Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

How Physician Reviews Affect Online Consultation Demand: An Innovative Small Language Model with Fine-Tuning

Bin Zhang, Haijing Hao, Yongcheng Zhan, Jiang Wu

To cite this article:
Bin Zhang, Haijing Hao, Yongcheng Zhan, Jiang Wu (2026) How Physician Reviews Affect Online Consultation Demand: An Innovative Small Language Model with Fine-Tuning. Information Systems Research

Published online in Articles in Advance 08 May 2026

. https://doi.org/10.1287/isre.2024.1183

https://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Articles in Advance, pp. 1–17
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

How Physician Reviews Affect Online Consultation Demand: An Innovative Small Language Model with Fine-Tuning

Bin Zhang,[a,*] Haijing Hao,[b] Yongcheng Zhan,[c] Jiang Wu[d,*]

[a] Department of Information and Operations Management, Texas A&M University, College Station, Texas 77843; [b] Computer Information Systems Department, Bentley University, Waltham, Massachusetts 02452; [c] Orfalea College of Business, California Polytechnic State University, San Luis Obispo, California 93407; [d] School of Information Management, Wuhan University, Wuhan, Hubei 430072, China
*Corresponding authors

Contact: bzhang@mays.tamu.edu, https://orcid.org/0000-0003-0675-2222 (BZ); hhao@bentley.edu, https://orcid.org/0000-0003-3071-2538 (HH); yozhan@calpoly.edu, https://orcid.org/0000-0002-5029-0961 (YZ); jiangw@whu.edu.cn, https://orcid.org/0000-0002-3342-9757 (JW)

Received: May 17, 2024
Revised: January 9, 2025; November 1, 2025
Accepted: March 26, 2026
Published Online in Articles in Advance: May 8, 2026

https://doi.org/10.1287/isre.2024.1183

Abstract. Large language models (LLMs) are advancing rapidly but remain expensive to train and deploy, especially in specialized domains such as healthcare. To improve efficiency and reduce costs, we develop a small language model (SLM) that can be fine-tuned for effective sentiment analysis in this context. Our SLM operationalizes a quality evaluation framework for healthcare service to extract providers' service quality scores from online physician reviews, and then we estimate the relationship between the quality scores and the online consultation demand. Using a panel data set from one of China's largest online physician review and consultation platforms, we show that higher service-quality scores are associated with greater consultation demand. Our research contributes in three ways. Methodologically, we develop an SLM (Doc-BERT) tailored to the healthcare context and demonstrate that a task-aligned, domain-tuned SLM can outperform a wide range of existing NLP approaches, including recent general LLMs, for multidimensional sentiment analysis of physician reviews. Theoretically, we adapt and refine a healthcare service quality assessment framework at provider level that has been widely used in healthcare practice but rarely applied in information systems research. Empirically, we identify specific dimensions of service quality that most strongly predict online consultation demand, offering actionable guidance for healthcare professionals and administrators seeking to optimize services and increase uptake.

History: Indranil Bardhan, Senior Editor; Atanu Lahiri, Associate Editor.

Funding: This work was supported by the Wuhan University, National Natural Science Foundation of China [Grant 72232006]. The views expressed in this article are those of the authors and do not necessarily reflect the views of the funding agency.

Supplemental Material: The online appendices are available at https://doi.org/10.1287/isre.2024.1183.

Keywords: healthcare providers' service quality • online consultation demand • natural language processing • deep learning • sentiment analysis • small language model • large language model • online physician review

1. Introduction

Large language models (LLMs) have been advancing rapidly since the introduction of the transformer architecture (Vaswani et al. 2017); however, they remain costly to train and deploy because they contain billions to trillions of parameters as general-purpose sequence models (Cottier et al. 2024). To address this challenge, we develop a small language model (SLM) with a fine-tuning feature that is customized for healthcare applications, demonstrating the SLM's cost and time efficiency when applied to a specific task within a defined domain. We design our SLM by combining the strengths of Doc2Vec and bidirectional encoder representations from transformers (BERT) to examine how online reviews reflect physicians' service quality, and how this, in turn, influences the demand for online consultations. The algorithm is designed to extract service quality scores from online physician reviews using a health service quality framework established by health authorities to evaluate healthcare provider performance (AHRQ 2019), which avoids random text mining based on probabilistic methods without a theory supporting guideline as previous studies have done (Wan et al. 2021, Xu et al. 2021). We then analyze how these quality

scores affect patients' decisions to make appointments for online consultation.

We choose online consultation as our study instance because it is in one of the highest demands among all healthcare services. Online consultation allows physicians to deliver healthcare service through electronic channels (Dorsey and Topol 2016). This service has experienced an astonishing 154% surge in April 2020 compared with previous year (Koonin et al. 2020). The global telehealth market is projected to experience substantial growth, with a compound annual growth rate of 15.8% (Grand View Research 2024).

However, patients often face information asymmetry, making it challenging to assess the quality of a physician's service when deciding for online consultation. Online physician reviews have gained prominence for patients to share their personal experiences since the advent of Web 2.0, which can reduce the information asymmetry, and have been prevalent in many countries (Gao et al. 2012, Emmert and Meier 2013, Hao 2015). With the rapid advancements in natural language processing (NLP) technology, researchers have begun to utilize physician reviews to gain rich and previously inaccessible insights into patient experiences (Hao et al. 2017) and to examine how patients perceive healthcare service quality based on online ratings (Jiang et al. 2024). Research has found that physician reviews can effectively assess a physician's service quality (Gao et al. 2015), can impact patients' choices of primary care providers (Yaraghi et al. 2018) and can influence *offline* consultation demand (Xu et al. 2021) or online consultation demand (Wan et al. 2021). However, existing research lacks a systematic approach to utilize the latest NLP technology within theory-supported healthcare service quality evaluation frameworks to extract meaningful healthcare service quality assessments from online reviews and examine the service quality's impact on online consultations. First, prior studies rely on aspect-based sentiment analysis (ABSA) methods that mine service reviews without applying a theoretical or clinically grounded framework (Wan et al. 2021, Xu et al. 2021). As a result, they fail to capture the clinically validated dimensions of healthcare quality that regulatory authorities and medical professionals consider essential for healthcare effectiveness. Second, prior studies' sentiment analysis has not leveraged the latest NLP technologies tailored specifically for healthcare contexts (Wan et al. 2021, Xu et al. 2021). To overcome these challenges, we adopt the widely used health service evaluation framework developed by healthcare authorities as our health quality score mining guideline and develop an innovative SLM that can be fine-tuned with the customized healthcare service quality framework to analyze physician reviews for quality purpose more effectively.

Our study makes several contributions. Methodologically, we develop an SLM (Doc-BERT) tailored to the healthcare context and demonstrate that a task-aligned, domain-tuned SLM can outperform a wide range of existing NLP approaches, including recent general LLMs, for multidimensional sentiment analysis of physician reviews. This performance advantage reflects that SLM has the advantage of full domain adaptation and computational cost compared with LLMs in this context. Theoretically, we adopt and refine a quality evaluation framework for healthcare service at the provider level, ensuring that our SLM purposefully analyzes online reviews with a focus on providers' service quality rather than random text mining. To the best of our knowledge, the present study is among the first in the information systems field to employ this framework. Empirically, our findings demonstrate that physician reviews have a significant impact on the demand for online consultation, which offers practical implications for online healthcare providers and healthcare organizations seeking to enhance service quality and, consequently, increase online consultation demand.

2. Literature Review

2.1. Online Consultations and Physician Reviews

The onset of the COVID-19 pandemic has intensified discussions on telehealth or online consultation (Michel et al. 2023, Pan et al. 2023). Fan et al. (2023) suggested that factors like online reviews and physician rank positively impacted patient demand for these services. However, Saifee et al. (2020) found that there was no clear relationship between physician reviews and the readmission rate or emergency room visits, the common clinical quality measures of physician healthcare service. Online physician review platforms offer unparalleled convenience for patients in providing feedback of their experience because the forums are accessible to anyone with Internet access. Online consultations are especially beneficial for patients who are geographically distant and lack local healthcare resources, making such services their primary source of care. The online reviews also allow prospective patients to benefit from the experiences of previous patients, an opportunity not available with traditional patient surveys administrated by healthcare organizations internally (Schlesinger et al. 2015, Lee 2017). Consequently, prospective patients can make more informed decisions about which physicians to see based on these public reviews. Research has found that both the quality and quantity of online physician reviews impact patients' perceptions of physicians (Grabner-Kräuter and Waiguny 2015). Existing literature has begun to explore the impact of physician reviews on patient choice of healthcare providers (Vennik et al. 2014). However,

prior research has predominantly utilized basic techniques, such as naïve Bayes, support vector machine, and latent Dirichlet allocation (LDA) to mine the content of textual online reviews using probabilistic methods, or aspect-based sentiment analysis, without a specific framework for healthcare service quality (Hao et al. 2017, Wan et al. 2021, Xu et al. 2021).

It has been widely accepted that healthcare services should be patient centered, with patient satisfaction being at the core (Epstein and Street 2011, NEJM Catalyst 2017), and patients evaluated providers' performance based on technical and interpersonal qualities (Fung et al. 2005). Traditionally, the evaluation of a physician's interpersonal skills relied on paper-based patient surveys administered by hospitals (Carr-Hill 1992, Grøndahl et al. 2013, Al-Abri and Al-Balushi 2014). However, research acknowledged the inherent limitations of paper-based surveys, including small sample sizes and low response rates, which can undermine the statistical validity of the surveys (Scaletta 2015). There are needs to explore alternative methods, such as online reviews, to evaluate the health service quality (Evans et al. 2020). Siegrist (2013) also has highlighted the value of analyzing textual comments in patient surveys, as this can uncover meaningful information.

2.2. NLP for Multidimensional Healthcare Sentiment Analysis

Using NLP to perform sentiment analysis can play a crucial role to determine the emotional tone or sentiment behind a paragraph of text (Liu 2022). This task is particularly useful for analyzing large volumes of online reviews or social media content, providing underlying information from public opinions or customer satisfaction (Pang and Lee 2008, Ko et al. 2019).

Evaluating patient opinions within a healthcare service evaluation framework requires meticulous analysis because sentiment can vary significantly across different aspects of service quality (Singh et al. 2020). To address these challenges, researchers have developed various language modeling techniques (Mikolov et al. 2013). Among these, BERT, as a SLM with approximately 110 million parameters, compared with LLMs with 175 billion parameters in GPT-3, excels at capturing complex language structures, enabling accurate, and multidimensional sentiment analysis across large data sets (Devlin et al. 2019). BERT is particularly effective in specialized domains, for example, healthcare, where task-specific smaller models can outperform LLM because of their adaptability and cost efficiency (Lepagnol et al. 2024). The versatility of BERT-based models for fine-grained text analytics has also been demonstrated in the information systems literature, including applications such as personality trait prediction from user-generated text (Yang et al. 2023), as well as to derive strategic insights from conversation data for managerial decision support (Chen et al. 2024). Meanwhile, advances in deep learning, particularly LLMs like GPT (OpenAI 2023), Gemini (Google 2024), Llama (Touvron et al. 2023), and Claude (Anthropic 2023), offer broad applications, although their extensive parameters, for example, 175 billion parameters in GPT-3, require significant computational resources, making them better suited for industrial-scale tasks. Thus, the present study aims to develop a research-appropriate SLM to extract sentiment scores from physician reviews.

Although these advancements in NLP have enabled sophisticated sentiment analysis across various domains, healthcare presents unique challenges that general-purpose ABSA methods cannot adequately address. Unlike traditional product or service reviews that focus on general aspect such as "product quality" or "shipping speed," healthcare reviews require extraction of sentiment across dimensions that are specifically validated by healthcare regulatory agencies such as AHRQ.

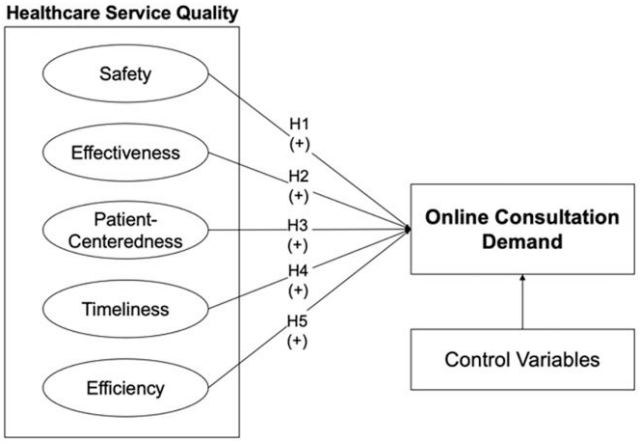
3. Model and Hypotheses Development

3.1. Healthcare Provider's Service Quality Evaluation Framework

The Agency for Healthcare Research and Quality (AHRQ) is an independent agency within the U.S. Department of Health and Human Services tasked with producing evidence that can make healthcare safer, higher quality, more accessible, equitable, and affordable (AHRQ 2024). AHRQ developed a healthcare quality evaluation framework, STEEEP, which has been utilized by various healthcare systems (Ballard 2014, Thomas Craig et al. 2020, Tahoe Forest Health System 2025). Based on STEEEP, AHRQ then proposed a six-dimension system to evaluate physicians' service quality: safety, effectiveness, patient-centeredness, timeliness, efficiency and descriptive measures (SEPTED) (AHRQ 2019). In the present study, we adopt and refine the SEPTED framework as the foundation for assessing a physician's service quality based on physician reviews.

In SEPTED, safety measures refer to the provider's responsibility to avoid harm to patients when providing the care that is intended to help them, which reflects a physician's technical skills. Effectiveness measures require healthcare providers to base their services on scientific knowledge to achieve desired treatment outcomes, which is also part of a physician's technical skills. Patient-centeredness measures refer to healthcare providers being respectful and responsive to patients' preferences, needs, and values, ensuring that patients' interests guide all clinical decisions. Timeliness measures require healthcare providers to minimize waiting time or harmful delays for patients. Efficiency measures require healthcare providers to

Figure 1. Conceptual Model for Healthcare Service Quality: SEPTE



avoid all kinds of waste, including equipment, supplies, ideas, and energy when caring for a patient. The sixth dimension of the SEPTED model is descriptive measures, which suggests using descriptive information to assess physicians' capacity for providing care, such as the number of medical procedures performed. Although potentially useful, this dimension is considered ad hoc and not an official domain defined by the Institute of Medicine according to medical sciences (AHRQ 2019). Additionally, common descriptive measures, such as percentage of certified physicians and number of surgical procedures performed, are typically not accessible to patients. Hence, we refine the SEPTED model to be SEPTE by removing the descriptive dimension for the present research endeavor as Figure 1 shows.

3.2. Online Consultation Demand Model and Hypotheses

In the present study, we assume that positive reviews can increase physicians' online consultation demand because literature shows that positive customer reviews can increase purchase (Singh et al. 2017, Guo et al. 2020). As illustrated in Figure 1, our conceptual model outlines the association between a physician's service quality and the subsequent patient demand for online consultations. To empirically examine this relationship, we construct the following model represented by Equation (1). This model posits that the online demand for physician i at time t is a function of the service quality provided by that physician, along with relevant control variables. The equation can be expressed as

$$\begin{aligned}&\text{OnlineDemand}_{i,t}\\&\quad = g(\text{ServiceQuality}_{i,t}; \text{ControlVariables}_{i,t}).\end{aligned} \tag{1}$$

Furthermore, the health service quality is estimated by SEPTE, so Equation (1) can be represented as

$$\begin{aligned}&\text{OnlineDemand}_{i,t}\\&\quad = \beta_1 \text{Safety}_{i,t} + \beta_2 \text{Effectiveness}_{i,t}\\&\qquad + \beta_3 \text{PatientCenteredness}_{i,t} + \beta_4 \text{Timeliness}_{i,t}\\&\qquad + \beta_5 \text{Efficiency}_{i,t} + \vec{\gamma} \text{ControlVariables}_{i,t}\\&\qquad + \alpha_i + \delta_t + \varepsilon_{i,t}.\end{aligned} \tag{2}$$

In Equation (2), $OnlineDemand_{i,t}$ denotes the online consultation demand for physician i at time t. The coefficients β_1–β_5 capture the effects of the five service quality dimensions, and $\vec{\gamma}$ is the vector of coefficients on the control variables. The term α_i denotes physician fixed effects and δ_t denotes time fixed effects, respectively. All standard errors are clustered at the physician level to account for within-physician correlation over time.

We hypothesize that all five dimensions of physician service quality positively influence online consultation demand, as illustrated in Figure 1. The complete hypothesis development is provided in Online Appendix A.

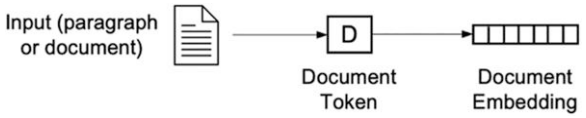
4. Methods

4.1. Language Modeling Methods

To evaluate physician service quality based on reviews by patients, precise sentiment analysis methods are essential for extracting sentiments across the SEPTE framework dimensions. Sentiment analysis of unstructured text allows us to infer patient sentiments toward service quality, and language models are currently the most effective approaches for this task (Devlin et al. 2019). Scholars in information systems have utilized these models to address various business and societal challenges, including healthcare, finance, social networks, privacy, and misinformation (Menon and Sarkar 2016, Hendershott et al. 2017, Li et al. 2017, Fang et al. 2021). In this context, we employ two prominent and customizable algorithms, Doc2Vec and BERT, to develop a method that extracts valuable insights from physician reviews and accurately assesses sentiments associated with healthcare service quality.

Doc2Vec is a distributed representation learning algorithm that generates vector embeddings for entire documents, including sentences, paragraphs, and even complete sets of documents (Le and Mikolov 2014). It has become popular for various language modeling tasks such as document classification and sentiment analysis, outperforming traditional NLP methods like bag-of-words (BoW) and Term Frequency-Inverse Document Frequency (Dai et al. 2015), and demonstrating competitiveness with approaches like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). One advantage of Doc2Vec is its

Figure 2. Structure of Doc2Vec Algorithm



ability to capture the meaning of lengthy texts. However, it may miss nuances and subtleties at the sentence or phrase level, struggling to recognize local contextual variations (Lau and Baldwin 2020). In the Doc2Vec process, the document is represented as a special token, and the model iteratively refines its output vector by considering each word in the document, generating an embedding that effectively captures the document's semantic content. The structure of Doc2Vec algorithm is shown in Figure 2.

Another popular algorithm is BERT, introduced by. BERT has significantly advanced NLP tasks like sentiment analysis, question answering, and machine translation by leveraging the transformer architecture Vaswani et al. (2017). It uses bidirectional context and masked language modeling to capture deep contextual representations, achieving state-of-the-art performance in various applications (Raffel et al. 2020). As shown in Figure 3, BERT processes input text by creating word embeddings for each word, which are passed through a transformer encoder to generate a document-level embedding used for downstream tasks. Its network architecture allows extensive customization to meet specific requirements.

However, BERT has limitations, such as a fixed input length that can lead to the omission of important text beyond its capacity, potentially missing relationships between distant parts of a document (Zaheer et al. 2020). Additionally, pretrained BERT models use a general vocabulary built from sources like Wikipedia and news articles, which may lack medical terminology and clinical jargon. This lack of domain-specific knowledge can hinder BERT's understanding of medical texts. To address these challenges, we customize BERT into a smaller language model with 110 million parameters, in contrast to the billions in LLMs, making it more focused and efficient for the healthcare domain. By integrating Doc2Vec and customized BERT embeddings, along with masked language modeling techniques, we enhance our algorithm's adaptability and effectiveness for this study.

4.2. Our Small Language Model: Doc-BERT

4.2.1. Algorithm Design. To leverage the advantages of both Doc2Vec and BERT while mitigating their respective limitations, we have developed a novel SLM algorithm called Doc-BERT. Doc2Vec excels at generating document-level embeddings, capturing the overall context and meaning of the text, and BERT's proficiency lies in understanding contextual information and modeling long-term dependencies between words. By combining these advantages, Doc-BERT aims to create a comprehensive and accurate representation of documents, making it well-suited for tasks requiring deep semantic understanding and context-aware analysis. These features are then fed into a deep neural network (DNN) for sentiment analysis. Doc-BERT shines when interpreting reviews that can be lengthy and carry complex sentiment structures. Despite an overarching positive or negative sentiment, reviews often contain context-dependent sentiments that swing the other way. Doc-BERT comprehends the broader meaning of extensive documents while recognizing the semantic variations of words in diverse contexts.

The architecture of Doc-BERT is illustrated in Figure 4. In this algorithm, each physician review is treated as a document and passed through both a Doc2Vec module and an enhanced and fine-tuned BERT module. Upon processing a document, the Doc2Vec module generates a vector, also known as an embedding, for the document. This embedding captures the fundamental semantics or overall meaning of a document. In essence, Doc2Vec can represent the global meaning of a review at the document level.

Besides the intrinsic advantages of BERT, we also innovatively customize it by integrating the disease type information into the embedding process. This modification is achieved by implementing a `link_disease_type()` function in the BERT module as an additional embedding layer, as illustrated in Pseudocode 2 of Online Appendix B. By incorporating the disease type, our algorithm effectively captures domain-specific contextual nuances and seamlessly integrates them with the document-level embeddings, resulting in

Figure 3. Structure of BERT Algorithm

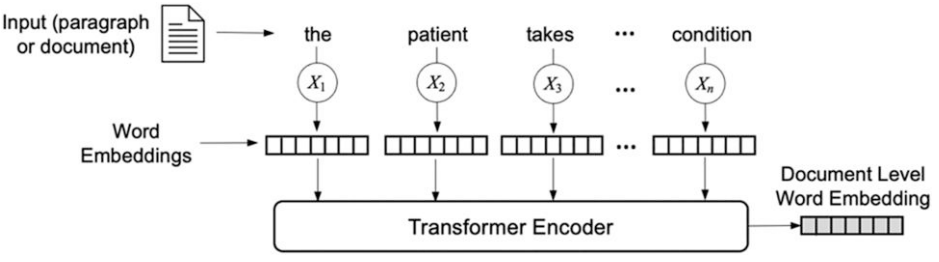
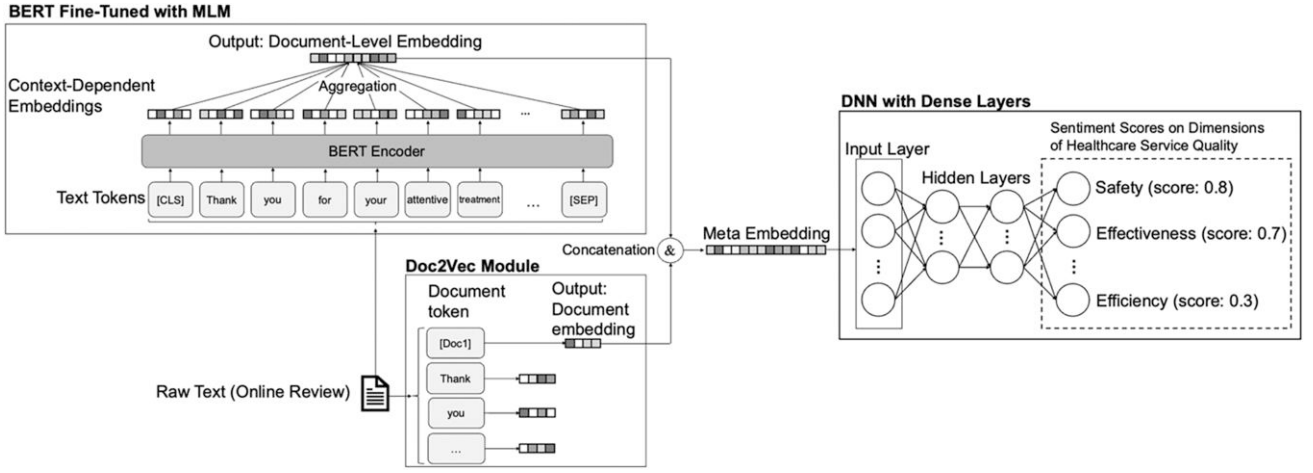


Figure 4. Architecture of Our Newly Designed Small Language Model: Doc-BERT



a deeper understanding and analysis of physician reviews.

In addition to integrating the disease type into the BERT module, we have employed a fine-tuning method, masked language modeling (MLM). This fine-tuning process enhances the BERT model's ability to predict masked tokens with greater accuracy, leading to a thorough understanding of the domain context (Salazar et al. 2020). MLM is designed to generate precise predictions to restore a corrupted input sequence. When training a language model with MLM, a certain percentage of the input tokens in a given text are randomly masked, meaning they are replaced with the [MASK] token, a special placeholder. The objective is for the model to predict the original tokens that were masked based on the context of the surrounding words. Formally, consider an input sequence S derived from a set of physician reviews, D. The sequence S is represented as $S := \langle t_1, t_2, \ldots, t_n \rangle$, where n is the number of tokens and t_i denotes a token. Let S_{-i} denote the sequence of tokens after the ith token, t_i, is excluded and replaced with a special [MASK] token, as shown:

$$S_{-i} := \langle t_1, t_2, t_{i-1}, [\text{MASK}], t_{i+1} \ldots, t_n \rangle.$$

Then define $S_{-\mathcal{M}}$ as the sequence excluding m tokens and replacing them with the [MASK] token, $|\mathcal{M}| = m$, $\mathcal{M} \subset S$. MLM's objective is to learn a probabilistic model p_θ minimizing the loss function $\mathcal{L}(D)$.

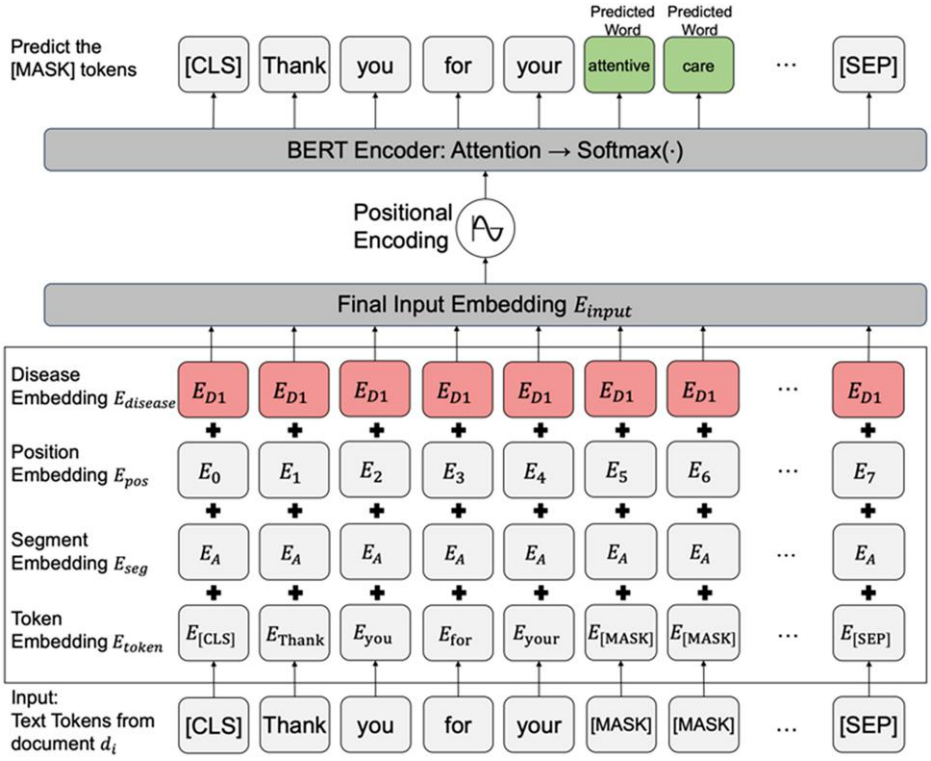
$$\mathcal{L}(D) = \mathop{\mathbb{E}}_{S \in D} \mathop{\mathbb{E}}_{\mathcal{M} \subset S} \left[\sum_{t_i \in \mathcal{M}} \log p_\theta(t_i | S_{-M}) \right]$$

The architecture of the proposed MLM fine-tuning process is illustrated in Figure 5. To initiate the process, the input document (an online review) is disseminated into a sequence of text tokens. These tokens then undergo a four-embedding procedure (Figure 5) to incorporate essential information such as word, sentence, position in the sentence, and disease type information. The final input embedding is obtained by summing these four individual embeddings. Following this embedding procedure, we employ a well-established positional encoding method, as delineated by Vaswani et al. (2017), to help BERT understand the sequence of tokens in a sentence, allowing it to understand the order and arrangement of words in a sentence. The sequence after positional encoding is sent to the BERT encoder and compute attention scores, which measures the relative importance of words in a sentence and helps BERT to focus on important words to understand the meaning of the whole sentence. The attention scores are fed into a softmax classifier to predict the original words of the [MASK] words. The next stage in the process involves the comparison of the ground truth words and the predicted [MASK] words, which is integral to the calculation of loss. This loss is then used to perform backpropagation within the deep neural network to enhance the model's predictive performance. For further reference, the specifics of the proposed BERT model fine-tuned with MLM are concisely outlined in Online Appendix B, Pseudocode 3.

The embeddings generated from the two modules, Doc2Vec and BERT, will be concatenated to create a more comprehensive and accurate representation of the entire document. Finally, these embeddings are fed into a DNN module for sentiment analysis, as illustrated in Figure 4. The DNN module utilizes this combined embedding to understand both the global context and detailed nuances of the reviews, allowing for a more holistic analysis of sentiment on multiple healthcare service quality dimensions. Detailed implementation of the DNN module is provided in Pseudocode 4 of Online Appendix B.

Doc-BERT offers several notable advantages. First and foremost, it excels in capturing both local and global contexts within long texts, allowing it to identify

Figure 5. (Color online) Architecture of BERT Fine-Tuning with Masked Language Model in Doc-BERT



scattered negative sentiments in predominantly positive reviews, and vice versa. This enhances its performance in sentiment analysis tasks that require a deep understanding of contextual nuances. Second, Doc-BERT effectively leverages medical context, even for rare words, by incorporating disease type information into BERT embeddings. This results in improved representations for medical terms, crucial for analyzing healthcare-related texts. Third, our approach is robust to noise through complementary mechanisms. Doc2-Vec captures global co-occurrence statistics, whereas BERT focuses on token-level dependencies; as a result, noise appears differently in each embedding, whereas true semantic meaning remains consistent across both. By concatenating these representations, aligned semantic components reinforce one another and mismatched noise is diluted. Additionally, the disease-type embedding layer constrains the model's attention to medically relevant relationships. Lastly, Doc-BERT is highly efficient because of its significantly fewer parameters compared with LLMs, allowing it to achieve superior performance with reduced computational resources. With its notable advantages, including its ability to understand the meaning of long texts, capture local contextual nuances, handle rare words, and generate a multimodal representation of the text, Doc-BERT is well suited for extracting the sentiment scores and the five dimensions of the SEPTE healthcare service quality model from a vast collection of physician reviews.

4.2.2. Algorithm Walkthrough with Example.

To illustrate the mechanism of our algorithm, considering the following review as an example.

> **Example Review:** "Dr. XYZ was very attentive during the online consultation about my depression, listening to my feelings and concerns and explaining his diagnosis and treatment plan patiently. He ensures I understood the plan and fosters a strong sense of attachment and trust that made me feel relaxed and supported. Based on my condition, he prescribed a psychological therapy treatment plan to support me, which instilled confidence in my recovery. My weekly therapy always started on time. During the consultation with Dr. XYZ, I felt very comfortable, and my privacy was protected. The prescribed therapy sessions were effective, and I noticed improvements in my mood, which became stable over time. Overall, the online consultation and the prescribed therapy by Dr. XYZ were both efficient. Thank you, Dr. XYZ, for your attentive care and warm attitude!"

The Doc2Vec module processes this review example to generate a document-level embedding v_d, which is a vector that encapsulates the global meaning and overall semantics of the text. By default, the Doc2Vec algorithm generates a vector with 300 elements, providing a comprehensive representation of the review's content. An example of such a vector is: $v_d = [0.37, -0.24, -0.41, 0.65, \ldots, 0.22]$. The 300 elements collectively represent the algorithm's understanding of the review.

Each element in the vector, v_{d_i}, contributes to representing one or more features, such as a word or phrase, or aspect of the five dimensions of healthcare service quality: safety, effectiveness, patient-centeredness, timeliness, and efficiency. In essence, each element v_{d_i} has a weighted representation contributing to the overall understanding of these dimensions such as safety.

To address Doc2Vec's limitations in capturing local contextual variations of semantics, we enhance the document representation by incorporating a customized BERT method. BERT embeds a sequence of tokens from three perspectives: the token ID in the vocabulary, token's sentence ID, and token's position in the sentence. Our BERT module processes each review and generate a document-level embedding v_b, which captures the nuanced contextual information and semantic relationships within the text.

An example of the embedding generated by BERT is $v_b = [0.34, 0.46, 0.76, -0.33, \ldots, 0.41]$. By default, this vector has 768 elements, each corresponding to a specific feature or aspect of the text, such as the context in which a word is used or its relationship to other words. This allows BERT to represent the contextual meaning of individual words and their interdependencies, capturing the intricacies of the five dimensions of healthcare service quality. Using the same example review as the Doc2Vec's, in the phrase "He (Dr. XYZ) … fosters a strong sense of attachment and trust …," Doc2Vec might interpret "attachment" as "accessory," which could misrepresent the patient's sentiment and seem disrespectful. In contrast, our BERT module understands "attachment" in the context of an "emotional connection," accurately reflecting the patient's true feelings.

To demonstrate the advantages of embedding disease type information, consider the following excerpt from the example review:

> "Based on my condition, he (Dr. XYZ) prescribed a [MASK] therapy treatment plan to [MASK] me, which instilled confidence in my recovery."

During the training stage, the model predicts masked words based on the surrounding context. The accuracy of the algorithm is demonstrated by its ability to recover the original words. A standard BERT model would predict the masked words as "appropriate" and "aid," which are grammatically correct but not accurate. However, with the inclusion of the patient's disease type as depression, the masked words are accurately predicted as "psychological" and "support." When the words "psychological" and "support" are correctly understood, the sentence reflects the use of an accurate treatment method and evaluates the Effectiveness dimension. Conversely, if the masked words are interpreted as "appropriate" and "aid," the sentence emphasizes tailoring care to individual patient needs and preferences, which could mistakenly align the review with the patient-centeredness dimension. After the disease-type layer is added, the BERT embedding for the example review v_b will be changed to $v'_b = [0.79, 0.68, -0.090, 0.76, \ldots, 0.59]$. This vector reflects a deeper understanding of physician reviews once the disease type is considered.

The MLM fine-tuning process substantially enhances the model's ability to capture SEPTE domain-specific nuances, especially in contexts where accurate word predictions are critical for understanding sentiment and medical relevance. For example, in the excerpt from the example review:

> "Thank you, Doctor XYZ, for your [MASK] [MASK] and warm attitude!"

A standard BERT predicts the masked words "good technology," which, although plausible, do not reflect any dimension of healthcare service quality. In contrast, the MLM-fine-tuned BERT module predicts the masked words as "attentive care," accurately reflecting the service quality in the dimension of patient-centeredness, as depicted in Figure 5. The embedding generated by the fine-tuned module $v''_b = [0.67, 0.55, -0.33, -0.04, \ldots, 0.49]$, maintaining the same number of elements as BERT but capturing nuanced meanings that are more attuned to the context of our medical-background data set. This capability ensures that our model can handle complex and contextually rich inputs, making it particularly effective for tasks that demand precise semantic interpretation.

The embeddings from Doc2Vec, v_d, and MLM-fine-tuned BERT, v''_b, are concatenated to form $v_c = [v_d, v''_b] = [0.37, -0.24, -0.41, 0.65, \ldots, 0.22, 0.67, 0.55, -0.33, -0.04, \ldots, 0.49]$, with a total length of 1068. This combined representation leverages the global meaning captured by Doc2Vec and the local contextual meaning learned from BERT. Then v_c is fed to the DNN module for sentiment analysis.

In the example review, the DNN interprets the overarching sentiment of trust and satisfaction reflected in statements like "Dr. XYZ was very attentive" and "(He fosters … trust that) … made me feel relaxed and supported" using the Doc2Vec embedding. Meanwhile, the fine-tuned BERT embedding helps DNN interpret the subtle contextual meanings in phrases such as "psychological therapy treatment plan to support me" and "fosters a strong sense of attachment and trust" accurately identifying their relevance to effectiveness and patient-centeredness dimensions. This approach is particularly beneficial when sentiments related to multiple dimensions, such as safety ("I felt very comfortable, and my privacy was protected."), timeliness ("My weekly therapy always started on time"), and efficiency ("Overall, the online consultation and the prescribed therapy … were both efficient"), are

all present in the same review. Finally, the DNN module provides sentiment scores for each dimension simultaneously, offering a detailed understanding of the review's sentiment aligned with the SEPTE healthcare service quality model. For the example review, the DNN module generates the following sentiment scores: safety = 0.86; effectiveness = 0.91; patient-centeredness = 0.88; timeliness = 0.81; efficiency = 0.84. More examples comparing the generated embeddings and sentiment scores between Doc-BERT and benchmarks are available in the Online Appendix C.

In summary, Doc-BERT is, to our knowledge, the first SLM (approximately 110M parameters) to combine global Doc2Vec embeddings with token-level BERT attention plus a disease-type embedding layer, allowing it to capture global coherence, local nuance, and medical context simultaneously. Unlike conventional ABSA approaches that either rely on attention mechanisms alone or predefined aspect lexicons, Doc-BERT integrates theory-grounded healthcare quality dimensions directly into its architecture. Its aspect identification capability achieves higher than 90% accuracy in mapping sentences to SEPTE dimensions, confirming that it consistently retrieves the authoritative categories that healthcare regulators and patients use when evaluating online healthcare service quality.

5. Data

5.1. Data Preprocessing

We assembled a data set from HaoDF.com (https://www.haodf.com/), one of the largest online physician consultation service platforms established in 2006 in China, which includes millions of reviews on 936,975 physicians from 10,526 major hospitals across China. Data were collected via web crawling over a 24-month period from February 2022 to February 2024. Both online and offline consultation reviews are available on the website; however, we exclusively utilized reviews related to online consultation services. This approach ensures that our analysis accurately reflects the specific dynamics and patient decision-making processes associated with online healthcare interactions. To control for disease-related confounding factors, we focused on physicians specializing in the five most common low-risk diseases (hypertension, rheumatoid arthritis, gastritis, depression, and menoxenia) and the five most common high-risk diseases (leukemia, lung cancer, cirrhosis, coronary heart disease, and diabetes), aligning with mortality rates from prior research (Yang et al. 2015). This resulted in a data set of 5,258 physicians and 1,316,771 reviews. The average length of all reviews is 86 Chinese characters, and 13.4% of reviews fall into the long-review category (>200 characters). A screenshot of the website interface, along with a sample physician profile and corresponding reviews, is provided in Online Appendix D. We obtained each physician's profile information, including specialty area, technical title (categorized into four levels: senior physician, associate senior physician, attending physician, and resident physician), hospital location, and engagement details such as posts, reviews, thank-you letters, and virtual gifts received over time. Weekly platform-assigned recommendation scores and patient satisfaction ratings were also collected.

For the textual reviews, we first removed stop words and other irrelevant content to retain meaningful information and then linked each review to its corresponding physician ID for averaging sentiment scores. Although transformer models such as BERT can process raw text, this filtering step improves the Doc2Vec component's performance; to keep inputs consistent, we therefore applied the same preprocessing to both Doc2Vec and BERT. We also incorporated a disease-specific medical dictionary covering all included conditions to enhance term recognition. A random sample of 10,000 reviews was manually labeled by two domain experts. They independently assessed each review across the SEPTE model dimensions and assigned sentiment scores on a scale from −2 to +2, following a predefined sentiment lexicon with emotional weights for various words (Dong and Dong 2003). Interrater reliability was high (kappa = 0.92), indicating strong agreement (Fleiss et al. 2013). The labeled data were then used to train and test our Doc-BERT algorithm against established methods.

5.2. Algorithm Performance Evaluation

5.2.1. Performance Comparison. In this section, we conduct four sets of performance assessment experiments to evaluate the proposed Doc-BERT model against 19 baseline algorithms, as summarized in Table 1. The models are compared under both zero-shot and few-shot learning settings using standard regression metrics for numeric outcomes: mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE), which together provide a comprehensive view of prediction accuracy and error magnitude.

For the performance evaluations, we randomly split the labeled data set of 10,000 reviews into training and testing sets using an 80:20 ratio. The detailed setup of performance assessments of Doc-BERT with other algorithms is provided in Online Appendix E.1. To ensure robustness, we repeat the experiment process 100 times, each time with different random splits for training and testing. All the methods were trained and tested on the same data splits in each repetition. As the output of all methods is numerical sentiment scores for the five dimensions of the SEPTE service quality model, we adopt MSE, RMSE, and MAE as the performance metrics. Table 1 displays the average MSE, RMSE, and MAE values obtained from 100 runs

Table 1. Performance Comparison Between Doc-BERT and Other Benchmark Methods

		Metrics			
	Algorithms	MSE	RMSE	MAE	No. of parameters
Experiment 1	Linear regression	0.32***	0.57***	0.36***	
	Regression tree	0.46***	0.68***	0.50***	
	Support vector regression	0.33***	0.57***	0.36***	
	Lasso regression	0.34***	0.58***	0.39***	
Experiment 2	Word2Vec	0.30***	0.55***	0.43***	
	GloVe	0.32***	0.57***	0.42***	
	Doc2Vec	0.22***	0.47***	0.36***	
	Sentence transformer	0.19***	0.44***	0.30***	
Experiment 3	BERT	0.18***	0.42***	0.30***	110 million
	BERT with disease embedding	0.18***	0.42***	0.28***	110 million
	BERT with MLM fine-tuning	0.17***	0.41***	0.28***	110 million
	ALBERT	0.19***	0.44***	0.29***	12 million
	DistilBERT	0.18***	0.42***	0.28***	66 million
Experiment 4	ChatGPT zero-shot	0.34***	0.58***	0.38***	175 billion
	ChatGPT few-shot	0.28***	0.53***	0.32***	175 billion
	Llama2 zero-shot	0.37***	0.61***	0.40***	70 billion
	Llama2 few-shot	0.32***	0.57***	0.37***	70 billion
	Gemini Nano zero-shot	0.41***	0.64***	0.42***	3 billion
	Gemini Nano few-shot	0.37***	0.61***	0.37***	3 billion
Our proposed method	Doc-BERT	0.13	0.36	0.22	110 million

Note. $n = 2{,}000$.
***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$ by t-test.

for each method. To assess the significance of differences between Doc-BERT and the baseline algorithms, we conducted paired t-tests. Remarkably, our Doc-BERT method achieved the lowest MSE value, indicating its superior performance compared with the other approaches. All t-test results showed significant differences at the 0.001 level. Experiment 1 demonstrates that Doc-BERT achieved significantly lower MSE, RMSE, and MAE values compared with conventional machine learning models, demonstrating its superior predictive power through deep learning. In Experiment 2, Doc-BERT outperforms text embedding models like Word2Vec, GloVe, and Doc2Vec, highlighting its effectiveness in capturing and representing healthcare service quality data. Experiment 3 demonstrates that Doc-BERT outperforms multiple BERT-based variants, including standard BERT, BERT with disease embeddings, BERT fine-tuned with MLM, and two more efficient compact models ALBERT and DistilBERT, because of its refined architecture that more effectively captures nuanced, dimension-level sentiment in healthcare reviews. Additional evidence that this performance advantage persists across short, medium, and long reviews is provided in Online Appendix E.2.

In Experiment 4, Doc-BERT performs better than the tested LLM baselines—ChatGPT and Llama2 in zero-shot and few-shot configurations. Consistent with recent work showing that larger models are not always superior for domain-specific classification problems, this result illustrates that a resource-efficient, task-aligned SLM can be highly competitive in a focused setting (Lepagnol et al. 2024). The performance advantage of Doc-BERT reflects its specialized attention to healthcare text, enabling it to discern subtle semantic nuances that broader LLMs may overlook. Furthermore, the domain-specific fine-tuning enhances its understanding of healthcare-related sentiment and service quality, yielding consistently high predictive accuracy relative to the LLM baselines we consider. Table 1 also reports the number of training parameters required by each model. This metric is crucial as it directly impacts both the training time and computational resource requirements, which in turn determine overall cost. A smaller number of parameters typically leads to faster training times and reduced computational demands, making our model potentially more time and cost efficient for practical applications, especially in settings with limited computational resources. Notably, our proposed method, Doc-BERT, requires 110 million parameters, significantly fewer than LLMs with billions of parameters, highlighting its efficiency and suitability for resource-constrained environments. Together, these results highlight Doc-BERT's effectiveness and efficiency, underscoring its promise as a practical tool for comprehensive healthcare service quality assessment, particularly in resource-constrained environments.

5.2.2. Architecture Comparison. Although LLMs such as ChatGPT, Llama, and Gemini offer powerful general-purpose capabilities, their performance in our

task of aspect-specific sentiment extraction in healthcare reviews was consistently outperformed by our proposed Doc-BERT model. This performance difference is attributable to three main factors:

First, Doc-BERT benefits from domain-specific fine-tuning on medical reviews. LLMs are pretrained on broad corpora such as Wikipedia and Common Crawl, which lack sufficient healthcare-specific sentiment nuances and clinical vocabulary. Doc-BERT incorporates disease-type embeddings and is fine-tuned on labeled review data aligned with a healthcare quality framework, allowing it to capture semantic subtleties that LLMs overlook.

Second, our architecture is more task-aligned and parameter-efficient. LLMs typically require prompt engineering and are optimized for text generation or general question answering, not structured document-level regression across multiple predefined dimensions. Their large context windows are underutilized in our setting, while their billions of parameters increase the risk of overfitting when fine-tuned on smaller labeled data sets. Doc-BERT, by contrast, integrates lightweight contextual and global embeddings, resulting in stronger generalization with lower computational cost.

Third, the hybrid structure of Doc-BERT, combining Doc2Vec for global document meaning with fine-tuned BERT for contextual nuance, proves more effective in interpreting lengthy reviews that contain mixed or contradictory sentiments across different service dimensions. Reviews often contain both positive and negative expressions regarding different aspects of healthcare quality. Our model is designed to capture such sentiment contradictions across dimensions, whereas LLMs often average sentiment across the entire review unless explicitly prompted otherwise.

To further demonstrate reliability, we extended our sentiment accuracy tests to include an independent RateMDs data set, where Doc-BERT again consistently outperformed all baselines (details in Online Appendix F). Given Doc-BERT's outstanding performance, we utilize it to process more than 1 million reviews in our data set, enabling a comprehensive healthcare quality evaluation. This ensured that we obtained accurate and reliable assessments across a larger sample of reviews.

5.3. Descriptive Statistics

Table 2 presents the descriptive statistics of our data set, encompassing various personal attributes of physicians, both time varying and invariant, as well as the sentiment scores pertaining to their service dimensions of 5,258 physicians over 24 months for a total of 546,832 observations. These sentiment scores are derived using our Doc-BERT method and aggregated at the weekly level. Online consultation demand is defined as the number of weekly online appointments a physician receives on the platform. On average, physicians receive 1.8 online appointments per week, with a maximum of 74 appointments. Approximately 50% of physicians hold the highest technical title of

Table 2. Descriptive Statistics of Our Data Set

Variable	Mean	Standard deviation	Median	Minimum	Maximum
Dependent variable: *Online consultation demand (number of appointments per physician per week)*	1.80	3.77	0	0	74
Physician attributes (individual level)					
Risky	0.65	0.47	1	0	1
Title rank 1	0.0080	0.080	0	0	1
Title rank 2	0.12	0.31	0	0	1
Title rank 3	0.39	0.47	0	0	1
Title rank 4	0.48	0.51	0	0	1
Platform tenure	5.77	2.83	6	0	9
Physician attributes (weekly)					
No. of posts	0.68	24.67	0	0	46
No. of reviews	2.41	6.05	1	0	102
No. of thank you letters	0.36	1.01	0	0	38
No. of gifts	0.92	3.21	0	0	66
Numerical ratings (weekly)					
Numerical rating score by patient	0.87	0.19	1	0.48	1
Recommend score by platform	4.30	0.34	4	0	5
Service quality dimensions (weekly)					
Safety	0.25	0.21	0	0	0.43
Effectiveness	0.15	0.20	0	0	0.43
Patient-centeredness	0.61	0.48	1	−0.080	1
Timeliness	0.79	0.22	1	−0.32	1.60
Efficiency	0.59	0.46	1	−0.12	1.86

senior physician. Physicians have been using the platform for an average of 5.77 years, with a maximum of 9 years. On average, physicians receive 2.41 reviews per week, with a maximum of 102 reviews.

Online Appendix G presents the correlation matrix of variables used in the regression model. Note that some variables are categorical (e.g., physician title rank); therefore, we use Pearson's R for continuous-continuous variables, correlation ratio for categorical-continuous variables, and Cramer's V for categorical-categorical variables.

6. Results

6.1. Results of Primary Empirical Models

To examine our research questions, we estimate a fixed-effects regression model based on Equation (2) in Section 3.2, incorporating five time-varying control variables and physician fixed effects α_i to control for unobserved heterogeneity. The five control variables are: the number of healthcare-related posts shared by the physician ($Posts_{i,t}$), the number of reviews received ($Reviews_{i,t}$), the number of thank-you letters ($ThankYou\ Letters_{i,t}$), the number of virtual gifts ($Gifts_{i,t}$) during week t, and the number of online consultations received in the previous week ($OnlineDemand_{i,t-1}$).

We used Poisson regression to estimate the model, as the dependent variable is a count with similar mean and variance. The results are presented in Table 3. We also conducted negative binomial regression analyses, which yielded similar results (see Online Appendix H). To further examine the effects of variables across different aspects, we specified three distinct models in a stepwise manner.

Model 1 includes five time-varying control variables that are directly observable on the platform without reading textual reviews. The number of posts shared by a physician is statistically insignificant, suggesting this physician-initiated behavior does not significantly influence patients' appointment decisions, likely because it conveys limited information about service quality. In contrast, patient-initiated engagement—the total number of reviews (which includes both positive and negative reviews, $\beta = 0.50$), thank-you letters ($\beta = 0.17$), and virtual gifts ($\beta = 0.14$)—are all statistically significant. Notably, the number of reviews has a substantial larger effect on online consultation demand than thank-you letters and virtual gifts. This suggests that the content of reviews may influence consultation demand more than numerical indicators of positive feedback (such as thank-you letters or virtual gifts), aligning with prior studies

Table 3. Regression Results of Our Empirical Models

	Model 1	Model 2	Model 3
Control variables (physician's time-varying attributes)			
Number of Posts Created Per Week	0.00031	0.00083	0.00084
	(0.00081)	(0.00091)	(0.00076)
Number of Online Reviews Received Per Week	0.50***	0.26***	0.19***
	(0.023)	(0.025)	(0.023)
Number of Thank You Letters Received Per Week	0.17***	0.15***	0.12***
	(0.040)	(0.041)	(0.040)
Number of Virtual Gifts Received Per Week	0.14***	0.082***	0.081***
	(0.0079)	(0.0073)	(0.0082)
Number of Online Consultations of Previous Week	0.85***	0.84***	0.85***
	(0.061)	(0.048)	(0.044)
Numerical rating scores			
Average Numerical Rating Score by Patient Per Week		0.039	0.22
		(0.12)	(0.26)
Recommendation Score by Platform Per Week		2.88***	3.00***
		(0.072)	(0.078)
Service quality scores			
Safety			0.11***
			(0.013)
Effectiveness			0.78***
			(0.10)
Patient-centeredness			0.66***
			(0.077)
Timeliness			0.23**
			(0.083)
Efficiency			0.18**
			(0.056)
Pseudo-R^2	0.31	0.35	0.72

Note. $N = 546{,}832$.
***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

emphasizing the importance of analyzing qualitative comments (Siegrist 2013). The control variable, the number of online consultation appointments in the previous week, is statistically significant ($\beta = 0.85$) because of the high correlation between these variables. However, no endogeneity issue arises because the dependent variable (current appointments) does not influence the independent variable (previous week appointments).

In Model 2, we added two additional time-varying variables: the average patient rating (numerical) each week and the platform's recommendation score for each physician. The five variables from Model 1 showed slightly reduced coefficient magnitudes but remained statistically significant with consistent directional effects. Notably, the platform's recommendation score is a significant predictor of online consultation demand ($\beta = 2.88$). This finding is unsurprising given that the platform prominently displays the recommendation score in the most visible position on the website. This score is algorithmically assigned by the platform based on multiple performance metrics, serving as an authoritative quality signal. Conversely, patient ratings are not statistically significant, likely because of limited variance ($\bar{x} = 0.87, s = 0.19$). These results provide initial evidence that patients rely on certain quality signals—particularly the platform's recommendation score, review volume, and patient appreciation signals such as thank-you letters and virtual gifts. Notably, the coefficient for the review volume decreased from 0.50 in Model 1 to 0.26 in Model 2, suggesting that the platform recommendation score partially captures information also reflected in review volume. Neither patient ratings nor physician-shared posts significantly influence consultation demand. Model 2 demonstrates that patients respond to aggregate quality signals that published by the platform. However, these variables represent black-box measures that do not reveal which specific service quality dimensions patients value. This limitation motivates our primary analysis: extracting interpretable service quality dimensions from textual reviews using Doc-BERT to identify which aspects of physician performance drive patient choice.

In Model 3, we incorporate the five service quality dimensions from the SEPTE framework alongside all variables from Model 2. These dimensions—safety, effectiveness, patient-centeredness, timeliness, and efficiency—are measured as sentiment scores extracted from textual reviews using Doc-BERT. The results indicate that all five dimensions have statistically significant effects on online consultation demand, supporting our hypotheses and the SEPTE framework. Notably, the coefficient for review volume is further attenuated (from $\beta = 0.26$ in Model 2 to $\beta = 0.19$ in Model 3), underscoring that the content of reviews matters more than review quantity alone. Among the five dimensions, effectiveness has the largest impact ($\beta = 0.78$), followed by patient-centeredness ($\beta = 0.66$), timeliness ($\beta = 0.23$), efficiency ($\beta = 0.18$), and safety ($\beta = 0.11$). These findings align with prior health service quality research (Fung et al. 2005, Dagger et al. 2007, Hirpa et al. 2020, Lamprecht et al. 2020) but our results derived from NLP of online reviews rather than surveys, demonstrating the value of naturally occurring patient feedback. The lagged dependent variable ($\beta = 0.85$) indicates strong demand persistence, reflecting reputation effects. These results reveal that to increase online consultation demand, physicians should prioritize two key service quality dimension: effectiveness (technical competence in diagnosis and treatment) and patient-centeredness (empathy and attentiveness to patients' needs).

6.2. Robustness Checks

6.2.1. Difference in Difference with Propensity Score Matching. To address potential selection bias, we employ a difference-in-differences (DID) method combined with propensity score matching (PSM). This method ensures that physicians with higher service quality scores are comparable to those with lower scores on observable characteristics. Using the effectiveness dimension as an illustrative example, we define the treatment group as physicians with effectiveness sentiment scores at or above the 75th percentile. The control group is constructed using PSM to select physicians from the remaining pool, matching on three observable characteristics: title rank, account tenure, and geographic location. We select these three characteristics because they are platform-provided attributes that reflect physicians' professional status and service capacity. We estimate propensity scores using logistic regression and employ nearest neighbor matching to pair each physician in the treatment group with a control physician having the closest propensity score. The DID estimates show positive effects of each service quality dimension on online consultation demand, consistent with our main findings in Model 3. Coarsened exact matching (CEM) yielded similar results, further reinforcing the robustness of our findings (see Online Appendix I.1 for detailed results).

We verify covariate balance after matching and find that the standardized differences for all matching variables fall below the 5% threshold, confirming successful mitigation of selection bias (see Online Appendix I.2 for balance diagnostics). Additionally, we employ a generalized method of moments (GMM) estimator to address potential endogeneity arising from the lagged dependent variable. The Arellano-Bond test for second-order autocorrelation ($p = 0.594$) and the Sargan-Hansen test for instrument validity ($\chi = 1.04, p = 0.31$) support the model specification, confirming no problematic correlation between the

lagged dependent variable and the error term, validating our main model (details in Online Appendix I.3).

6.2.2. Impact of Review Recency. We investigate whether the recency of physician reviews moderates the relationship between service quality dimensions and consultation demand. Specifically, we test whether recent reviews (posted in the current week) exert a stronger influence on patient choice than the cumulative historical record. The results reveal significant positive interaction effects for all five SEPTE dimensions, indicating that patients place greater weight on recent feedback when evaluating physician service quality. These findings complement our main analysis and demonstrate that the influence of service quality on online consultation demand is amplified when fresh reviews are available (see Online Appendix I.4).

6.2.3. Impact of Review Sentiment Asymmetry. We explore asymmetry in review sentiment by examining whether positive and negative reviews exert differential effects on consultation demand. Specifically, we investigate whether a single positive review can improve demand for physicians with generally low service quality scores, and whether a negative review disproportionately reduces demand for highly rated physicians. The results indicate that isolated positive or negative reviews do not significantly alter consultation demand patterns, suggesting that patients rely on aggregate quality signals rather than individual reviews when making consultation decisions. This finding underscores the robustness of our sentiment-based quality measures, which capture overall service quality rather than being driven by outlier reviews (details in Online Appendix I.5).

7. Contributions and Future Work

In the present study, we developed a fine-tuned SLM specifically designed to extract service quality scores from physician reviews by utilizing a refined healthcare service quality evaluation framework, and then we empirically demonstrated that these quality dimensions significantly impact online consultation demand, using data from one of the largest online health platforms in the world.

Our algorithm, Doc-BERT, offers several significant advantages that enhance its impact and utility in sentiment analysis in the healthcare domain. First, Doc-BERT, as an SLM, is designed with fewer parameters compared with LLMs. This efficiency translates into quicker training times and reduces computational costs, making it ideal for real-time analysis and practical applications, especially in environments with limited computational resources. Second, our Doc-BERT is specially fine-tuned for the healthcare domain via MLM on the medical terminology, improving its ability to understand clinical language and contextual nuances. This tuning ensures that Doc-BERT can accurately interpret and analyze physician reviews. Third, by integrating embeddings from both Doc2Vec and BERT, Doc-BERT captures both global document-level meaning and local contextual detail, allowing it to model domain-specific jargon and sentiment in medical service discussions and improving the accuracy of sentiment analysis.

Our research contributes to the information systems field in three ways. Theoretically, we refine the SEPTED framework into the SEPTE model, specifically tailored to evaluate healthcare providers' service quality in an online consultation context. This refinement is grounded in standards ratified by healthcare authorities and governmental agencies, ensuring both conceptual validity and practical relevance. Our study is among the first to adapt and apply this framework to open-ended physician reviews, thereby extending theory on how patients evaluate provider quality in digital health settings.

From a methodological perspective, we develop Doc-BERT, an SLM specifically optimized for healthcare sentiment analysis that captures both the overall meaning of lengthy texts and local contextual nuances, even for rare medical terms. Doc-BERT integrates document-level embeddings from Doc2Vec with token-level contextual embeddings from BERT and adds an embedding layer for disease type, enabling more nuanced interpretation of each review and richer semantic representations of service quality. A fine-tuned MLM strategy further deepens linguistic comprehension in this domain and enhances the model's ability to interpret complex sentiment structures within reviews. In our experiments, these design choices allow Doc-BERT to outperform 19 widely used text-mining methods in analyzing sentiment in physician reviews. Although our focus is on healthcare, the design illustrates how a task-aligned, domain-tuned SLM can be an effective and resource-efficient option for sentiment analysis of user-generated content across online platforms. By successfully applying a customized SLM to physician reviews, our research addresses gaps in the literature and offers a template for studying other under-researched domains where nuanced textual understanding is crucial. This approach highlights the potential of SLMs to provide detailed, domain-specific insights and encourages further exploration of task-aligned, fine-tuned models in diverse contexts. Moreover, the design of Doc-BERT is adaptable beyond healthcare, suggesting that similar architectures could support effective sentiment analysis of user-generated content in other online forums and social media platforms. Together, these contributions not only deepen understanding of healthcare service quality but also lay the groundwork for innovative applications of sentiment analysis in additional domains.

Our research offers several implications for practitioners by clarifying how the five SEPTE dimensions shape online consultation demand. For physicians, understanding the relative contribution of each SEPTE dimension enables physicians to prioritize targeted improvements in service delivery and develop strategies to enhance both consultation demand and patient satisfaction. The refined SEPTE framework, tailored for online consultations, can also serve as a foundation for integrating patient-specific data—such as health status, demographics, and preferences—into more personalized decision-support tools and recommendation systems that reveal how individual factors influence patients' choices.

For healthcare administrators and platform operators, our findings provide an evidence-based lens for evaluating physician services and understanding their effects on online consultation demand, thereby informing resource allocation, performance management, and quality-improvement initiatives. Looking ahead, incorporating economic variables such as insurance coverage and pricing models would allow future research to examine the joint role of service quality and financial considerations, supporting more nuanced policy recommendations for healthcare platforms and payers, particularly around optimizing consultation offerings and improving access in cost-sensitive environments.

Our study has several limitations that suggest future research. First, our analysis relies exclusively on online platform data and does not incorporate offline clinical outcomes such as readmission rates, complication rates, or treatment success metrics. Integrating online review data with objective clinical quality measures would provide a more comprehensive assessment of physician service quality and enable validation of patient-reported quality perceptions against clinical performance indicators. Second, a lack of patient data may introduce endogeneity in our model of patients' online consultation choices; some patients might prefer certain physicians for reasons not captured in our data, such as favoring local physicians for potential in-person visits in the future. Third, we currently lack insights into offline consultations. To address this, we plan to gather data from a local hospital where some physicians also provide consultation services on HaoDF.com, which will help us gain additional insights and further distinguish these two formats of consultation. Lastly, the structural differences among countries' online health consultation or health systems, such as health insurance policy, may affect the interpretation and applicability of our findings.

Acknowledgments

The authors thank the senior editor, associate editor, and anonymous reviewers for constructive comments that significantly improved this manuscript.

References

AHRQ (2019) Examples of physician quality measures for consumers. Retrieved December 22, https://www.ahrq.gov/talkingquality/measures/setting/physician/examples.html.

AHRQ (2024) Mission and budget. Retrieved December 22, https://www.ahrq.gov/cpi/about/mission/index.html.

Al-Abri R, Al-Balushi A (2014) Patient satisfaction survey as a tool towards quality improvement. *Oman Medical J.* 29(1):3–7.

Anthropic (2023) Model card and evaluations for Claude models. Retrieved May 7, https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf.

Ballard DJ (2014) *The Guide to Achieving STEEEP Health Care: Baylor Scott & White Health's Quality Improvement Journey* (Productivity Press, Boca Raton, FL).

Carr-Hill RA (1992) The measurement of patient satisfaction. *J. Public Health Medicine* 14(3):236–249.

Chen Y, Rui H, Whinston AB (2024) Conversation analytics: Can machines read between the lines in real-time strategic conversations? *Inform. Systems Res.* 36(1):440–455.

Cottier B, Rahman R, Fattorini L, Maslej N, Besiroglu T, Owen D (2024) The rising costs of training frontier AI models. Preprint, submitted May 31, https://arxiv.org/abs/2405.21015.

Dagger TS, Sweeney JC, Johnson LW (2007) A hierarchical model of health service quality: Scale development and investigation of an integrated model. *J. Service Res.* 10(2):123–142.

Dai AM, Olah C, Le QV (2015) Document embedding with paragraph vectors. Preprint, submitted July 29, https://doi.org/10.48550/arXiv.1507.07998.

Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. Burstein J, Doran C, Solorio T, eds. *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Tech.* (Association for Computational Linguistics, Kerrville, TX), 4171–4186.

Dong Z, Dong Q (2003) Hownet: A hybrid language and knowledge resource. Zong C, ed. *Proc. Internat. Conf. Natural Language Processing Knowledge Engrg.* (IEEE Press, Piscataway, NJ), 820–824.

Dorsey ER, Topol EJ (2016) State of telehealth. *New England J. Medicine* 375(2):154–161.

Emmert M, Meier F (2013) An analysis of online evaluations on a physician rating website: Evidence from a German public reporting instrument. *J. Medical Internet Res.* 15(8):e2655.

Epstein RM, Street RL (2011) The values and value of patient-centered care. *Ann. Family Medicine* 9(2):100–103.

Evans R, Berman S, Burlingame E, Fishkin S (2020) It's time to take patient experience measurement and reporting to a new level: Next steps for modernizing and democratizing national patient surveys. Retrieved May 3, https://www.healthaffairs.org/do/10.1377/forefront.20200309.359946.

Fan W, Zhou Q, Qiu L, Kumar S (2023) Should doctors open online consultation services? An empirical investigation of their impact on offline appointments. *Inform. Systems Res.* 34(2):629–651.

Fang X, Gao Y, Hu PJH (2021) A prescriptive analytics method for cost reduction in clinical decision making. *MIS Quart.* 45(1):83–115.

Fleiss JL, Levin B, Paik MC (2013) *Statistical Methods for Rates and Proportions* (John Wiley & Sons, Hoboken, NJ).

Fung CH, Elliott MN, Hays RD, Kahn KL, Kanouse DE, McGlynn EA, Spranca MD, et al. (2005) Patients' preferences for technical versus interpersonal quality when selecting a primary care physician. *Health Services Res.* 40:957–977.

Gao G, Greenwood BN, Agarwal R, McCullough J (2015) Vocal minority and silent majority: How do online ratings reflect population perceptions of quality? *MIS Quart.* 39(3):565–589.

Gao GG, McCullough JS, Agarwal R, Jha AK (2012) A changing landscape of physician quality reporting: Analysis of patients' online ratings of their physicians over a 5-year period. *J. Medical Internet Res.* 14(1):e38.

Google (2024) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Retrieved May 7, https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

Grabner-Kräuter S, Waiguny MK (2015) Insights into the impact of online physician reviews on patients' decision making: Randomized experiment. *J. Medical Internet Res.* 17(4):e93.

Grand View Research (2024) Healthcare IT market size, share & trends analysis report by application (EHR, CPOE, electronic prescribing systems, medical imaging information), by delivery mode, by end use, by region, and segment forecasts, 2024-2030. Retrieved October 25, https://www.grandviewresearch.com/industry-analysis/healthcare-it-market.

Grøndahl VA, Wilde-Larsson B, Karlsson I, Hall-Lord ML (2013) Patients' experiences of care quality and satisfaction during hospital stay: A qualitative study. *Eur. J. Personality Centered Healthcare* 1(1):185–192.

Guo J, Wang X, Wu Y (2020) Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *J. Retailing Consumer Services* 52:101891.

Hao H (2015) The development of online doctor reviews in China: An analysis of the largest online doctor review website in China. *J. Medical Internet Res.* 17(6):e134.

Hao H, Zhang K, Wang W, Gao G (2017) A tale of two countries: International comparison of online doctor reviews between China and the United States. *Internat. J. Medical Inform.* 99:37–44.

Hendershott T, Zhang MX, Zhao JL, Zheng E (2017) Call for papers—Special issue of information systems research fintech: Innovating the financial industry through emerging information technologies. *Inform. Systems Res.* 28(4):885–886.

Hirpa M, Woreta T, Addis H, Kebede S (2020) What matters to patients? A timely question for value-based care. *PLoS One* 15(7):e0227845.

Jiang L, Hou J, Ma X, Pavlou PA (2024) Punished for success? A natural experiment of displaying clinical hospital quality on review platforms. *Inform. Systems Res.* 36(1):285–306.

Ko D, Mai F, Shan Z, Zhang D (2019) Operational efficiency and patient-centered health care: A view from online physician reviews. *J. Oper. Management* 65(4):353–379.

Koonin LM, Hoots B, Tsang CA, Leroy Z, Farris K, Jolly B, Antall P, et al. (2020) Trends in the use of telehealth during the emergence of the Covid-19 pandemic—United States, January–March 2020. *Morbidity and Mortality Weekly Rep.* 69(43):1595–1599.

Lamprecht R, Struppek J, Heydecke G, Reissmann DR (2020) Patients' criteria for choosing a dentist: Comparison between a university-based setting and private dental practices. *J. Oral Rehabilitation* 47(8):1023–1030.

Lau JH, Baldwin T (2020) An empirical evaluation of Doc2vec with practical insights into document embedding generation. Blunsom P, Cho K, Cohen S, Grefenstette E, Hermann KM, Rimell L, Weston J, Yih SW, eds. *Proc. 1st Workshop Representation Learn. NLP* (Association for Computational Linguistics, Kerrville, TX), 78–86.

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. Xing EP, Jebara T, eds. *Proc. 31st Internat. Conf. Machine Learn.* (PMLR, New York), 1188–1196.

Lee V (2017) Transparency and trust: Online patient reviews of physicians. *New England J. Medicine* 376(3):197–199.

Lepagnol P, Gerald T, Ghannay S, Servan C, Rosset S (2024) Small language models are good too: An empirical study of zero-shot classification. Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, eds. *Proc. Joint Internat. Conf. Comput. Linguistics, Language Resources Evaluation* (ELRA, Marseille, France).

Li M, Tan CH, Wei KK, Wang K (2017) Sequentiality of product review information provision. *MIS Quart.* 41(3):867–892.

Liu B (2022) *Sentiment Analysis and Opinion Mining* (Springer Nature, Cham, Switzerland).

Menon S, Sarkar S (2016) Privacy and big data. *MIS Quart.* 40(4):963–982.

Michel J, Kawonga M, Rubin H (2023) Pandemic-driven telehealth uptake: The missing healthcare provider, system and patient voices. *Frontiers Digital Health* 5:1293921.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 3111–3119.

NEJM Catalyst (2017) What is patient-centered care? Retrieved January 1, https://catalyst.nejm.org/doi/full/10.1056/CAT.17.0559.

OpenAI (2023) Gpt-4 technical report. Retrieved May 7, https://arxiv.org/pdf/2303.08774.

Pan X, Zhou X, Yu L, Hou L (2023) Switching from offline to online health consultation in the post-pandemic era: The role of perceived pandemic risk. *Frontiers Public Health* 11:1121290.

Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Foundations Trends Inform. Retrieval* 2(1–2):1–135.

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, et al. (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Machine Learn. Res.* 21(1):5485–5551.

Saifee DH, Zheng Z, Bardhan IR, Lahiri A (2020) Are online reviews of physicians reliable indicators of clinical outcomes? A focus on chronic disease management. *Inform. Systems Res.* 31(4):1282–1300.

Salazar J, Liang D, Nguyen TQ, Kirchhoff K (2020) Masked language model scoring. Jurafsky D, Chai J, Schluter N, Tetreault J, eds. *Proc. 58th Annual Meeting Assoc. Comput. Linguistics* (Association for Computational Linguistics, Kerrville, TX), 2699–2712.

Scaletta T (2015) Timely patient satisfaction surveys: No longer an option. Retrieved May 3, https://www.fiercehealthcare.com/healthcare/timely-patient-satisfaction-surveys-no-longer-option.

Schlesinger M, Grob R, Shaller D (2015) Using patient-reported information to improve clinical practice. *Health Services Res.* 50(S2):2116–2154.

Siegrist RB (2013) Patient satisfaction: History, myths, and misperceptions. *Virtual Mentor* 15(11):982–987.

Singh A, Prasher A, Kaur N (2020) Assessment of hospital service quality parameters from patient, doctor and employees' perspectives. *Total Quality Management Bus. Excellence* 31(13–14):1467–1486.

Singh A, Alryalat MAA, Alzubi JA, Sarma HK (2017) Understanding Jordanian consumers' online purchase intentions: Integrating trust to the Utaut2 framework. *Internat. J. Appl. Engrg. Res.* 12(20):10258–10268.

Tahoe Forest Health System (2025) Quality and safety. Retrieved September 20, https://www.tfhd.com/about/quality-safety/#toc_Quality_and_Safety_Defined.

Thomas Craig KJ, McKillop MM, Huang HT, George J, Punwani ES, Rhee KB (2020) US hospital performance methodologies: A scoping review to identify opportunities for crossing the quality chasm. *BMC Health Services Res.* 20(1):640.

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, et al. (2023) Llama: Open and efficient foundation language models. Preprint, submitted February 27, https://doi.org/10.48550/arXiv.2302.13971.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, et al. (2017) Attention is all you need. von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, eds. *Proc. 31st Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY),

Vennik FD, Adams SA, Faber MJ, Putters K (2014) Expert and experiential knowledge in the same place: Patients' experiences with online communities connecting patients and health professionals. *Patient Ed. Counseling* 95(2):265–270.

Wan Y, Peng Z, Wang Y, Zhang Y, Gao J, Ma B (2021) Influencing factors and mechanism of doctor consultation volume on online

medical consultation platforms based on physician review analysis. *Internet Res.* 31(6):2055–2075.

Xu Y, Armony M, Ghose A (2021) The interplay between online reviews and physician demand: An empirical investigation. *Management Sci.* 67(12):7344–7361.

Yang H, Guo X, Wu T (2015) Exploring the influence of the online physician service delivery process on patient satisfaction. *Decision Support Systems* 78(C):113–121.

Yang K, Lau RY, Abbasi A (2023) Getting personal: A deep learning artifact for text-based measurement of personality. *Inform. Systems Res.* 34(1):194–222.

Yaraghi N, Wang W, Gao G, Agarwal R (2018) How online quality ratings influence patients' choice of medical providers: Controlled experimental survey study. *J. Medical Internet Res.* 20(3):e99.

Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, Pham P, et al. (2020) Big bird: Transformers for longer sequences. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Proc. 34th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 17283–17297.

Bin Zhang is an associate professor of information systems at Mays Business School, Texas A&M University. He has a PhD in information systems management and MS in machine learning. His work has appeared in premier journals including Information Systems Research, MIS Quarterly, Production and Operations Management, and Journal of Management Information Systems. His research has been funded by federal agencies such as the National Science Foundation and National Institutes of Health.

Haijing Hao research includes technology adoption, online health communities/online doctor reviews, econometrical modeling, Bayesian learning models, and text mining. She has published in Information Systems Research, Journal of Medical Internet Research, International Journal of Medical Informatics, International Journal of Public Health, Electronic Commerce Research, IEEE Transactions on Engineering Management, and so on. She received a PhD in information systems management.

Yongcheng Zhan is an assistant professor of information systems at the Orfalea College of Business, California Polytechnic State University. He holds a PhD in Management Information Systems. His research focuses on artificial intelligence and data analytics in healthcare, education, and business processes. His work has appeared in the Journal of Medical Internet Research, the Journal of the American Medical Informatics Association, and the Journal of Information Systems Education.

Jiang Wu is a professor at the School of Information Management, Wuhan University, and director of the Research Center for Digital Commerce and Industry. His research interests include business data intelligence, social network computing, digital S&T innovation, and smart healthcare. He has published in journals such as Journal of the Association for Information Science and Technology, Information Processing & Management, International Journal of Electronic Commerce, and Technovation.