 86 10 8280 5901
Email: jblei@hsc.pku.edu.cn

## Abstract

**Background:** Mobile learning apps are increasingly used in medical education, but cross-national evidence on their longitudinal development and rating-based satisfaction patterns remains limited.

**Objective:** This study aimed to examine the longitudinal evolution of medical education apps in China and the United States and identify app-related factors associated with positive and negative rating deviations among app store reviewers.



XSL•FO
RenderX

**Methods:** We conducted a longitudinal app store analysis of 467 medical education apps and 232,598 user reviews from major Chinese and US app stores. Joinpoint regression was used to characterize temporal trends in app development. Latent Dirichlet allocation (LDA) topic modeling, with large language model–assisted topic labeling, was used to identify themes in user reviews. Tobit regression, Wald tests, and Kano model classification were used to examine asymmetric associations with positive and negative rating deviations and to classify review topics into functional attributes.

**Results:** Among the 232,598 analyzed reviews, 219,105 were rated 4 stars or higher, corresponding to an overall rating-based satisfaction proxy rate of 94.20%. This rating-based satisfaction proxy rate was lower in China than in the United States (128,018/139,416, 91.82% vs 91,087/93,182, 97.75%). Within the 2025-observable analytic cohort, cumulative observable app counts increased significantly in both countries, with average annual percentage changes of 41.69% in China and 41.17% in the United States. The joinpoint analysis identified temporal inflection points within this observable cohort around 2017 and 2021-2022. User concerns also shifted over time. In China, the dominant review theme changed from “Clinical practicality” in 2009-2017 (2735/7385 reviews, 37.03%) to “Practice test function” in 2018-2025 (53,929/132,031 reviews, 40.85%). In the United States, user attention shifted from “Functional availability” in 2009-2017 (20,736/26,118 reviews, 79.39%) to “Cost” in 2018-2025 (27,401/67,064 reviews, 40.86%). Tobit regression and Kano classification showed that factors associated with positive and negative rating deviations differed by country and developmental stage, with essential and charm factors varying across temporal phases.

**Conclusions:** This longitudinal app store analysis revealed distinct but partially converging patterns of observable app development and rating-based satisfaction among app store reviewers in China and the United States. These findings suggest that future medical education apps should prioritize reliable content, technical stability, affordable exam preparation, and personalized learning support. They may also inform the design of future AI-supported medical education platforms while requiring further validation with learning process and outcome data.

***(JMIR Med Educ 2026;12:e107560)*** doi: 10.2196/107560

**KEYWORDS**

digital health informatics; medical education; mobile health apps; text mining; app store analysis; user-generated content; rating-based satisfaction; cross-national comparison

## Introduction

With the rapid advancement of mobile internet, AI, and digital technologies, coupled with the growing demands of professional education for continuous learning and self-renewal capabilities, digital learning methods are gradually evolving from traditional teaching aids into essential infrastructure supporting lifelong learning. Medical education, as a field highly dependent on specialized knowledge, clinical skills, and situational decision-making abilities, demands rapid knowledge updates and navigates complex practical scenarios, placing heightened demands on learning flexibility and accessibility [1]. However, the long-standing training model centered on fixed classroom instruction and offline clinical rotations remains constrained in terms of scheduling, learning continuity, and personalized support. This vulnerability becomes particularly pronounced during public health emergencies [2,3]. The COVID-19 pandemic has further accelerated the shift toward digital, self-directed, and contextualized medical education models [4,5]. Against this backdrop, the development of mobile learning and online educational resources has enabled learning activities to be more flexibly integrated into daily teaching and clinical practice [6]. Medical education apps support diverse educational activities including coursework, clinical guideline access [7,8], case reasoning training, and formative assessment [9-13], providing on-demand learning support for medical students, residents, and practicing physicians [14].

Despite the rapid global adoption of medical education apps, their practical implementation faces numerous challenges. Existing research indicates that learners frequently encounter issues such as outdated content, insufficient system stability, and distractions from multifunctional mobile devices when using medical education apps. These factors negatively impact learning outcomes, user experience, and willingness for sustained use [15-18]. Given the high diversity in developers, target users, and application scenarios of medical education apps, quality issues do not manifest uniformly but exhibit distinct characteristics across different markets and cultural contexts [19]. Differences in medical education systems, learning approaches, and cultural backgrounds across countries and regions also lead to significant variations in learners’ perceptions and demands regarding app functionality, content quality, and “user satisfaction” [20]. This implies that evaluations of medical education apps that overlook market and cultural context differences may fail to accurately reflect their true utility and areas for improvement.

These contextual differences are particularly salient in a China-US comparison. Both countries represent large digital app markets, but medical education apps are embedded in different institutional, educational, and sociocultural environments. Established technology adoption theories, including the technology acceptance model (TAM) and the unified theory of acceptance and use of technology, suggest that users’ acceptance and continued use of digital technologies are shaped by perceived usefulness, ease of use, social influence, and facilitating conditions [21,22]. Cross-cultural technology acceptance research further indicates that national and organizational cultures may moderate how users perceive and adopt digital technologies [23]. In China, medical education app use is situated within standardized medical training



XSL•FO
RenderX

pathways, national medical education reform, and examination-oriented learning pressures within the medical education system [24]. In the United States, medical education app use occurs within a credentialing environment shaped by the United States Medical Licensing Examination (USMLE) and residency selection processes [25]. These differences provide a rationale for comparing how app store reviewers in the 2 countries evaluate content reliability, technical stability, cost, exam preparation, and personalized learning support. Therefore, this study examines medical education apps not only as digital market products but also as sociotechnical educational tools embedded in distinct policy, institutional, and learning contexts.

Nevertheless, current research on user experience and satisfaction with medical education apps still primarily relies on methods such as questionnaire surveys [26,27], standardized scales, and on-site usability testing [28]. Common tools such as the Mobile App Rating Scale (MARS), developed by Stoyanov et al [29], and the Mobile App Rubric for Learning (MARuL), developed by Gladman et al [28], are mainly used to quantify an app's functionality, learning support effectiveness, usability, and overall satisfaction [30,31]. Such studies typically rely on participants' subjective evaluations, feature limited sample sizes, and often focus on single institutions or regions. Research perspectives also tend to be narrow, frequently emphasizing app effectiveness [32], usability [33], or overall user experience [34], while insufficient attention is given to cross-national differences and large-scale user feedback in real-world usage scenarios [35]. A growing body of research indicates that dynamic analysis based on longitudinal data is more effective than static, single-point-in-time assessments for revealing patterns in user experience and satisfaction over time and across app development stages [36,37]. Overall, existing studies remain constrained by limitations in sample size, data types, and temporal dimensions, making it challenging to systematically characterize the evolutionary process of medical education apps and the dynamic shifts in user satisfaction needs across different markets and cultural contexts.

Beyond these methodological limitations, a related conceptual gap is that medical education apps are often evaluated as functional toolkits rather than as sociotechnical interventions. Digital tools in professional education may reshape users' learning behaviors, daily routines, role expectations, and professional identity. In medical education, mobile technologies are not merely static tools for accessing learning materials; their design and use are shaped by educational theory, preclinical and clinical training contexts, and the needs of institutions, teachers, and learners [38]. Similar role-related effects have been observed in other professional service settings. For example, research on smart technologies in nursing homes showed that digital implementation changed caregivers' role expectations, work routines, performance norms, and adaptation strategies [39]. This perspective suggests that medical education apps should be examined not only as functional toolkits, but also as digital interventions embedded in professional training environments and institutional workflows.

Given these limitations, it is necessary to examine the evolution of medical education apps and the app-related factors associated with rating-based satisfaction patterns among app store reviewers from a systematic and dynamic perspective. As technology-driven educational products, medical education apps are also closely tied to institutional and cultural contexts in their usage and development. Therefore, understanding the evolution and use of medical education apps requires consideration of multiple dimensions, including changes in app quantity and functionality, as well as user needs and evaluation feedback. This study innovatively proposes a bidirectional "supply-demand" analytical framework. By integrating market evolution with user reviews, it explores functional adjustments, demand evolution, and cross-national differences in rating-based satisfaction patterns among app store reviewers in China and the United States, providing insights for app evaluation and design optimization. Based on the aforementioned research background and analytical framework, this study poses the following research questions (RQs):

- RQ1: What are the temporal evolution characteristics of market supply for medical education apps?
- RQ2: What app-related factors are associated with positive rating deviations (PD) and negative rating deviations (ND) among reviewers of medical education apps, and do these associations exhibit asymmetric patterns?
- RQ3: How can asymmetric patterns in PD and ND be used to classify review topics into Kano functional attributes across country-specific samples and temporal phases?

## Methods

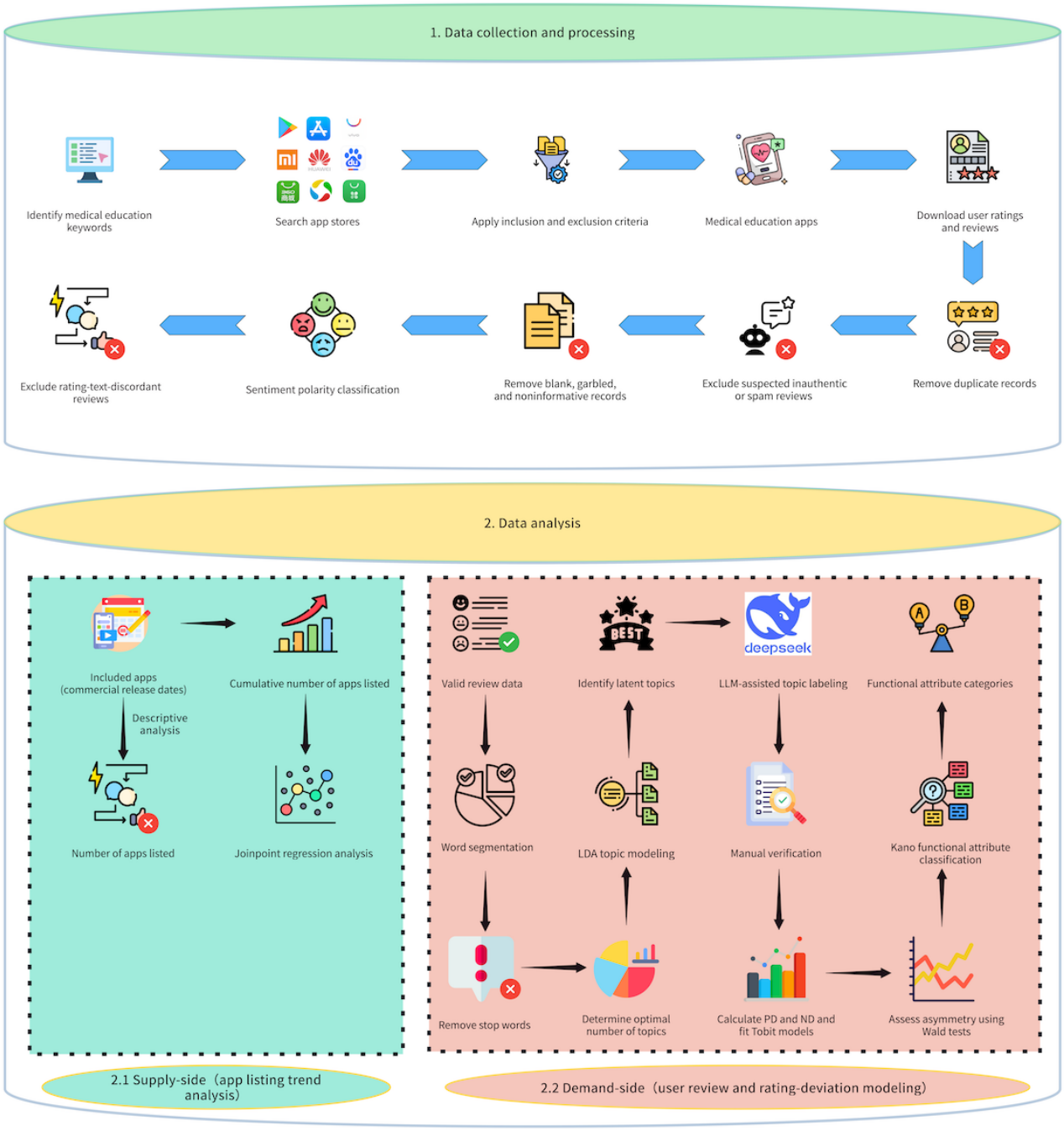
### Overall Study Design and Analytical Framework

This study uses a longitudinal, cross-national mixed methods research design to systematically characterize the observable developmental features of medical education apps and rating-based satisfaction patterns among app store reviewers. The overall research approach introduces a bidirectional "supply-demand" analytical framework (Figure 1), integrating structural changes at the app market level with user experience feedback from real-world usage scenarios. This enables simultaneous capture of both the developmental trajectory of medical education apps and the dynamics of user perceptions. The analytical workflow comprised app identification and screening, extraction and preprocessing of ratings and reviews, and supply- and demand-side analyses.



XSL•FO
RenderX

**Figure 1.** Supply-demand analytical framework of medical education apps in China and the United States.




On the supply side, this study analyzes annual cumulative counts and temporal trends of observable and analytically eligible medical education apps from 2009 to 2025 based on app store data from China and the United States. Joinpoint regression models identify data-derived temporal segments and potential inflection points in observable app growth, thereby describing temporal patterns within the 2025-observable analytic cohort.

On the demand side, this study analyzes app store reviewers' primary concerns and rating patterns using large-scale user-generated review data. First, by combining latent Dirichlet allocation (LDA) topic modeling with large language model–assisted topic labeling using DeepSeek-R1, we identified key themes in app store reviews. Subsequently, Tobit regression models examined the associations between topic probabilities and PD and ND, while Wald tests assessed asymmetry between these associations. Building on these results, the Kano model was used to classify the identified topics into functional attributes according to their observed associations with rating deviations.

By integrating the supply-side analysis of observable app development patterns with the demand-side analysis of review topics and rating deviations, this study provides a unified framework for comparing medical education apps in China and the United States. This framework supports the comparison of data-derived temporal patterns and topic-rating deviation associations across the 2 country-specific samples, directly addressing the research questions posed in this study.



XSL•FO
RenderX

## Protocol Registration and Deviations

This study was registered on the Open Science Framework [40]. The overall research objectives, cross-national comparative design, study scope, and supply-demand analytical framework remained consistent with the registered protocol. However, several deviations and methodological refinements were introduced during implementation because the availability, structure, and quality of the app store data could not be fully determined in advance. First, app identification was supplemented by screening related or recommended apps displayed by the app store platforms in addition to retrieval using the prespecified search terms. This refinement was introduced to reduce the risk of missing eligible apps that used atypical titles or descriptions. Second, the eligibility criteria and review preprocessing procedures were operationalized in greater detail, including the requirements concerning recent app updates, minimum review counts, duplicate and noninformative records (including developer replies), suspected inauthentic or spam reviews, and rating-text concordance. These procedures were specified to ensure that the final sample was observable, analytically eligible, and suitable for linking textual content with rating-based outcomes. Third, after LDA topic modeling, AI-assisted candidate topic labels were subjected to consensus-based review by 3 research team members and independent structured assessment by 2 domain experts. This additional human validation step was introduced to improve the interpretability and semantic validity of the topic labels. These deviations and refinements did not alter the study's primary objectives or overall analytical framework; they were made to improve app coverage, data quality, analytic feasibility, reproducibility, and methodological transparency.

## Data Sources and Sample Construction

### *App Selection*

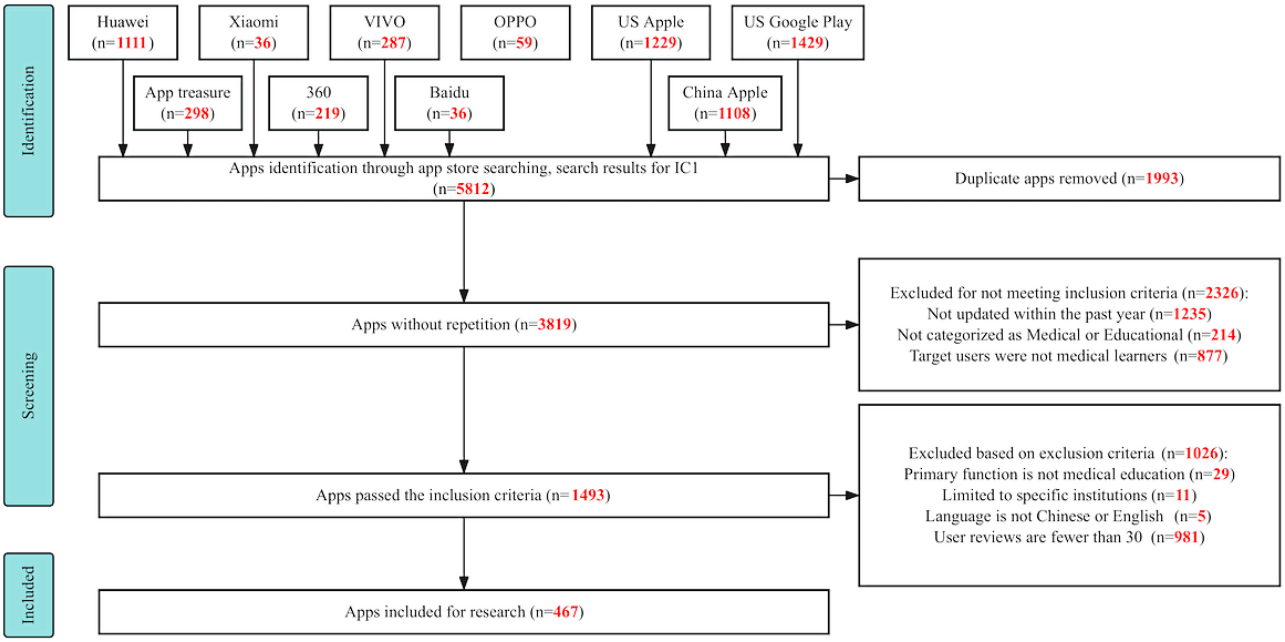
To comprehensively cover medical education apps in China and the United States, this study systematically searched mainstream app stores in both countries in September 2025. In China, the search encompassed the Apple App Store and 7 major Android platforms (Huawei, Xiaomi, OPPO, VIVO, Baidu, 360, and App Treasure). In the United States, it covered the Apple App Store and Google Play. Searches used Chinese and English keywords (eg, "medical education," "medical examinations," "USMLE"; detailed strategy in Table S1 in Multimedia Appendix 1), initially identifying 5812 apps. This keyword-based app store search strategy was consistent with prior systematic searches of medical education apps, which used predefined search terms followed by title-and-description screening and eligibility assessment to identify relevant educational apps [31]. To reduce the risk of missing apps that used alternative descriptions, the keyword set was designed to include broad bilingual terms related to medical education, medical examinations, clinical learning, and professional training. In addition to keyword-based retrieval, related or recommended apps displayed by the app-store platforms were reviewed where available and screened using the same eligibility criteria. Retrieved apps were then screened according to app titles, descriptions, categories, and predefined eligibility criteria rather than keywords alone.

The predefined eligibility criteria were as follows. Apps were included if they (1) were identified through the prespecified Chinese or English search terms listed in Table S1 in Multimedia Appendix 1 or through related or recommended apps displayed by the app store platforms during the search process; (2) had an update record within the past year as shown in the app store at the September 2025 search cutoff; (3) were categorized under "Medical" or "Education"; and (4) primarily targeted medical learners, including medical students, residents, medical graduate school applicants, or users preparing for the National Medical Licensing Examination or the USMLE. Apps were excluded if they (1) primarily provided noneducational or nonsystematic medical or health-related functions, such as pharmaceutical electronic commerce, wellness or health preservation, news or information services, medical social networking, or doctor-patient communication; (2) were institution-restricted internal systems of hospitals or medical schools; (3) operated in languages other than Chinese or English; or (4) had fewer than 30 platform reviews as of the search cutoff. These criteria are also summarized in Table S2 in Multimedia Appendix 1. To ensure rigorous and objective screening, 2 researchers (JM and J Liang) independently conducted selection and evaluation based on these predefined criteria [41], with consistency verified via the kappa coefficient (Cohen $\kappa$=0.87). Disputed results were adjudicated by a third expert (YC). After removing 1993 duplicate apps, 3819 apps remained for screening. Of these, 2326 did not meet the inclusion criteria, leaving 1493 apps. A further 1026 apps were excluded based on the exclusion criteria. Ultimately, 467 medical education apps were included, comprising 192 apps from China and 275 apps from the United States (refer to Figure 2 for the complete screening process). This rigorous screening process ensured the systematic nature, comparability, and reproducibility of the sample. The final analytic sample was defined based on information available at the September 2025 search cutoff. Therefore, apps that had been discontinued before the search cutoff and were no longer searchable or had no retrievable metadata could not be systematically included in the historical denominator. The update-record and review-count criteria were used to define apps that were observable and analytically eligible at the time of data collection.



XSL•FO
RenderX

**Figure 2.** Flowchart of the medical education apps screening process. IC1: inclusion criterion 1; n: the number of apps.



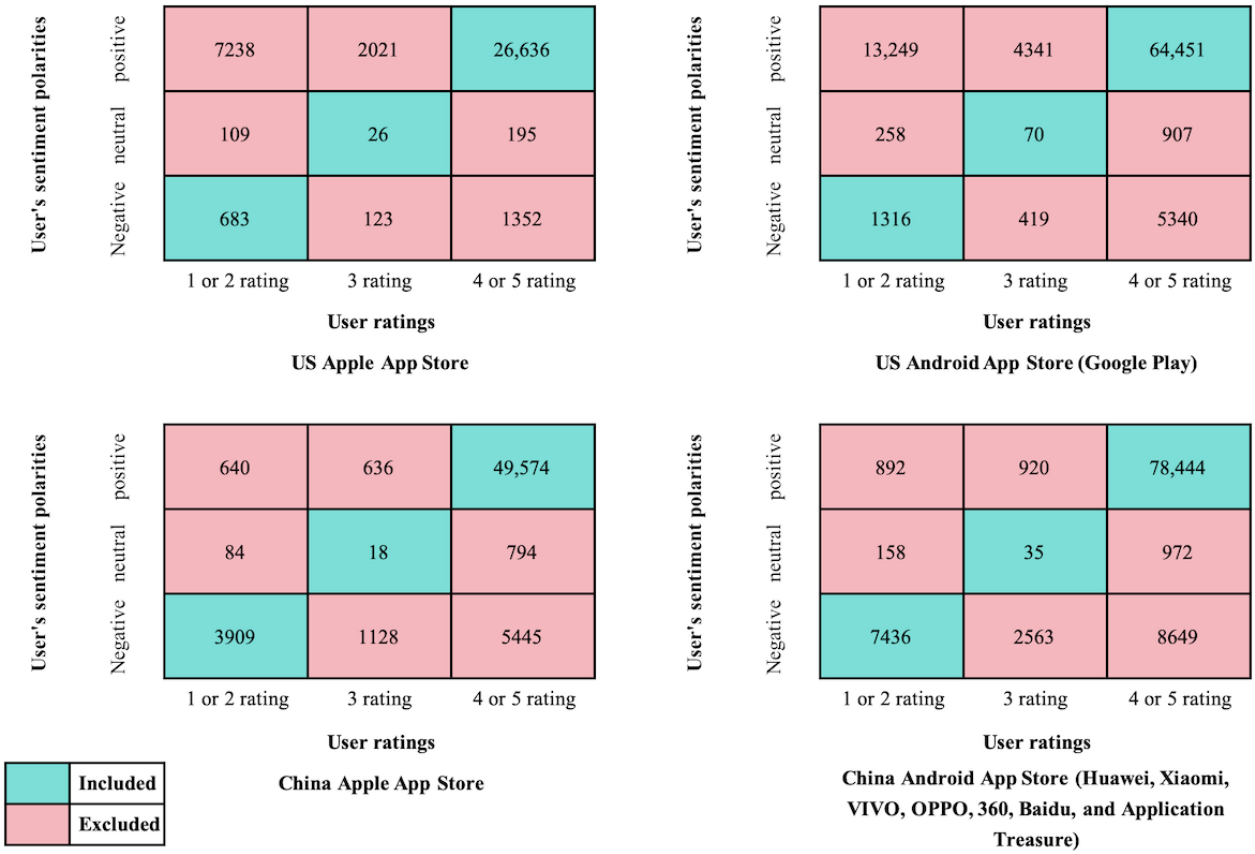

### *User-Generated Data*

To obtain real-world user feedback, this study collected user ratings and review data for the 467 selected apps across 10 major app stores in China and the United States. As of October 2025, a total of 498,760 retrievable reviews were collected (China: 321,911; United States: 176,849). All reviews underwent standardized preprocessing, including removing duplicate records and records with missing or noninformative content (including official developer replies), as well as excluding suspected inauthentic or spam reviews [42]. Sentiment polarity was classified as negative, neutral, or positive using the Baidu AI Cloud Sentiment Analysis API [43]. Rating-text discordant reviews were then excluded as a quality-control step to construct a rating-text-concordant analytic subset (refer to Figure 3 and Note S1 in Multimedia Appendix 1). Figure 3 presents 4 platform groups: the Apple App Store in China, the 7 Chinese Android app stores combined, the Apple App Store in the United States, and Google Play in the United States. Rows represent negative, neutral, and positive sentiment, and columns represent ratings of 1-2 stars, 3 stars, and 4-5 stars. Some low-rated reviews in the Chinese Android app stores expressed positive sentiment, whereas some high-rated reviews in the US Apple App Store expressed negative sentiment. Rating-text discordant reviews were not interpreted as definitively invalid records, because prior online review studies have shown that numerical ratings and textual sentiment do not always align [44,45]. Such discordance may reflect mixed evaluations, feature-specific complaints, sarcasm, or rating noise. The detailed exclusion criteria, the number of reviews removed during each quantitative preprocessing step, and the proportions excluded by the rating-text concordance screen are provided in Note S1 in Multimedia Appendix 1. After preprocessing, the final rating-text-concordant analytic subset comprised 232,598 reviews (China: 139,416; United States: 93,182). Drawing from prior research [46], we operationalized the rating-based satisfaction proxy rate as the proportion of reviews with ratings of 4 stars or higher among app store reviewers. This proxy was not interpreted as a direct measure of satisfaction in the full user population. To eliminate rating scale differences across apps, user ratings were converted into deviation values relative to the corresponding app's overall rating. Based on this, PD and ND were calculated, both ranging from 0 to 4. To ensure analytic consistency, the final review-level dataset retained only records with valid app identifiers, review text, user ratings, and review dates. All subsequent review-level analyses, including topic modeling, Tobit regression, Wald tests, and Kano classification, were conducted using this rating-text-concordant analytic subset.



XSL•FO
RenderX

**Figure 3.** Matching between user sentiment polarities (rows) and user ratings (columns).




## Sample Comparability

The Chinese and US samples maintained consistency in search time, platform coverage, inclusion criteria, and data processing procedures to ensure comparability across cultures. Through this methodology, this study constructed a 2-sided dataset that links observable app-market supply with app-store review-based demand signals, laying the foundation for subsequent supply-side trend analysis and demand-side rating-deviation modeling. The same search period, screening criteria, preprocessing workflow, and analytic procedures were applied to the Chinese and US samples to improve cross-national comparability. However, individual-level demographic information of app-store reviewers, such as age, sex, education level, professional background, and training stage, was not available from the app stores. Therefore, direct demographic adjustment was not possible. To reduce platform-related heterogeneity, we applied identical eligibility criteria, preprocessing procedures, rating-deviation construction, and analytic workflows across all app stores, and all rating-deviation models were estimated separately by country and developmental phase.

## Supply-Side Analysis: Cumulative Observable App Counts and Temporal Trends

To characterize supply-side temporal patterns, this study used the cumulative count of apps that remained observable and analytically eligible at the September 2025 search cutoff as the core metric. In this study, "survival" was operationalized as an app being retrievable from the included app stores or Qimai Apps Data platform (Beijing QiMai Technology Co, Ltd) at the search cutoff, having an update record within the preceding year, and meeting the minimum review-count criterion. Accordingly, the annual cumulative survival counts should be interpreted as temporal patterns within the 2025-observable analytic cohort, rather than as a complete historical census of all medical education apps that had ever existed in each year. Therefore, the terms "survival" and "inflection points" in this study refer to app availability and trend shifts within the 2025-observable analytic cohort, not to the survival history of all medical education apps ever released. In time-series research, relevant literature has used joinpoint regression analysis to model trends in publicly available data (such as app counts, user interest, or technological attention) to identify significant inflection points in developmental trajectories and reveal phased market shifts [47]. This method avoids predefined time segmentation by identifying trend structure shifts directly from the data, effectively mitigating subjective classification biases. Drawing on this research paradigm and considering the potential phased development characteristics of the medical education app market, this study uses joinpoint regression to analyze the temporal trends in cumulative counts of observable and analytically eligible apps. This approach identifies data-derived temporal segments, providing methodological support for subsequent phase-based analysis grounded in user reviews.

Specifically, we used Joinpoint software (version 5.4.0; National Cancer Institute, November 2025) to construct regression models on a logarithmic scale. Using annual cumulative counts of observable and analytically eligible medical education apps as



XSL•FO
RenderX

the dependent variable, we estimated annual percentage changes (APCs) and average annual percentage changes (AAPCs) for each time segment, supplemented by pairwise tests to assess trend parallelism. The model allowed 0 to 3 joinpoints, selected based on data-driven Bayesian information criterion (BIC) with a global significance level of 0.05. Poisson SEs were used to estimate variability under the assumption of Poisson-distributed count data. Finally, the AAPC and its 95% CI for each trend segment were calculated from the slope of the log-transformed regression line, systematically revealing temporal dynamics.

Based on critical inflection points identified through joinpoint regression analysis, user review data underwent temporal segmentation. Analysis revealed key inflection points occurring in both China and the United States in 2017, dividing the data into 2 developmental phases: “2009-2017” and “2018-2025.” Although the first included Chinese app appeared in 2010, the first phase was labeled 2009-2017 to maintain a unified cross-national comparison window; the Chinese sample had no included apps in 2009. APC estimates that did not reach statistical significance were also considered in each phase to ensure that the subsequent analysis covered the entire temporal trajectory. Through this segmentation strategy, the study examined app-related factors associated with PD and ND and assessed how these associations varied across temporal phases.

## Demand-Side Analysis: User Reviews and Rating-Deviation Modeling

### *Review Topic Identification (Text Processing and LDA Modeling)*

To extract core themes from user reviews of medical education apps, the study first performed text preprocessing on Chinese and English reviews. This included tokenization using the jieba library [48] and the Natural Language Toolkit (NLTK), stopword removal using Chinese and English stopword lists, and cleaning numerical values, punctuation, and emoticons. Subsequently, the LDA topic model [49] was applied to analyze review data across different phases, identifying core themes and their distribution. The number of themes was determined comprehensively by balancing perplexity and coherence metrics with the interpretability of thematic clusters, ensuring themes were both statistically robust and understandable (Figures S1-S4 in Multimedia Appendix 1). This step structured review texts into quantifiable topic probability distributions, providing a quantitative foundation for the subsequent rating-deviation analysis. Visualization results, detailed text processing workflows, parameter settings, and thematic modeling specifics are documented in Figures S5-S8 and Note S2 in Multimedia Appendix 1. To address cross-linguistic differences, topic modeling was conducted separately for Chinese and English review corpora rather than by pooling translated texts. Chinese and English reviews were preprocessed using the same standardized preprocessing workflow, and topic labels were generated and validated separately within each country-phase corpus. For reporting and cross-national comparison, Chinese topic labels were translated into English after modeling and checked by bilingual researchers against the original high-probability keywords. Therefore, topics were compared at the level of conceptual domains, such as exam preparation, cost, technical stability, and learning support, rather than being treated as strictly one-to-one equivalent linguistic categories across languages.

### *Topic Naming and Semantic Validation (Assisted by DeepSeek-R1)*

To enhance topic interpretability, DeepSeek-R1 was used to assist with the semantic labeling of topics generated by the LDA models. For each topic, the top 30 high-probability keywords were entered into the model to generate candidate topic names and brief descriptions (Table S3 in Multimedia Appendix 1). The model was accessed on October 26, 2025, with the temperature set to 0.6 and all other parameters kept at their default settings. After AI-assisted label generation, 3 members of the research team conducted a consensus-based semantic review by comparing each candidate topic name and description with the corresponding high-probability keywords. Disagreements were resolved through discussion, and the final topic labels were determined by human consensus. To provide additional verification, 2 domain experts independently assessed the finalized topic labels using a structured 5-point rubric for name accuracy and description interpretability (Table S4 in Multimedia Appendix 1). Because this validation task involved a small number of rubric-based semantic assessments of finalized topic names and descriptions, rather than independent categorical coding of individual reviews into topic categories, Cohen κ was not calculated. Instead, exact agreement between the two experts and expert validity scores were reported as descriptive indicators of topic-label validation quality. This human-in-the-loop approach was used to combine the efficiency of AI-assisted semantic labeling with the contextual judgment of human reviewers [50]. Generative AI was not used for data collection, review preprocessing, topic modeling, statistical analysis, result interpretation, or conclusion generation. To strengthen the computational validity and robustness of the text-mining procedure, we adopted a staged human-in-the-loop workflow. LDA was used as an unsupervised topic modeling approach to reduce high-dimensional review text into latent topic probability distributions, rather than to generate final conceptual categories directly. The number of topics was determined by combining perplexity, coherence, and human interpretability. AI-assisted candidate labels were subsequently reviewed through the consensus-based and expert-validation process described above. Similar natural language processing approaches have been used to analyze unstructured user-generated digital text in health-related contexts [51]. Therefore, the resulting topics should be interpreted as structured approximations of recurrent user-review themes rather than exhaustive or definitive categories of all user needs.

### *Asymmetric Modeling of Rating Deviations (Tobit Regression and Wald Test)*

Guided by Herzberg two-factor theory, which distinguishes positive and negative evaluations, this study modeled PD and ND. PD and ND were used as dependent variables to represent the extent to which an individual review rating deviated above or below the corresponding app’s overall rating, while LDA topic probability distributions were used as independent variables to quantify the thematic composition of each review.



XSL•FO
RenderX

Because PD and ND are bounded dependent variables with a lower bound of zero, Tobit regression was used to examine the associations between review topics and rating deviations. Before model estimation, multicollinearity was assessed using variance inflation factors (VIFs); all VIF values were below 5, indicating no substantial multicollinearity [52] among the independent variables (Tables S5-S12 in Multimedia Appendix 1). Tobit regression is commonly used for limited dependent variables with restricted distributions [53,54]. Wald tests were subsequently conducted to test the null hypothesis that the PD coefficient for each topic equaled the sign-reversed coefficient from the corresponding ND model ($H_0$: $\beta_{PD,k} = -\beta_{ND,k}$), thereby evaluating asymmetry between the PD and ND associations. Detailed procedures for constructing PD and ND, model specification, and statistical testing are provided in Note S3 in Multimedia Appendix 1.

### *Functional Attribute Classification Using the Kano Model*

Based on the asymmetric associations identified in the PD and ND models, each review topic was further classified into a functional attribute using the Kano model. The resulting categories included essential factors, expectation factors, charm factors, no differential factors, and reverse factors. The classification criteria integrated statistical characteristics from the PD and ND models, including coefficient significance, coefficient direction, and the significance of coefficient differences [55]. This classification was used to translate the observed rating-deviation patterns into potentially actionable categories for app design and optimization. Detailed classification criteria and procedures are provided in Figure S9, Table S13, and Note S4 in Multimedia Appendix 1.

### *Methodological Integration and Analytical Advantages*

This study systematically analyzed reviews of medical education apps by integrating phase-specific topic modeling, asymmetric rating-deviation modeling, and Kano functional attribute classification. This methodological framework aligns the review analysis with the data-derived temporal segments identified from the observable app cohort, thereby supporting longitudinal comparisons. By modeling PD and ND separately, the analysis avoided collapsing the 2 directions of rating deviation into a single outcome. The Kano model then translated the observed topic-rating deviation associations into functional attribute categories that may inform the design and optimization of medical education apps.

## Statistical Analysis

Data entry and organization for this study were performed using Microsoft Excel 2024. Data preprocessing, LDA topic modeling, and visualization were implemented using R 4.3.3 (R Core Team) and Python (*Gensim*, *PyLDAvis*, and other packages). The temporal trend analysis of cumulative counts of observable and analytically eligible medical education apps used the joinpoint regression method, executed using Joinpoint software (version 5.4.0, National Cancer Institute). Tobit regression and Wald tests of the associations between review topics and PD and ND were performed using Stata 18.0 (StataCorp LLC). All statistical tests were 2-tailed, with the significance level set at $\alpha$=.05.

## Ethical Considerations

All data used in this study were sourced from publicly accessible websites, and all user information was anonymized. Sensitive data were not collected for this study, and no compensation was provided for participants. All data relevant to this study were stored on a password-encrypted computer, and only the researchers had access to the data. Ethics approval was not required according to the ethical review by the ethics committee of the Second Affiliated Hospital Zhejiang University School of Medicine.

# Results

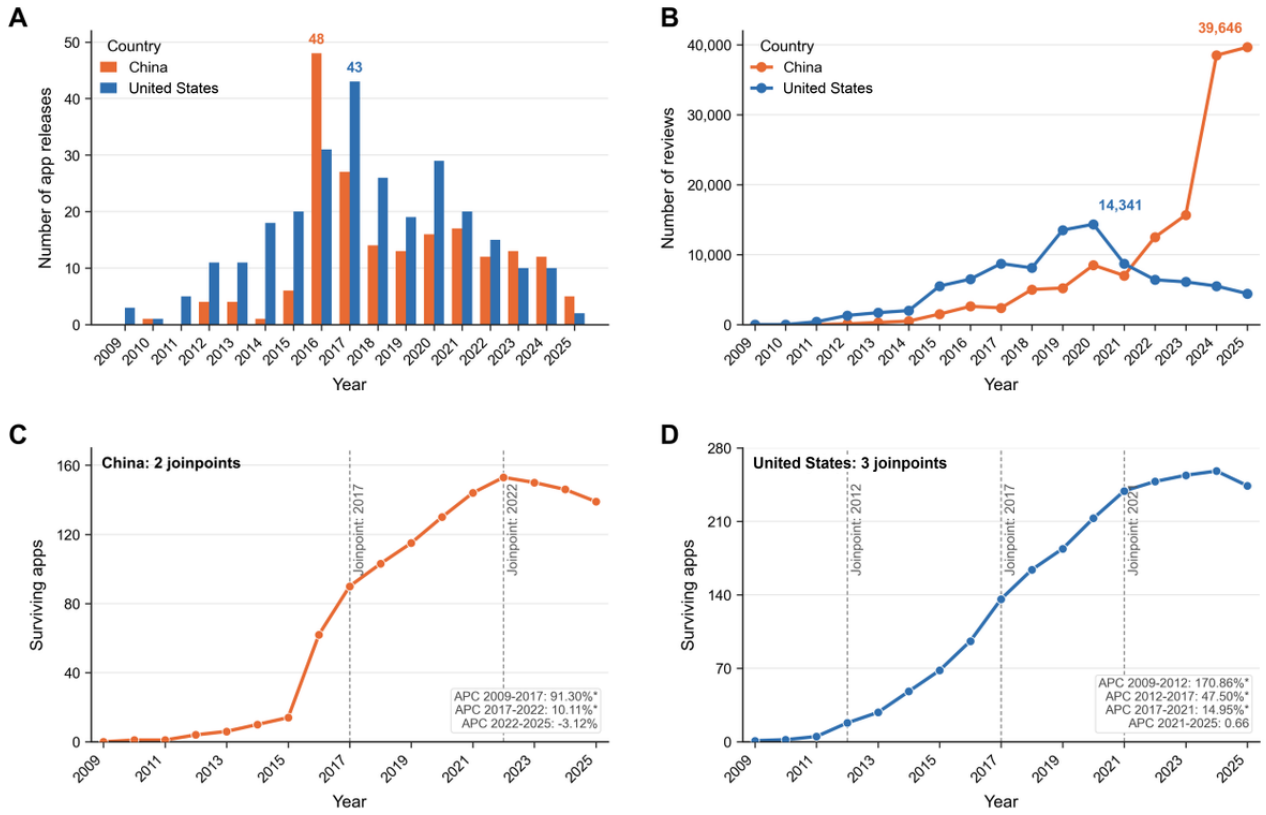
## Market Evolution and Phased Changes of Medical Education Apps in China and the United States

The final analytic sample included 467 medical education apps, comprising 192 apps from China and 275 apps from the United States. Among all reviews, 219,105 were rated 4 stars or higher, corresponding to an overall rating-based satisfaction proxy rate of 94.20%. By region, the rating-based satisfaction proxy rates for medical education apps in China and the United States were 91.82% (128,018/139,416) and 97.75% (91,087/93,182), respectively. This baseline difference in the rating-based satisfaction proxy rate was interpreted as a descriptive cross-national contrast rather than as direct evidence of higher or lower educational quality in either market. Because app store ratings are generated by self-selected reviewers, the observed difference may reflect multiple factors, including app quality, platform review moderation practices, reviewer composition, and culturally shaped rating behaviors. Therefore, we further examined whether the baseline rating difference was accompanied by distinct review topics and asymmetric rating-deviation patterns across countries and temporal phases. Cumulative downloads reached 361,367,013 for Chinese apps and 28,189,607 for US apps. Figure 4 presents the annual numbers of app releases (Figure 4A), annual user review counts (Figure 4B), and joinpoint regression analyses of annual cumulative observable app counts for China (Figure 4C) and the United States (Figure 4D) within the 2025-observable analytic cohort. Apps in the Chinese market first appeared in 2010, reaching an annual release peak of 48 in 2016, followed by a gradual decline in new app releases each year thereafter. The US market began in 2009, peaking in 2017 (43 apps), and also showed a declining trend thereafter. Therefore, the 2009-2017 phase for China should be understood as a harmonized comparative period, with no included Chinese apps observed in 2009. Regarding user reviews, the Chinese market exhibited sustained overall growth; the US market reached a review peak in 2020 (14,341 reviews) before gradually declining.



XSL•FO
RenderX

**Figure 4.** Comparison of development trends in medical education apps between China and the United States: annual number of app releases, user review volume, and joinpoint regression analysis results. APC: annual percentage change.




Joinpoint results indicate that, within the 2025-observable analytic cohort, the annual cumulative counts of observable and analytically eligible medical education apps in both China and the United States showed significant growth ($P$<.001) throughout the observation period. The AAPCs were 41.69% (95% CI 31.69-52.45) and 41.17% (95% CI 31.61-51.42) for China and the United States, respectively, indicating similar overall growth rates. Specifically, China's cumulative observable app counts exhibited distinct inflection points in 2017 and 2022, revealing a 3-phase evolutionary pattern: rapid growth from 2009 to 2017 (APC=91.30%, 95% CI 64.36-122.66; $P$<.001), decelerating growth from 2017 to 2022 (APC=10.11%, 95% CI 0.68-20.43; $P$=.04), and a nonsignificant decline after 2022 (APC=–3.12%, $P$=.59). The cumulative observable app counts in the United States exhibited turning points in 2012, 2017, and 2021, following a 4-phase trajectory: a steep rise from 2009 to 2012 (APC=170.86%, 95% CI 72.86-324.44; $P$=.002), followed by sustained high growth from 2012 to 2017 (APC=47.50%, 95% CI 38.78-56.77; $P$<.001), a further deceleration in growth from 2017 to 2021 (APC=14.95%, 95% CI 9.36-20.83; $P$<.001), and no significant change after 2021 ($P$=.58).

## Topic Distribution in User Reviews of Medical Education Apps

Tables 1 and 2 illustrate the thematic distribution of user reviews for medical education apps in China and the United States across different developmental stages. Overall, user-focused topics exhibit phased evolution over time. In China, from 2009 to 2017, users primarily emphasized "Clinical practicality" and "Technical stability," whereas from 2018 to 2025, the focus shifted toward "Practice test function." In the United States, users primarily focused on "Functional availability" from 2009 to 2017, while from 2018 to 2025, attention shifted toward "Cost" and "Exam preparation and practice questions." Detailed analysis results are provided in Note S5 in Multimedia Appendix 1.



XSL•FO
RenderX

**Table 1.** Review topics and keywords identified using latent Dirichlet allocation topic modeling (China).

| Stage and topics | Keywords | Reviews, n (%) |
|---|---|---|
| 2009-2017 | | |
| Topic 1: Clinical practicality | clinical, practical, guide, job, function, content | 2735 (37.03) |
| Topic 2: Professional learning community | platform, communication, discussion, share, recommended, medical students | 1070 (14.49) |
| Topic 3: Technical stability | flashback, unable, solve, update, installation, professional version | 1555 (21.06) |
| Topic 4: Medical knowledge question bank | medications, exam, e-book, complete, physician, instructions | 694 (9.40) |
| Topic 5: Cost | charge, cheap, offline, version, time, use | 203 (2.75) |
| Topic 6: Content reliability | professional, certification, medical school, course, case, bone | 325 (4.40) |
| Topic 7: Learning and practice support | postgraduate entrance examination, medical record, patient, work, growth, companionship | 511 (6.92) |
| Topic 8: Infant medication inquiry | baby, medicine, lactation, internet, department, not complete | 292 (3.95) |
| 2018-2025 | | |
| Topic 1: Practice test function | practice problems, title, analysis, past exam papers, question mistakes, postgraduate entrance exam | 53,929 (40.85) |
| Topic 2: Audio-visual teaching function | explanation, course, video, lecture, teacher, interesting | 11,197 (8.48) |
| Topic 3: Medical knowledge question bank | question bank, question type, content-rich, software, exam preparation, easy to understand | 12,518 (9.48) |
| Topic 4: Functional usability | like, gospel, with heart, satisfied, treasure, value for money | 10,309 (7.81) |
| Topic 5: Cost | free, advertisement, charges, purchase, unlock, money | 10,860 (8.23) |
| Topic 6: Technical stability | crash, collapse, unable to open, operation, resolve, type | 9230 (6.99) |
| Topic 7: Professional learning community | study, share, companionship, good helper, appreciate, skill | 4980 (3.77) |
| Topic 8: Nurse qualification examination | nursing certification exam, occupation, nursing licensing examination, anatomy, attend lectures, subject | 6250 (4.73) |
| Topic 9: User-perceived value | recommended, worth, premium, support, knowledge, anatomy | 1837 (1.39) |
| Topic 10: Personalized learning | precision, relaxed, smoothly, test center, function, convenient and quick | 1063 (0.81) |
| Topic 11: User service | customer service, update, refund, cancel registration, reply, store, answering questions | 2331 (1.76) |
| Topic 12: Cross-screen adaptation function | tablet, landscape mode, full screen, adapt, switch, play | 7527 (5.70) |



XSL•FO
RenderX

**Table 2.** Review topics and keywords identified using latent Dirichlet allocation topic modeling (United States).

| Stage and topics | Keywords | Reviews, n (%) |
|---|---|---|
| 2009-2017 | | |
| Topic 1: Practice test function | questions, quiz, practice, exam, explanation, study | 72 (0.28) |
| Topic 2: Usability | simple, guide, step, learn, fun, educational | 1283 (4.91) |
| Topic 3: Functional availability | feature, add, search, offline, mobile, flashcard | 20,736 (79.39) |
| Topic 4: Interface design and visual clarity | interface, picture, organize, easy, understand, quick | 1083 (4.15) |
| Topic 5: Content accuracy and reliability | accurate, reliable, medical, resource, professional, student | 677 (2.59) |
| Topic 6: Clinical teaching practice | doctor, surgery, patient, informative, teach, reference | 957 (3.66) |
| Topic 7: Cost | free, pay, price, buy, purchase, offer | 126 (0.48) |
| Topic 8: Learning management and note-taking | set, create, note, edit, delete, sync | 1045 (4) |
| Topic 9: Technical stability | update, issue, download, load, Android, check | 139 (0.53) |
| 2018-2025 | | |
| Topic 1: Cost | free, pay, money, subscription, premium, purchase | 27,401 (40.86) |
| Topic 2: Exam preparation and practice questions | quiz, exam, test, practice, pass, game | 9000 (13.42) |
| Topic 3: Error analysis and memorization function | wrong, explanation, rationale, review, flash, revision | 6000 (8.95) |
| Topic 4: Personalized learning services | feature, track, progress, option, set, enjoy | 4413 (6.58) |
| Topic 5: Nursing qualification exam | nursing, preparation, aspirant, exercise, learn, knowledge | 3948 (5.89) |
| Topic 6: Clinical teaching practice | anatomy, physiology, clinical, flashcard, skill, understanding | 4137 (6.17) |
| Topic 7: Convenience | simple, quick, resource, reference, platform, recommend | 2911 (4.34) |
| Topic 8: Teaching quality | easy, understand, explain, lecture, teacher, teach | 1373 (2.05) |
| Topic 9: Learning outcomes and confidence building | grade, confident, effective, accurate, simply, ready | 7156 (10.67) |
| Topic 10: Clinical decision support | patient, treatment, medication, quality, science, AI | 725 (1.08) |

## Associations Between Review Topics and Rating Deviations

### *Tobit Regression Results for PD and ND*

The Tobit models revealed country-specific and phase-specific patterns in the associations between review topics and rating deviations. In China, reviews from the early phase were more strongly characterized by clinical practicality, content reliability, technical stability, cost, and learning support, whereas reviews from the later phase were more concentrated around examination preparation, question banks, user service, technical stability, and cost-related experiences. In the United States, reviews from the early phase were more strongly characterized by functional availability, interface clarity, content reliability, and usability, whereas reviews from the later phase showed greater attention to cost, exam preparation, convenience, and learning support. These patterns indicate that rating deviations were associated with different app-related topics across countries and temporal phases. Tables 3 and 4 present the Tobit regression results for medical education apps in China and the United States. These tables report complete estimates for all LDA-derived topics before subsequent factor discrimination and Kano classification. Overall, both the PD and ND models showed that multiple review topics were significantly associated with rating deviations. In China, during Phase I (2009-2017), content reliability showed the strongest positive association with PD (regression coefficient β=0.488, $P$<.001), whereas cost showed the strongest negative association with PD (β=–0.459, $P$<.001). In the ND model, cost showed the strongest positive association with ND (β=2.704, $P$<.001), whereas technical stability showed the strongest negative association with ND (β=–4.464, $P$<.001). During Phase II (2018-2025), user service showed the strongest positive association with PD (β=0.196, $P$<.001), whereas user-perceived value showed the strongest negative association with PD (β=–0.134, $P$<.001). In the ND model, user-perceived value showed the strongest positive association with ND (β=3.368, $P$<.001), whereas the nurse qualification examination showed the strongest negative association with ND (β=–1.774, $P$<.001).



XSL•FO
RenderX

**Table 3.** Tobit regression results for rating deviations in China.

| Stage and review topic | Model 1 (PD[a]) β | SE | *P* value | Model 2 (ND[b]) β | SE | *P* value |
|---|---|---|---|---|---|---|
| 2009-2017 | | | | | | |
| Topic 1: Clinical practicality | 0.254 | 0.034 | <.001 | –0.051 | 0.179 | .77 |
| Topic 2: Professional learning community | –0.445 | 0.036 | <.001 | 2.432 | 0.170 | <.001 |
| Topic 3: Technical stability | 0.180 | 0.032 | <.001 | –4.464 | 0.240 | <.001 |
| Topic 4: Medical knowledge question bank | 0.380 | 0.047 | <.001 | –0.988 | 0.210 | <.001 |
| Topic 5: Cost | –0.459 | 0.071 | <.001 | 2.704 | 0.226 | <.001 |
| Topic 6: Content reliability | 0.488 | 0.070 | <.001 | 0.779 | 0.232 | .001 |
| Topic 7: Learning and practice support | 0.126 | 0.037 | .001 | –1.876 | 0.254 | <.001 |
| Topic 8: Infant medication inquiry | 0.024 | 0.040 | .55 | –1.166 | 0.308 | <.001 |
| 2018-2025 | | | | | | |
| Topic 1: Practice test function | 0.031 | 0.004 | <.001 | –1.214 | 0.039 | <.001 |
| Topic 2: Audio-visual teaching function | –0.108 | 0.005 | <.001 | 3.286 | 0.040 | <.001 |
| Topic 3: Medical knowledge question bank | –0.025 | 0.004 | <.001 | 0.584 | 0.045 | <.001 |
| Topic 4: Functional usability | 0.087 | 0.005 | <.001 | –1.503 | 0.025 | <.001 |
| Topic 5: Cost | 0.042 | 0.004 | <.001 | –1.441 | 0.051 | <.001 |
| Topic 6: Technical stability | 0.153 | 0.006 | <.001 | –1.127 | 0.054 | <.001 |
| Topic 7: Professional learning community | 0.115 | 0.006 | <.001 | –1.672 | 0.067 | <.001 |
| Topic 8: Nurse qualification examination | 0.097 | 0.005 | <.001 | –1.774 | 0.065 | <.001 |
| Topic 9: User-perceived value | –0.134 | 0.009 | <.001 | 3.368 | 0.052 | <.001 |
| Topic 10: Personalized learning | 0.021 | 0.009 | .03 | 1.329 | 0.073 | <.001 |
| Topic 11: User service | 0.196 | 0.008 | <.001 | –0.774 | 0.070 | <.001 |
| Topic 12: Cross-screen adaptation function | 0.124 | 0.005 | <.001 | –0.626 | 0.049 | <.001 |

[a]PD: positive rating deviations. The log likelihoods for Model 1 were –9967.3345 (2009-2017) and –10993.34 (2018-2025).

[b]ND: negative rating deviations. The log likelihoods for Model 2 were –5175.8596 (2009-2017) and –62667.978 (2018-2025).



XSL•FO
RenderX

**Table 4.** Tobit regression results for rating deviations in the United States.

| Stage and review topic | Model 1 (PD[a]) β | SE | *P* value | Model 2 (ND[b]) β | SE | *P* value |
|---|---|---|---|---|---|---|
| 2009-2017 | | | | | | |
| Topic 1: Practice test function | –0.883 | 0.017 | <.001 | 0.271 | 0.064 | <.001 |
| Topic 2: Usability | 0.127 | 0.013 | <.001 | –0.413 | 0.053 | <.001 |
| Topic 3: Functional availability | 0.128 | 0.031 | <.001 | –0.342 | 0.107 | .001 |
| Topic 4: Interface design and visual clarity | 0.216 | 0.013 | <.001 | –0.490 | 0.057 | <.001 |
| Topic 5: Content accuracy and reliability | 0.185 | 0.014 | <.001 | –0.605 | 0.062 | <.001 |
| Topic 6: Clinical teaching practice | –0.018 | 0.013 | .16 | 0.216 | 0.050 | <.001 |
| Topic 7: Cost | –0.072 | 0.019 | <.001 | 1.537 | 0.069 | <.001 |
| Topic 8: Learning management and note-taking | –0.121 | 0.015 | <.001 | 0.863 | 0.051 | <.001 |
| Topic 9: Technical stability | –0.017 | 0.015 | .27 | 0.164 | 0.064 | .01 |
| 2018-2025 | | | | | | |
| Topic 1: Cost | –0.033 | 0.004 | <.001 | 0.382 | 0.027 | <.001 |
| Topic 2: Exam preparation and practice questions | 0.066 | 0.005 | <.001 | –0.449 | 0.032 | <.001 |
| Topic 3: Error analysis and memorization function | 0.125 | 0.005 | <.001 | –0.804 | 0.038 | <.001 |
| Topic 4: Personalized learning services | 0.062 | 0.006 | <.001 | –0.041 | 0.035 | .24 |
| Topic 5: Nursing qualification exam | –0.188 | 0.006 | <.001 | 1.157 | 0.034 | <.001 |
| Topic 6: Clinical teaching practice | –0.131 | 0.006 | <.001 | 0.394 | 0.032 | <.001 |
| Topic 7: Convenience | 0.168 | 0.006 | <.001 | –0.574 | 0.041 | <.001 |
| Topic 8: Teaching quality | –0.004 | 0.008 | .64 | 0.112 | 0.045 | .01 |
| Topic 9: Learning outcomes and confidence building | –0.070 | 0.004 | <.001 | –0.888 | 0.034 | <.001 |
| Topic 10: Clinical decision support | –0.019 | 0.011 | .08 | 0.286 | 0.057 | <.001 |

[a]PD: positive rating deviations. The log likelihoods for Model 1 were –18560.334 (2009-2017) and –44306.823 (2018-2025).

[b]ND: negative rating deviations. The log likelihoods for Model 2 were –14143.986 (2009-2017) and –39305.998 (2018-2025).

In the United States, during Phase I (2009-2017), interface design and visual clarity showed the strongest positive association with PD (β=0.216, *P*<.001), whereas the practice test function showed the strongest negative association with PD (β=–0.883, *P*<.001). In the ND model, cost showed the strongest positive association with ND (β=1.537, *P*<.001), whereas content accuracy and reliability showed the strongest negative association with ND (β=–0.605, *P*<.001). During Phase II (2018-2025), convenience showed the strongest positive association with PD (β=0.168, *P*<.001), whereas the nursing qualification exam topic showed the strongest negative association with PD (β=–0.188, *P*<.001). In the ND model, the nursing qualification exam topic showed the strongest positive association with ND (β=1.157, *P*<.001), whereas learning outcomes and confidence building showed the strongest negative association with ND (β=–0.888, *P*<.001). Complete model results are provided in Note S6 in Multimedia Appendix 1.

### *Asymmetry Between PD and ND Coefficients*

The Wald tests compared the PD coefficient for each review topic with the sign-reversed coefficient from the corresponding ND model. A significant difference in this comparison indicated asymmetry between the positive and negative rating-deviation associations. In the Chinese sample, during Phase I (2009-2017), all topics included in the Wald comparison except “Clinical practicality” showed significant differences. During Phase II (2018-2025), all included topics showed significant differences. In the US sample, all topics included in the Wald comparison showed significant differences during Phase I (2009-2017). During Phase II (2018-2025), all included topics except “Personalized learning services” showed significant differences. Detailed results are presented in Tables S14 and S15 and Note S7 in Multimedia Appendix 1.

### *Kano Classification of Review Topics*

Furthermore, the Kano classifications of review topics differed across countries and temporal phases. In the Chinese sample, Phase I (2009-2017) included 1 essential factor (infant medication inquiry) and 6 charm factors, whereas Phase II (2018-2025) included 1 essential factor (user-perceived value) and 10 charm factors. In the US sample, Phase I (2009-2017) included 2 essential factors (clinical teaching practice and technical stability) and 7 charm factors, whereas Phase II (2018-2025) included 2 essential factors (teaching quality and clinical decision support) and 7 charm factors. Classification results and the corresponding classification indicators are presented in Figure S10 and Tables S16-S19 in Multimedia Appendix 1, respectively. Detailed analysis is provided in Note S8 in Multimedia Appendix 1. Taken together, the Kano classification suggested that app store reviewers in both



XSL•FO
RenderX

countries increasingly distinguished between baseline requirements and value-added features. Essential factors reflected functions whose absence was more likely to generate dissatisfaction, such as technical stability, clinical teaching practice, teaching quality, or clinical decision support, depending on country and phase. Charm factors reflected features that could increase satisfaction when present, particularly exam preparation, learning management, personalized learning, convenience, user service, and cross-screen adaptation. This distinction helped translate the Tobit regression results into more actionable insights for app design by separating foundational quality expectations from features that may create additional perceived value.

# Discussion

## Principal Findings

This study aimed to characterize the longitudinal evolution of medical education apps in China and the United States and to identify app-related factors associated with PD and ND among app store reviewers. Overall, within the 2025-observable analytic cohort, the findings showed phased development patterns, with notable data-derived inflection points around 2017 and 2021-2022, while country-specific trajectories differed in observable app availability, review activity, and user-evaluation patterns. On the demand side, Chinese app store reviewers showed stronger associations between rating deviations and exam preparation, cost, question banks, personalized learning, and user service, whereas US app store reviewers placed greater emphasis on technical stability, functional availability, content reliability, and clinical decision support. These findings address the study objectives by linking supply-side observable app development patterns with demand-side user review patterns in 2 distinct medical education contexts.

## Phased Evolutionary Characteristics of Medical Education Apps in China and the United States

### *Phased Inflection Points and Synchronization*

Based on joinpoint regression results, the 2025-observable medical education app cohorts in China and the United States exhibited similar data-derived temporal inflection points during the study period, primarily concentrated in 2017 and 2021-2022. The apparent similarity in these temporal patterns may be interpreted in relation to shared contextual developments, including the global diffusion of mobile internet technology, changes in digital medical education platforms, and shifts in the external environment. However, these factors were not directly measured, and the observed similarity should not be interpreted as evidence that they caused the temporal patterns. From a technological perspective, policy updates on the iOS platform may provide contextual support for interpreting the observed temporal patterns. Starting with iOS 11, all 32-bit apps were no longer supported, requiring developers to update their apps to a 64-bit architecture for operation [56]. As early as 2015, Apple mandated that new apps and updates support 64-bit architecture [57]. This policy may have contributed to the withdrawal of outdated or unmaintained applications by increasing technical requirements for app development and maintenance, and may therefore have influenced the composition of the observable app cohort during the study period. From a technology adoption perspective, while the TAM was introduced to medical education in 2010, it remained in an exploratory phase until 2016, with research primarily focused on electronic learning and mobile learning [58]. These developments in technology adoption may have created contextual conditions that supported growing awareness and use of mobile learning tools around 2017. Institutionally, China formally advanced its integrated clinical medical education system (“5+3” [5-year undergraduate + 3-year residency] and “3+2” [3-year specialty + 2-year assistant general practice training]) alongside medical education coordination reforms in 2017, emphasizing seamless integration between undergraduate education, residency training, and clinical practice [59]. This reform may have increased attention to standardized and practice-oriented clinical skills training and may have created an institutional context supportive of digital and mobile learning tools.

Changes in the external environment may also provide contextual explanations for the observed temporal patterns. Previous research has reported increased use of mobile medical education apps among medical students during COVID-19 lockdowns, with use remaining elevated after the pandemic [60]. This evidence may provide contextual support for interpreting the data-derived temporal changes observed around 2021-2022. More broadly, the increasing use of mobile learning may reflect its transition from a supplementary resource toward a more established educational approach.

### *Divergent Market Trajectories in China and the United States*

Despite similar data-derived inflection points, the observable cohort-based trajectories in China and the United States showed notable differences. Within the 2025-observable analytic cohort, China showed rapid early expansion followed by slower growth, while user reviews continued to grow. This may suggest that review activity and expressed engagement increased among app-store reviewers even as observable app growth slowed. In contrast, the US sample showed declining app listings after 2017, with review volumes peaking around 2020 before gradually declining. This pattern may be consistent with product consolidation within the observable app cohort and changing review activity in a more mature platform environment. Overall, the similarity in inflection points may reflect shared pressures in the global digital transformation of medical education, whereas the divergent trajectories may be related to differences in market maturity, institutional environments, and platform competition dynamics.

### *Dynamic Differences in Rating-Based Satisfaction Patterns for Medical Education Apps in China and the United States*

The observed pattern of rating-based satisfaction among Chinese app store reviewers showed distinct phased changes. During the early stage (2009-2017), reviewers emphasized clinical practicality, technical stability, and professional learning community. These patterns indicate that the directions and magnitudes of associations with PD and ND varied across review topics. In the later phase, reviewer attention was more



XSL•FO
RenderX

frequently directed toward exam preparation drills, question banks, personalized learning, user services, and cross-screen compatibility. These findings indicate that greater positive rating deviations were most strongly associated with user service, technical stability, and cross-screen adaptation, whereas greater NDs were most strongly associated with user-perceived value and audio-visual teaching function. This demand shift may be consistent with an assessment-driven learning context in high-stakes medical education. Because this study did not measure learning processes or outcomes, it cannot determine whether greater attention to question banks reflects rote memorization or is associated with superficial learning. As a broader educational design consideration, app-based exam preparation could be combined with activities such as problem-based learning (PBL) case discussions to support critical thinking and deep learning.

The observed pattern of rating-based satisfaction among US app store reviewers showed a somewhat different trajectory. In the early phase (2009-2017), reviewers appeared to prioritize functional availability, interface design, and content reliability, suggesting that these topics were associated with smaller negative rating deviations. In the later phase (2018-2025), greater PDs were associated with exam preparation and practice questions, error analysis and memorization, personalized learning services, and convenience, whereas greater NDs were associated with cost, nursing qualification exam, clinical teaching practice, teaching quality, and clinical decision support. These observed associations may be considered through self-determination theory as a possible interpretive lens [61]. From this perspective, stable technology and authoritative content could be perceived as supporting a trustworthy learning environment and a sense of competence, while access to clinical resources and guideline-search functions may be consistent with learner autonomy.

Cross-national comparisons indicated that app store reviewers in both countries showed a gradual shift in focus from core learning and technical functions toward exam preparation, question banks, and personalized learning. This observed pattern may be considered in relation to product iterations and the broader institutional contexts in which medical education apps are used. Country-specific patterns were also observed: among Chinese app store reviewers, rating deviations were more closely associated with service experience, cost, and exam-related and personalized features [62], whereas US app store reviewers showed stronger associations with the stability and professional reliability of core functions [63]. The shared attention to “personalized learning” among app store reviewers in both countries may suggest growing interest in adaptive learning models. Such systems could customize learning paths by adjusting content difficulty to learners’ ability levels, potentially helping manage cognitive load and learning efficiency. Future medical education apps could move beyond content aggregation toward intelligent tutoring systems that are responsive to learners’ real-time states, although such designs require further validation using learning process and outcome data. Developers could also consider optimizing exam support, personalized learning, and user service functions for different educational contexts, with the effects of such design changes requiring further evaluation.

From a functional classification perspective, the Kano model results indicated that review topics were primarily classified as essential or charm factors across country-phase models, whereas expectation factors showed no stable pattern. These classifications suggest that the associations between review topics and rating deviations varied across countries and temporal phases rather than following a single pattern. Essential factors differed across country-phase models, while exam support, learning management, and personalized features were more frequently classified as charm factors in particular models. This pattern suggests that functional innovation may contribute to competitive differentiation among medical education apps.

This study conducted a cross-national comparison of the observable development and rating-based satisfaction patterns of medical education apps in China and the United States. Within the 2025-observable analytic cohorts, data-derived temporal changes were identified around 2017 and 2021-2022. Although the timing of these changes was similar, the country-specific trajectories differed: the Chinese cohort showed rapid early expansion followed by slower growth while review activity continued to increase, whereas the US cohort showed declining app listings after 2017 and changing review activity after 2020. The similarity in timing may be considered in relation to shared contextual developments, including the global diffusion of mobile internet technology, institutional reforms, and shifts in the external environment, whereas the divergent trajectories may be considered in relation to differences in institutional environments, market maturity, and platform competition dynamics.

Regarding rating-based satisfaction patterns, this study identified phased shifts in app store reviewers’ priorities between China and the United States. In the Chinese sample, reviewers’ attention shifted from clinical utility and interactive learning in the early phase toward exam preparation, question banks, personalized learning, user service, and cross-screen compatibility in the later phase, while the directions of topic-rating-deviation associations varied across topics and phases. In the US sample, reviewers’ attention shifted from functional availability, interface design, and content reliability in the early phase toward cost, exam preparation, and practice questions, and convenience-related features in the later phase. Cross-national comparisons suggest that while functional evolution paths differed between countries, reviewers in both samples showed increasing attention to personalized learning, which may be consistent with growing interest in adaptive learning models in educational technology.

This study holds significant theoretical and practical implications. Theoretically, this study extends the cross-national evidence on mobile learning and rating-based satisfaction and situates the observed review patterns in relation to the TAM and educational evaluation theories. Educational systems and cultural contexts may provide useful interpretive context for the observed differences in user review patterns. Practically, the findings identify potential design priorities for medical education app developers, including core functional stability,



XSL•FO
RenderX

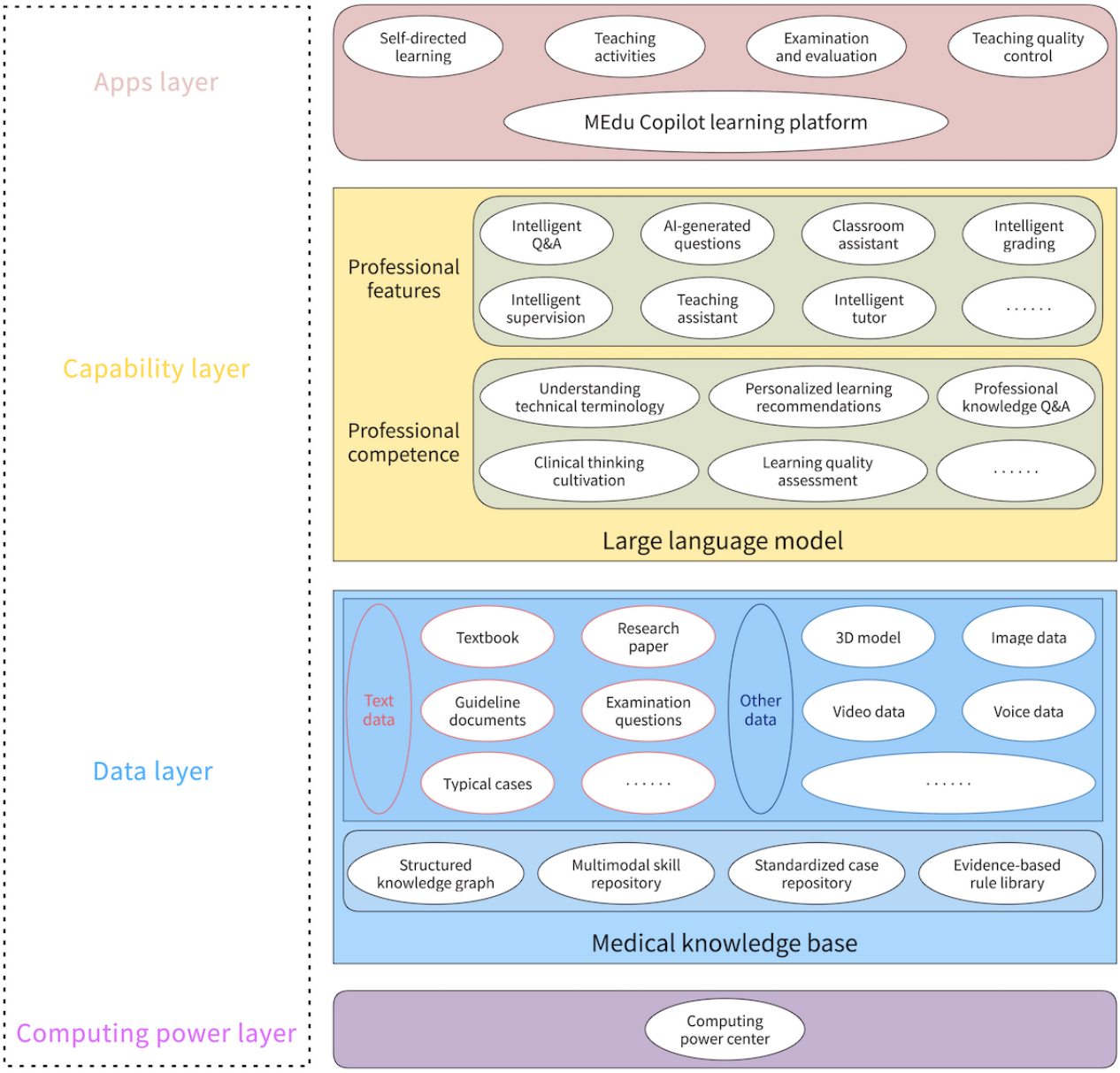
professional reliability, exam preparation support, personalized learning, and user service features. For educational administrators and policymakers, the observed patterns may inform the planning and evaluation of digital teaching resources in different institutional and educational contexts.

## Design Implications for AI-Supported Medical Education Platforms

Before translating these findings into a design framework, it is important to interpret the observed cross-national differences within their institutional and educational contexts. In China, the stronger association of rating deviations with exam preparation, question banks, cost, and user service may be considered in relation to a standardized and assessment-oriented medical training environment, in which affordable and efficient exam-related support may be particularly salient. In the United States, the stronger associations with technical stability, functional availability, content reliability, and clinical decision support may be considered in relation to a credentialing and self-directed learning context, in which stable access to trustworthy resources may be particularly salient across clinical and educational settings.

On the basis of these empirically observed and contextually interpreted review and rating-deviation patterns, we proposed a conceptual design framework, termed "MEdu Copilot," to illustrate potential design implications for future AI-supported medical education platforms (Figure 5). This framework conceptually translates the user review patterns and topic-rating deviation associations identified in China and the United States into potential design strategies for intelligent recommendation. Building on these findings, it illustrates a goal-oriented design approach that could inform the development of intelligent assistance for knowledge acquisition, clinical reasoning, and professional competency development.

**Figure 5.** MEdu Copilot conceptual framework.






XSL•FO
RenderX

Figure 5 presents a proposed layered architecture for a future AI-supported medical education platform. The computing layer represents the computational infrastructure; the data layer represents a "Medical Knowledge Base" comprising multimodal and structured resources, such as textbooks, guidelines, clinical cases, 3D models, videos, images, and examination questions. The capability layer centers on a large language model that could support functions such as terminology assistance, clinical reasoning support, intelligent question answering, classroom assistance, and automated examination generation and grading. The application layer maps these potential capabilities to educational scenarios, including self-directed learning, instructional activities, assessment, and quality monitoring. The framework illustrates possible relationships among infrastructure, data, AI-supported capabilities, and educational apps; it does not represent a validated or implemented platform.

This conceptual framework follows a user-centered, goal-oriented approach that connects the user and technical design space through an iterative feedback mechanism. In a future implementation, user interaction data and learning outcomes could inform the adaptive refinement of learning resources, instructional support tools, and assessment strategies. Privacy protection, trust, and cultural sensitivity are proposed as guiding design principles that may help align future platform development with expectations for accountability and inclusivity in digital health interventions. Overall, the framework illustrates a potential pathway for the responsible integration of AI-supported capabilities into medical education.

## Limitations

This study has several limitations. First, the analytic sample was defined at the September 2025 search cutoff. Although broad bilingual keywords, related-app screening where available, and title-description screening were used to improve app identification, apps discontinued earlier, no longer searchable, lacking retrievable metadata, or using atypical titles or descriptions may still have been missed. Therefore, supply-side trends should be interpreted as patterns within the 2025-observable analytic cohort, rather than as a complete historical census of medical education apps. Because the temporal segmentation was derived from cumulative counts within the 2025-observable analytic cohort, phase-based demand-side comparisons should be interpreted as exploratory comparisons across data-derived temporal segments rather than as definitive comparisons across complete historical market stages.

Second, app store reviews provide a self-selected and rating-based source of user feedback. These reviews may underrepresent "silent users" who used apps but did not leave ratings or comments. They may also overrepresent users with particularly positive experiences or strong technical frustrations. Because reviewer-level demographic, professional, training-stage, and app-use metadata were unavailable, we could not directly adjust for demographic differences across app stores or assess subgroup-specific satisfaction patterns. Accordingly, rating-based satisfaction metrics and topic patterns should be interpreted as rating- and review-based indicators among app store reviewers, rather than population-level estimates for all medical education app users or learners. In addition, all review-level analyses were conducted using the rating-text-concordant analytic subset. Although this concordance criterion was intended to improve consistency between numerical ratings and review text, it may also have excluded meaningful mixed, sarcastic, or feature-specific evaluations. Accordingly, the possibility of selection bias should be considered when interpreting the review-level findings.

Third, topic modeling and cross-linguistic comparison involve interpretive uncertainty. Although Chinese and English reviews were modeled separately and topic labels were checked through bilingual semantic validation, differences in sentiment expression and topic wording may still affect comparability across countries. The AI-assisted topic labels should therefore be understood as structured representations of review themes, rather than definitive categories of user needs.

Finally, the observational design limits causal interpretation. Reviews reflect subjective perceptions rather than actual learning outcomes, and the data do not contain individual-level information on educational status, examination participation, institutional affiliation, or cultural orientation. Interpretations related to national educational systems, institutional reforms, cultural learning behaviors, or learner psychology should therefore be viewed as contextual explanations for observed associations rather than causal inferences. Future studies should integrate learning logs, examination outcomes, user interviews, subgroup analyses, and longitudinal or experimental designs to evaluate how app features influence learning processes and professional competency development.

## Conclusions

In summary, this longitudinal app store analysis identified phased evolution and country-specific rating-based satisfaction patterns among medical education apps in China and the United States. The findings suggest that app store reviewers' concerns shifted from basic functionality, clinical practicality, and technical stability toward exam preparation, cost, and personalized learning support. More broadly, this study highlights the need to understand medical education apps not only as digital products but also as sociotechnical educational tools situated within platform ecosystems, institutional contexts, and learner needs. The MEdu Copilot framework should therefore be interpreted as a conceptual synthesis of the empirical findings and contextual policy interpretations, rather than as a validated platform, and future studies should evaluate such designs using learning process and outcome data.

## Acknowledgments

The authors thank all contributors who supported data collection, analysis, figure preparation, and manuscript revision.



XSL•FO
RenderX

## Funding

This work was supported by the National Natural Science Foundation of China (grant number 72574006), the Noncommunicable Chronic Diseases-National Science and Technology Major Project (grant numbers 2023ZD0509800 and 2023ZD0509804), the Key Project of Higher Education Research of Zhejiang Higher Education Association in 2025 (grant number KT2025001), the Key Project of the 2024 Higher Education Science Research Plan of the Chinese Society of Higher Education (grant number 24BJ0311), and the 2025 Medical Education Research Project of the Medical Education Branch of the Chinese Medical Association and the National Medical Education Development Center (grant number 2025B349).

## Data Availability

The deidentified and aggregated datasets supporting the main analyses of this study are available from the corresponding author on reasonable request, where permitted by applicable data-use restrictions. The analysis code and model specifications are publicly available in a GitHub repository [64]. App metadata and user review data were obtained from 10 major app stores through the Qimai Apps Data platform [65] and web scraping technology. Because the raw app-store review data may be subject to platform terms of service and may contain user-generated text, the full raw review dataset is not publicly redistributed. App-store searches were conducted in September 2025, and user review data were collected as of October 2025.

## Generative AI Use

DeepSeek-R1 was used to assist with semantic labeling of latent Dirichlet allocation–generated topics by generating candidate topic names and brief descriptions from the top 30 high-probability keywords for each topic. All AI-assisted topic labels and descriptions were critically reviewed, edited, and verified by the authors through consensus-based semantic review and structured expert validation. The authors take full responsibility for the final topic labels, analyses, interpretations, and manuscript content. Generative AI was not used for data collection, review preprocessing, topic modeling, statistical analysis, result interpretation, or generation of scientific conclusions. It was used to assist with language editing and polishing of the manuscript, and all AI-assisted revisions were critically reviewed and approved by the authors.

## Authors' Contributions

J Lei and J Liang conceived the study. JM and J Liang independently screened and evaluated the apps according to the predefined eligibility criteria, and YC adjudicated disagreements. ZL and JZ collected developer and user feedback. YH and XL performed the data analysis. YL and JW supervised and validated the qualitative analysis using large language models. TS and WD completed the manuscript translation. QY, LY, and HP were responsible for figure preparation and design. JM drafted the manuscript. J Liang and QY critically revised the manuscript, and PX reviewed it. QY, J Liang, and J Lei are co-corresponding authors of this work. All authors reviewed and approved the final manuscript.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Supplementary methods, tables, figures, and results on medical education app store reviews and Kano classification in China and the United States.
[DOCX File , 7069 KB-Multimedia Appendix 1]

## References

1. Cernega A, Nicolescu DN, Mele canu Imre M, Ripszky Totan A, Arsene AL, erban RS, et al. Volatility, uncertainty, complexity, and ambiguity (VUCA) in healthcare. Healthcare (Basel). 2024;12(7):773. [FREE Full text] [doi: 10.3390/healthcare12070773] [Medline: 38610195]
2. Stojan J, Haas M, Thammasitboon S, Lander L, Evans S, Pawlik C, et al. Online learning developments in undergraduate medical education in response to the COVID-19 pandemic: a BEME systematic review: BEME Guide No. 69. Med Teach. 2022;44(2):109-129. [doi: 10.1080/0142159x.2021.1992373]
3. Alrashed FA, Ahmad T, Almurdi MM, Alderaa AA, Alhammad SA, Serajuddin M, et al. Incorporating technology adoption in medical education: a qualitative study of medical students' perspectives. Adv Med Educ Pract. 2024;15:615-625. [FREE Full text] [doi: 10.2147/AMEP.S464555] [Medline: 38975614]
4. Alsharif MA, Bahbah AA, Ghonaim MM, Omran JA, Rihan MM, Zayed MA, et al. Usage and attitude of medical students towards mobile medical applications during and after COVID-19 lockdown: repeated cross-sectional study. BMC Med Educ. 2024;24(1):233. [FREE Full text] [doi: 10.1186/s12909-024-05216-5] [Medline: 38438883]



XSL•FO
RenderX

5. Khan HU, Ali Y, Azeem Akbar M, Khan F. A comprehensive survey on exploring and analyzing COVID-19 mobile apps: meta and exploratory analysis. Heliyon. 2024;10(15):e35137. [FREE Full text] [doi: 10.1016/j.heliyon.2024.e35137] [Medline: 39170132]
6. Prober CG, Khan S. Medical education reimagined: a call to action. Acad Med. 2013;88(10):1407-1410. [doi: 10.1097/ACM.0b013e3182a368bd] [Medline: 23969367]
7. Zhou Y, Yang Y, Liu L, Zeng Z. Effectiveness of mobile learning in medical education: a systematic review. Nan Fang Yi Ke Da Xue Xue Bao. 2018;38(11):1395-1400. [doi: 10.12122/j.issn.1673-4254.2018.11.20] [Medline: 30514692]
8. Dornan T, Boshuizen H, King N, Scherpbier A. Experience-based learning: a model linking the processes and outcomes of medical students' workplace learning. Med Educ. 2007;41(1):84-91. [doi: 10.1111/j.1365-2929.2006.02652.x] [Medline: 17209896]
9. Yeung AWK, Parvanov ED, Hribersek M, Eibensteiner F, Klager E, Kletecka-Pulker M, et al. Digital teaching in medical education: scientific literature landscape review. JMIR Med Educ. 2022;8(1):e32747. [FREE Full text] [doi: 10.2196/32747] [Medline: 35138260]
10. Chandran VP, Balakrishnan A, Rashid M, Pai Kulyadi G, Khan S, Devi ES, et al. Mobile applications in medical education: a systematic review and meta-analysis. PLoS One. 2022;17(3):e0265927. [FREE Full text] [doi: 10.1371/journal.pone.0265927] [Medline: 35324994]
11. Sadler J, Wright J, Vincent T, Kurka T, Howlett D. What is the impact of apps in medical education? A study of CAPSULE, a case-based learning app. BMJ Simul Technol Enhanc Learn. 2021;7(5):293-296. [doi: 10.1136/bmjstel-2020-000593] [Medline: 35515743]
12. Samuel A, Durning S. From crisis to opportunity: reinventing medical education after COVID-19. Perspect Med Educ. 2025;14(1):371-382. [FREE Full text] [doi: 10.5334/pme.1672] [Medline: 40586112]
13. Zhang Z, Ji T. Application and progress of digital techniques in head and neck oncology reconstructive surgery. Holist Integ Oncol. 2025;4(1):16. [doi: 10.1007/s44178-024-00137-1]
14. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res. 2019;21(8):e12967. [FREE Full text] [doi: 10.2196/12967] [Medline: 31456579]
15. Almulhem JA, Aldekhyyel RN, Binkheder S. Evaluation of mobile apps used among medical students for learning and education: a mixed-method concurrent triangulation approach. Appl Clin Inform. 2024;15(4):717-726. [FREE Full text] [doi: 10.1055/s-0044-1788332] [Medline: 39231539]
16. Xu Q, Hou X, Xiao T, Zhao W. Factors affecting medical students' continuance intention to use mobile health applications. J Multidiscip Healthc. 2022;15:471-484. [FREE Full text] [doi: 10.2147/JMDH.S327347] [Medline: 35300179]
17. Johnson SG, Potrebny T, Larun L, Ciliska D, Olsen NR. Usability methods and attributes reported in usability studies of mobile apps for health care education: scoping review. JMIR Med Educ. 2022;8(2):e38259. [FREE Full text] [doi: 10.2196/38259] [Medline: 35767323]
18. Walsh K. Mobile learning in medical education: review. Ethiop J Health Sci. 2015;25(4):363-366. [FREE Full text] [doi: 10.4314/ejhs.v25i4.10] [Medline: 26949301]
19. Barteit S, Guzek D, Jahn A, Bärnighausen T, Jorge MM, Neuhann F. Evaluation of e-learning for medical education in low- and middle-income countries: a systematic review. Comput Educ. 2020;145:103726. [FREE Full text] [doi: 10.1016/j.compedu.2019.103726] [Medline: 32565611]
20. Gimpel H, Manner-Romberg T, Schmied F, Winkler TJ. Understanding the evaluation of mHealth app features based on a cross-country Kano analysis. Electron Mark. 2021;31(4):765-794. [FREE Full text] [doi: 10.1007/s12525-020-00455-y] [Medline: 35602116]
21. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly. 1989;13(3):319-340. [doi: 10.2307/249008]
22. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly. 2003;27(3):425-478. [doi: 10.2307/30036540]
23. Jan J, Alshare KA, Lane PL. Hofstede's cultural dimensions in technology acceptance models: a meta-analysis. Univ Access Inf Soc. 2024;23(2):717-741. [doi: 10.1007/s10209-022-00930-7]
24. Liu X, Feng J, Liu C, Chu R, Lv M, Zhong N, et al. Medical education systems in China: development, status, and evaluation. Acad Med. 2023;98(1):43-49. [FREE Full text] [doi: 10.1097/ACM.0000000000004919] [Medline: 35947483]
25. Patel OU, Haynes WK, Burge KG, Yadav IS, Peterson T, Camino A, et al. Results of a national survey of program directors' perspectives on a pass/fail US Medical Licensing Examination Step 1. JAMA Netw Open. 2022;5(6):e2219212. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.19212] [Medline: 35763299]
26. Singh K, Sarkar S, Gaur U, Gupta S, Adams OP, Sa B, et al. Smartphones and educational apps use among medical students of a smart university campus. Front Commun. 2021;6:649102. [doi: 10.3389/fcomm.2021.649102]
27. Elhassan SE, Sajid MR, Syed AM, Fathima SA, Khan BS, Tamim H. Assessing familiarity, usage patterns, and attitudes of medical students toward ChatGPT and other chat-based AI apps in medical education: cross-sectional questionnaire study. JMIR Med Educ. 2025;11:e63065. [FREE Full text] [doi: 10.2196/63065] [Medline: 39883912]




XSL•FO
RenderX

28. Gladman T, Tylee G, Gallagher S, Mair J, Rennie SC, Grainger R. A tool for rating the value of health education mobile apps to enhance student learning (MARuL): development and usability study. JMIR Mhealth Uhealth. 2020;8(7):e18015. [FREE Full text] [doi: 10.2196/18015] [Medline: 32735228]
29. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth. 2015;3(1):e27. [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
30. Owrangi S, Zarifsanaiey N, Rezaei M. Mobile apps to enhance student learning in medical education: a systematic search in app stores and evaluation using the mobile app rating scale. Comput Biol Med. 2025;196(Pt B):110740. [doi: 10.1016/j.compbiomed.2025.110740] [Medline: 40669285]
31. Gladman T, Tylee G, Gallagher S, Mair J, Grainger R. Measuring the quality of clinical skills mobile apps for student learning: systematic search, analysis, and comparison of two measurement scales. JMIR Mhealth Uhealth. 2021;9(4):e25377. [FREE Full text] [doi: 10.2196/25377] [Medline: 33890859]
32. Briz-Ponce L, Juanes-Méndez JA, García-Peñalvo FJ. Description of a session to assess the efectiveness of apps in medical education. ResearchGate. URL: https://www.researchgate.net/publication/280114331_Description_of_a_session_to_assess_the_efectiveness_of_apps_in_medical_education [accessed 2026-07-16]
33. S S, Vedam VAV, Ghanta MK. An assessment of utility of educational apps in medical education among undergraduate medical students - a cross-sectional survey. Natl J Physiol Pharm Pharmacol. 2024;14(1):182-187. [doi: 10.5455/njppp.2023.13.06317202322062023]
34. Darras KE, van Merriënboer JJG, Toom M, Roberson ND, de Bruin ABH, Nicolaou S, et al. Developing the evidence base for M-learning in undergraduate radiology education: identifying learner preferences for mobile apps. Can Assoc Radiol J. 2019;70(3):320-326. [FREE Full text] [doi: 10.1016/j.carj.2019.03.007] [Medline: 31300315]
35. Zhao S, Cao Y, Cao H, Liu K, Lv X, Zhang J, et al. Chinese version of the mHealth app usability questionnaire: cross-cultural adaptation and validation. Front Psychol. 2022;13:813309. [FREE Full text] [doi: 10.3389/fpsyg.2022.813309] [Medline: 35185732]
36. Jiao X, Jiang L, Zhao M, Ma J, Li Y, Shen T, et al. User satisfaction with pregnancy management apps in Mainland China: user-generated content analysis and text mining study. J Med Internet Res. 2025;27:e78828. [FREE Full text] [doi: 10.2196/78828] [Medline: 40839869]
37. Liew MS, Zhang J, See J, Ong YL. Usability challenges for health and wellness mobile apps: mixed-methods study among mHealth experts and consumers. JMIR Mhealth Uhealth. 2019;7(1):e12160. [FREE Full text] [doi: 10.2196/12160] [Medline: 30698528]
38. Masters K, Ellaway RH, Topps D, Archibald D, Hogue RJ. Mobile technologies in medical education: AMEE Guide No. 105. Med Teach. 2016;38(6):537-549. [doi: 10.3109/0142159x.2016.1141190]
39. Yang F, Zheng J, Li W, Liu S, Zhou Y, Lu S. Digitalizing caregiving for what? Role dynamics of caregivers in nursing homes implementing smart technologies. J Technol Hum Serv. 2025:1-28. [doi: 10.1080/15228835.2025.2585409]
40. Toward AI-enhanced medical education: cross-national evolution and user satisfaction in mobile learning apps. OSF Registries. URL: https://osf.io/hy2cj/ [accessed 2026-07-15]
41. Su Z, Figueiredo MC, Jo J, Zheng K, Chen Y. Analyzing description, user understanding and expectations of AI in mobile health applications. AMIA Annu Symp Proc. 2020;2020:1170-1179. [FREE Full text] [Medline: 33936493]
42. Martens D, Maalej W. Towards understanding and detecting fake reviews in app stores. Empir Software Eng. 2019;24(6):3316-3355. [doi: 10.1007/s10664-019-09706-9]
43. Baidu AI Cloud. Sentiment analysis. Baidu AI Cloud Documentation. URL: https://cloud.baidu.com/doc/NLP/s/zk6z52hds [accessed 2025-10-17]
44. Valdivia A, Luzon MV, Herrera F. Sentiment analysis in TripAdvisor. IEEE Intell Syst. 2017;32(4):72-77. [doi: 10.1109/mis.2017.3121555]
45. Valdivia A, Hrabova E, Chaturvedi I, Luzón MV, Troiano L, Cambria E, et al. Inconsistencies on TripAdvisor reviews: a unified index between users and sentiment analysis methods. Neurocomputing. 2019;353:3-16. [doi: 10.1016/j.neucom.2018.09.096]
46. He Y, Zhu W, Wang T, Chen H, Xin J, Liu Y, et al. Mining user reviews from hypertension management mobile health apps to explore factors influencing user satisfaction and their asymmetry: comparative study. JMIR Mhealth Uhealth. 2024;12:e55199. [FREE Full text] [doi: 10.2196/55199] [Medline: 38547475]
47. Patel VR, Liu M, Jena AB. Public interest in an AI-enabled clinical decision support tool. JAMA Netw Open. 2025;8(11):e2544672. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.44672] [Medline: 41264272]
48. Sun J. Jieba: Chinese words segmentation utilities. GitHub. URL: https://github.com/fxsjy/jieba [accessed 2026-07-15]
49. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993-1022. [FREE Full text]
50. Arora N, Chakraborty I, Nishimura Y. AI–human hybrids for marketing research: leveraging large language models (LLMs) as collaborators. J Mark. 2025;89(2):43-70. [doi: 10.1177/00222429241276529]
51. Liu S, Zheng J, Yang F, Li W, Zhou Y, Lu S, et al. Internet altruistic behaviors and affective well-being among family caregivers: a natural language processing approach. J Gerontol B Psychol Sci Soc Sci. 2026;81(1):gbaf230. [doi: 10.1093/geronb/gbaf230] [Medline: 41224706]




XSL•FO
RenderX

52. Wang T, Fan L, Zheng X, Wang W, Liang J, An K, et al. The impact of gamification-induced users' feelings on the continued use of mHealth apps: a structural equation model with the self-determination theory approach. J Med Internet Res. 2021;23(8):e24546. [FREE Full text] [doi: 10.2196/24546] [Medline: 34387550]
53. Wang T, Zheng X, Liang J, An K, He Y, Nuo M, et al. Use of machine learning to mine user-generated content from mobile health apps for weight loss to assess factors correlated with user satisfaction. JAMA Netw Open. 2022;5(5):e2215014. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.15014] [Medline: 35639374]
54. Nuo M, Zheng S, Wen Q, Fang H, Wang T, Liang J, et al. Mining the influencing factors and their asymmetrical effects of mHealth sleep app user satisfaction from real-world user-generated reviews: content analysis and topic modeling. J Med Internet Res. 2023;25:e42856. [FREE Full text] [doi: 10.2196/42856] [Medline: 36719730]
55. Wang T, Wang W, Feng J, Fan X, Guo J, Lei J. A novel user-generated content-driven and Kano model focused framework to explore the impact mechanism of continuance intention to use mobile APPs. Comput Hum Behav. 2024;157:108252. [doi: 10.1016/j.chb.2024.108252]
56. Barrett B. The iOS 11 32-bit appocalypse isn't actually that bad. Wired. URL: https://www.wired.com/story/the-ios-11-32-bit-appocalypse-isnt-actually-that-bad/ [accessed 2026-07-15]
57. 64-bit and iOS 8 requirements for app updates. Apple Developer News. URL: https://developer.apple.com/news/?id=12172014b [accessed 2026-07-15]
58. Lee JWY, Tan JY, Bello F. Technology acceptance model in medical education: systematic review. JMIR Med Educ. 2025;11:e67873. [FREE Full text] [doi: 10.2196/67873] [Medline: 40668663]
59. The general office of the state council issued the opinions on deepening medical-education collaboration and further promoting the reform and development of medical education. National Health Commission of the People's Republic of China. URL: https://www.nhc.gov.cn/wjw/gwywj/201707/8620cd5834ec47bd966f2d0acb67d4d3.shtml [accessed 2026-07-15]
60. Izquierdo-Condoy JS, Arias-Intriago M, Nati-Castillo HA, Gollini-Mihalopoulos R, Cardozo-Espínola CD, Loaiza-Guevara V, et al. Exploring smartphone use and its applicability in academic training of medical students in latin America: a multicenter cross-sectional study. BMC Med Educ. 2024;24(1):1401. [FREE Full text] [doi: 10.1186/s12909-024-06334-w] [Medline: 39616324]
61. Howard JL, Bradshaw EL, Ryan RM. Future directions for self-determination theory: introduction to the special issue. Motiv Emot. 2026;50(1):1-12. [doi: 10.1007/s11031-025-10160-y]
62. Shen X, Wang Y, Zhao D, Li M. Medical licensing examination-based reform of medical microbiology teaching. Med Educ. 2023;57(11):1122. [doi: 10.1111/medu.15206] [Medline: 37705373]
63. Liu L, Chachad N, Tadjalli A, Rajput V. Decoupling the United States Medical Licensing Examinations (USMLEs) from the medical curriculum to promote student well-being and professional identity development. Cureus. 2025;17(5):e83335. [doi: 10.7759/cureus.83335] [Medline: 40458346]
64. The source code for the medical education app analysis. GitHub. URL: https://github.com/junhao-ma/medical-education-app-analysis [accessed 2026-09-28]
65. Qimai apps data platform. Qimai. URL: https://www.qimai.cn [accessed 2025-10-19]


## Abbreviations

**AAPC:** average annual percentage change
**APC:** annual percentage change
**BIC:** Bayesian information criterion
**LDA:** latent Dirichlet allocation
**MARS:** Mobile App Rating Scale
**MARuL:** Mobile App Rubric for Learning
**ND:** negative rating deviation
**NLTK:** Natural Language Toolkit
**PD:** positive rating deviation
**PBL:** problem-based learning
**RQ:** research question
**TAM:** technology acceptance model
**USMLE:** United States Medical Licensing Examination
**VIF:** variance inflation factor



XSL•FO
RenderX

*Edited by T de Azevedo Cardoso; submitted 20.Jul.2026; peer-reviewed by F Yang, A Wahbeh, F Cui; comments to author 19.Aug.2026; revised version received 29.Sep.2026; accepted 05.Oct.2026; published 03.Nov.2026*

*Please cite as:*
*Ma J, Shen T, Xiang P, Li Y, Chen Y, Lu Z, Zeng J, Dai W, Pan H, Yang L, He Y, Li X, Wu J, Yu Q, Liang J, Lei J*
*Mobile Medical Education Apps as Digital Health Information Resources in China and the United States: Longitudinal App Store Metadata Analysis and Topic Modeling of User Review Narratives*
*JMIR Med Educ 2026;12:e107560*
*URL: https://mededu.jmir.org/2026/1/e107560*
*doi: 10.2196/107560*
*PMID:*

©Junhao Ma, Tian Shen, Peng Xiang, Yingjun Li, Yuanshuai Chen, Zhengyang Lu, Jingkai Zeng, Wei Dai, Hanwen Pan, Liping Yang, Yunfan He, Xingxi Li, Joyce Wu, Qi Yu, Jun Liang, Jianbo Lei. Originally published in JMIR Medical Education (https://mededu.jmir.org), 03.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 JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.



XSL•FO
RenderX