Original Paper

# Fine-Tuned Bidirectional Encoder Representations From Transformers Versus ChatGPT for Text-Based Outpatient Department Recommendation: Comparative Study

Eunbeen Jo[1*], BA; Hakje Yoo[2,3*], PhD; Jong-Ho Kim[4,5], PhD; Young-Min Kim[6], PhD; Sanghoun Song[7], PhD; Hyung Joon Joo[1,4,5], MD, PhD

[1]Department of Medical Informatics, Korea University College of Medicine, Seoul, Republic of Korea

[2]Department of Bio-Mechatronic Engineering, Sungkyunkwan University College of Biotechnology and Bioengineering, Gyeonggi, Republic of Korea

[3]Medical AI Research Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea

[4]Korea University Research Institute for Medical Bigdata Science, Korea University, Seoul, Republic of Korea

[5]Department of Cardiology, Cardiovascular Center, Korea University College of Medicine, Seoul, Republic of Korea

[6]School of Interdisciplinary Industrial Studies, Hanyang University, Seoul, Republic of Korea

[7]Department of Linguistics, Korea University, Seoul, Republic of Korea

[*]these authors contributed equally

**Corresponding Author:**
Hyung Joon Joo, MD, PhD
Department of Medical Informatics
Korea University College of Medicine
73, Inchon-ro
Seoul, 02841
Republic of Korea
Phone: 82 2 920 5445
Email: drjoohj@gmail.com

## Abstract

**Background:** Patients often struggle with determining which outpatient specialist to consult based on their symptoms. Natural language processing models in health care offer the potential to assist patients in making these decisions before visiting a hospital.

**Objective:** This study aimed to evaluate the performance of ChatGPT in recommending medical specialties for medical questions.

**Methods:** We used a dataset of 31,482 medical questions, each answered by doctors and labeled with the appropriate medical specialty from the health consultation board of NAVER (NAVER Corp), a major Korean portal. This dataset includes 27 distinct medical specialty labels. We compared the performance of the fine-tuned Korean Medical bidirectional encoder representations from transformers (KM-BERT) and ChatGPT models by analyzing their ability to accurately recommend medical specialties. We categorized responses from ChatGPT into those matching the 27 predefined specialties and those that did not. Both models were evaluated using performance metrics of accuracy, precision, recall, and $F_1$-score.

**Results:** ChatGPT demonstrated an answer avoidance rate of 6.2% but provided accurate medical specialty recommendations with explanations that elucidated the underlying pathophysiology of the patient's symptoms. It achieved an accuracy of 0.939, precision of 0.219, recall of 0.168, and an $F_1$-score of 0.134. In contrast, the KM-BERT model, fine-tuned for the same task, outperformed ChatGPT with an accuracy of 0.977, precision of 0.570, recall of 0.652, and an $F_1$-score of 0.587.

**Conclusions:** Although ChatGPT did not surpass the fine-tuned KM-BERT model in recommending the correct medical specialties, it showcased notable advantages as a conversational artificial intelligence model. By providing detailed, contextually appropriate explanations, ChatGPT has the potential to significantly enhance patient comprehension of medical information, thereby improving the medical referral process.

***(JMIR Form Res 2024;8:e47814)*** doi: 10.2196/47814

**KEYWORDS**

natural language processing; bidirectional encoder representations from transformers; large language model; generative pretrained transformer; medical specialty prediction; quality of care; health care application; ChatGPT; BERT; AI technology; conversational agent; AI; artificial intelligence; chatbot; application; health care

## Introduction

Natural language processing technology has the potential to transform the process of health care and further improve the quality of care [1]. Among natural language processing deep learning models, transformer-based models, including bidirectional encoder representations from transformers (BERT), GPT, and XLNet, have shown excellent performance in many health care applications, such as clinical coding [2], named entity recognition [3], and disease prediction based on clinical notes [4]. Both BERT and GPT are advanced deep learning models that use transformer architectures, but they are fundamentally different. BERT is designed for bidirectional understanding of text, while GPT is designed for generative tasks and uses a unidirectional approach [5,6]. In particular, ChatGPT is a large language model (LLM) developed by OpenAI as an instance of GPT-3.5 that generates human-like text responses to a wide range of prompts and questions [7-9]. ChatGPT performed at or near the passing threshold of 60% accuracy on the United States Medical Licensing Examination, suggesting the potential integration into clinical decision-making [8]. Recently, the application of ChatGPT for general users seeking medical information has been highlighted [10,11].

The disparity in medical knowledge and literacy between health care professionals and the general public, often termed as information asymmetry, may inadvertently result in an inappropriate allocation of medical services due to misunderstandings or lack of awareness about health conditions [12,13]. Identifying the right outpatient specialist for their symptoms can be challenging for patients and often results in added costs and time. This is exacerbated by the current referral system, which leads to delays and increased missed clinical appointments [14,15]. Improving the process of identifying suitable medical professionals can enhance the quality of care, reduce costs, and boost patients' satisfaction [16]. To address this issue, we developed Korean Medical BERT (KM-BERT), a medical domain–specific pretrained BERT model, which was trained on a corpus of 6 million sentences from medical textbooks, health information news, and medical research papers [17]. Furthermore, we developed the fine-tuned KM-BERT model capable of recommending medical specialties based on general user queries [18].

Comparing these models can reveal which types of tasks each model is better suited to in the health care domain. For instance, one model may excel at predicting disease outcomes based on patient notes, while the other might be better at generating human-like text for health-related chatbots. In this study, we compare the performance of this model with ChatGPT and a previously developed BERT model, in line with previous research.

## Methods

### Data Collection

The previous BERT study collected 82,312 health care counsel posts from the NAVER portal, a Korean portal that provides medical questions and answers to general users [18]. The data-set was collected from the NAVER portal, a Korean portal that provides medical questions and answers to general users. The medical question involves the portal user describing their symptoms and requesting medical advice and information, which includes laboratory tests, medications, procedures, presumptive diagnoses, and recommendations for health professionals and institutions. Medical questions posted by users of the portal are reviewed and responded to by certified doctors through the portal. Each post also includes a label indicating the relevant medical specialty. The dataset consisted of questions and medical specialty label pairs. Medical specialty labels for the questions were limited to 27 clinical departments for the development of the BERT model. The original dataset was divided into a training set consisting of 50,454 data pairs and a test set comprising 31,482 data pairs. The training set was used to develop the fine-tuned KM-BERT model through 5-fold cross-validation. From the original test set, wherein data pairs were posted between July 13, 2021, and September 13, 2021, this study used 31,482 data pairs after excluding 376 due to missing data (Table 1). The medical questions asked to ChatGPT are the same as the test set (31,482 data pairs) used to develop the fine-tuned KM-BERT model.

**Table 1.** The number of test data used to measure the performance of ChatGPT and the fine-tuned Korean Medical bidirectional encoder representations from transformers (KM-BERT) model (N=31,482).

| Specialty | Value, n (%) |
|---|---|
| Anesthesiology | 1980 (6.29) |
| Cardiac and thoracic surgery | 46 (0.15) |
| Cardiology | 184 (0.58) |
| Dentistry | 1980 (6.29) |
| Dermatology | 1980 (6.29) |
| Emergency medicine | 591 (1.88) |
| Endocrinology | 169 (0.54) |
| Family medicine | 1980 (6.29) |
| Gastroenterology and hepatology | 306 (0.97) |
| General surgery | 3268 (10.38) |
| Hematology and oncology | 156 (0.50) |
| Infectious diseases | 146 (0.46) |
| Nephrology | 67 (0.21) |
| Neurology | 558 (1.77) |
| Neurosurgery | 1980 (6.29) |
| Obstetrics and gynecology | 2644 (8.40) |
| Ophthalmology | 1980 (6.29) |
| Orthopedic surgery | 1980 (6.29) |
| Otolaryngology | 1980 (6.29) |
| Pediatrics | 389 (1.24) |
| Plastic surgery | 1980 (6.29) |
| Psychiatry | 500 (1.59) |
| Pulmonology | 43 (0.14) |
| Radiology | 422 (1.34) |
| Rehabilitation medicine | 1980 (6.29) |
| Rheumatology | 213 (0.68) |
| Urology | 1980 (6.29) |

## Generating ChatGPT Medical Specialty Recommendations for Questions

ChatGPT is based on the GPT-3.5 series, and this study used "text-davinci-003" model, the latest version of the GPT-3.5 models available from the OpenAI application programming interface service at the time of the study [8,9]. ChatGPT has a better understanding of English than low-resource languages [19,20]. The questions were translated from Korean to English using the Google Translation application programming interface [21]. Previous research has also been successful in translating medical words and sentences from Korean to English [17]. ChatGPT can improve question comprehension depending on the prompting strategy [22]. To prompt ChatGPT to answer the questions, the question was appended with the sentence, "In this case, which clinical department in the hospital would be better? Please recommend 3 in order of priority."

The training corpus used for ChatGPT has not been publicly disclosed, but it is understood that it was trained on a vast amount of text data from multiple languages and sources, including Korean [8,19,23]. However, for this study, only translated sentences were used as inputs, which means they were not part of the original training samples used to develop ChatGPT. Furthermore, the original questions were randomly cross-checked to ensure that they were not indexed on Google.

## Evaluating the Performance of KM-BERT and ChatGPT

This study was conducted in strict accordance with the "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research" as published by JMIR [24]. The performance of appropriate medical specialty recommendations for medical questions from fine-tuned KM-BERT and ChatGPT was evaluated based on the same test set and 27 medical specialty labels. A confusion matrix for the 27 specialties was created to compare the first recommendation

from each model to the correct medical specialty labels and to calculate true positives, false positives, true negatives, and false negatives [25]. With an imbalance of data for each medical specialty, the performance was evaluated using macro-averaging accuracy, macro-averaging precision, macro-averaging recall, and macro-averaging $F_1$-score. The last layer of the fine-tuned KM-BERT used the softmax activation function for multiclassification, and performance was measured by comparing the first predicted medical specialty to the correct medical specialty label. The responses from ChatGPT were categorized into those that corresponded to the 27 predefined specialties and those that did not. This categorization was necessary because ChatGPT provided some responses that did not fit within the 27 predefined specialties. Out of a total of 31,482 questions, ChatGPT supplied first-rank responses corresponding to the 27 medical specialties 29,534 times (93.8%), second-rank responses 21,191 times (67.3%), and third-rank responses 19,291 times (61.3%).

### Ethical Considerations

This research project, including the original data collection, was approved by the institutional review board of Korea University Anam Hospital (2024AN0315).

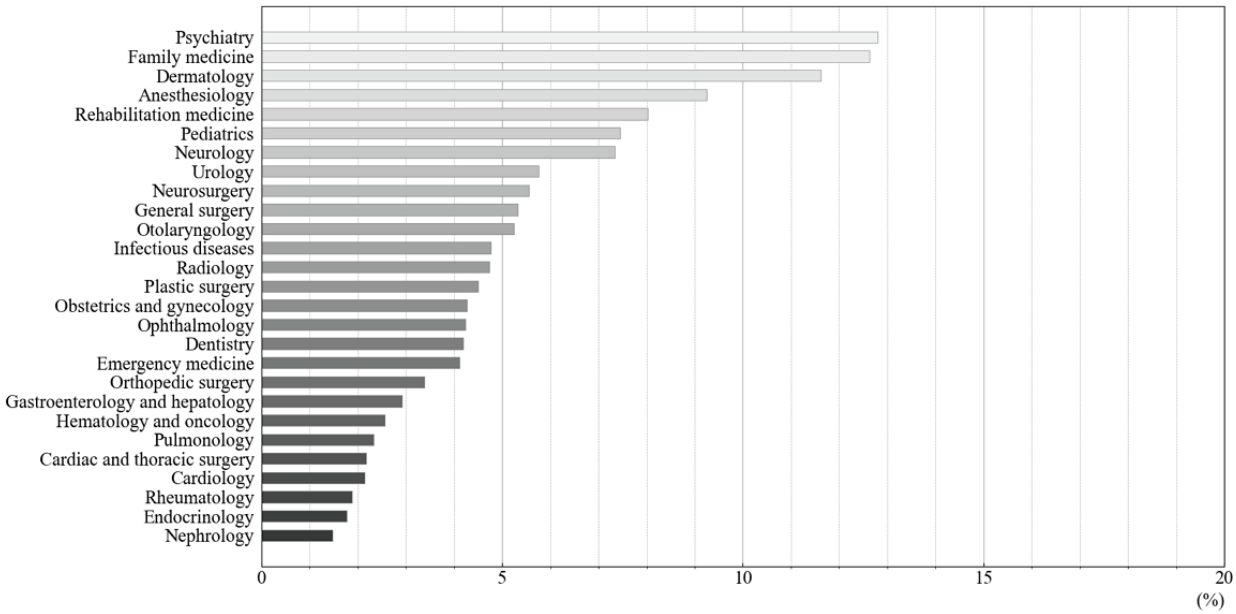
## Results

### Medical Specialty Recommendations by ChatGPT

ChatGPT was able to recommend medical specialties for 29,534 (93.8%) of the total 31,482 questions. ChatGPT declined to answer the rest of the questions (eg, “Unfortunately, I cannot answer your question as I am not a qualified medical professional and cannot provide legal advice”). The responses provided by ChatGPT covered a wide range of 1685 clinical departments, centers, clinics, hospitals, and medical specialists. However, some of the responses did not fit into the predefined 27 clinical departments, with “department of internal medicine” being a common general response. ChatGPT also provided some answers that were not classifiable, such as those relating to medical schools or hospitals that could not be categorized (eg, “Korea University College of Medicine,” “Seoul National University Bundang Hospital,” and “Johns Hopkins Hospital”). ChatGPT gave hallucinated answers relating to clinics that were not actual locations, like “K Dental Clinic” [19]. Overall, 842 of the 1685 distinct responses were able to be classified into 1 of the 27 clinical departments.

ChatGPT had an answer avoidance rate of 6.2% for inquiries regarding medical specialty recommendations. Figure 1 illustrates the response avoidance rate for each department of ChatGPT. Psychiatry had the highest avoidance rate, followed by family medicine and dermatology. On the other hand, nephrology, endocrinology, and rheumatology had the lowest avoidance rates, in that order.

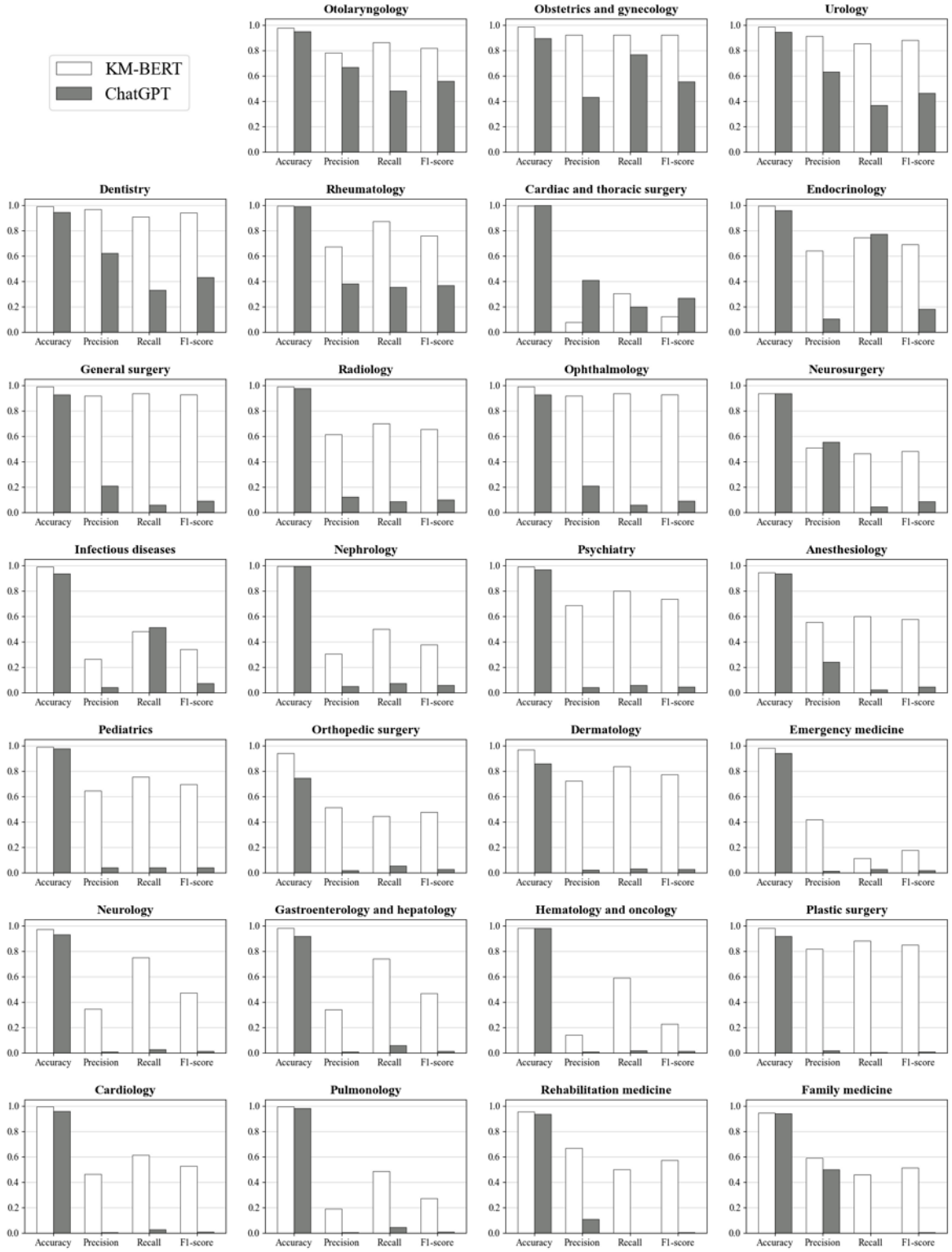
**Figure 1.** Answer avoidance rate of ChatGPT to the medical specialty recommendation.




### Performance of ChatGPT and KM-BERT

ChatGPT’s overall performance on medical specialty recommendations was lower than the fine-tuned KM-BERT model (accuracy 0.939 for ChatGPT vs 0.977 for KM-BERT, precision 0.219 for ChatGPT vs 0.570 for KM-BERT, recall 0.168 for ChatGPT vs 0.652 for KM-BERT, $F_1$-score 0.134 for ChatGPT vs 0.587 for KM-BERT). In ChatGPT, the departments with the highest $F_1$-score were otolaryngology, obstetrics and gynecology, and urology, in that order, and the departments with the lowest $F_1$-score were family medicine, rehabilitation medicine, and pulmonology (Figure 2).

**Figure 2.** Accuracy, precision, recall, and F1-score of ChatGPT and KM-BERT for each department of test set evaluation. KM-BERT: Korean Medical bidirectional encoder representations from transformers.



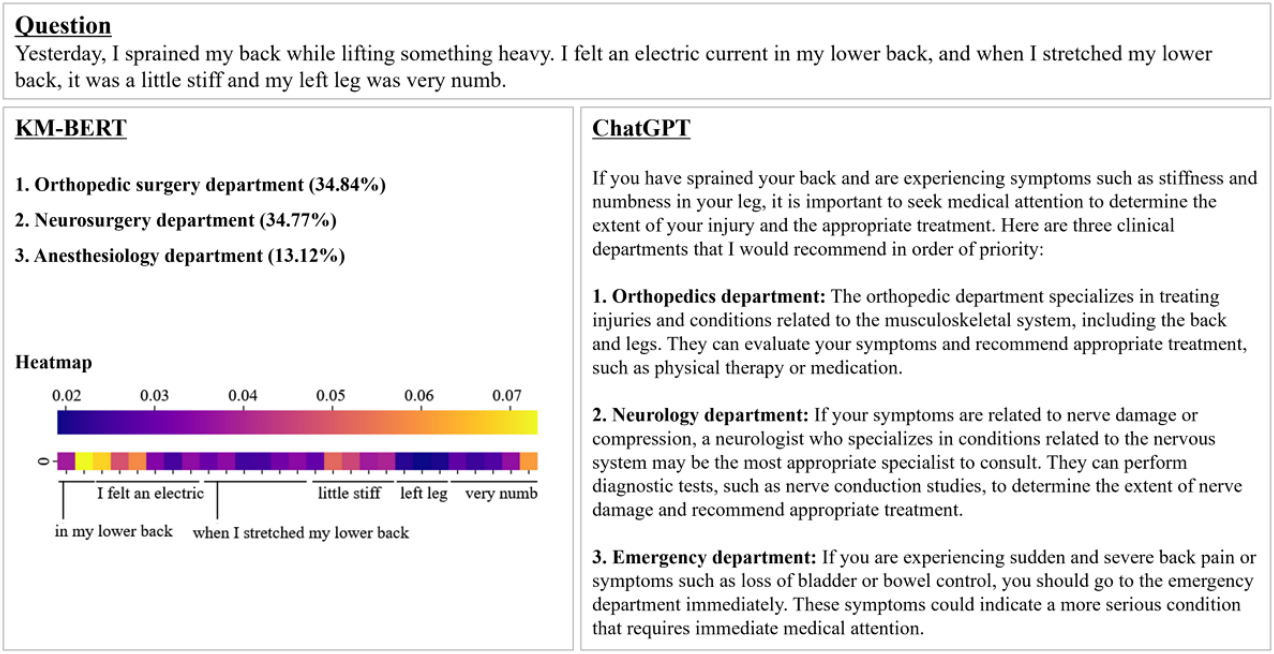

## Discussion

### Principal Findings

In the health care industry, it is crucial to provide patients with a clear justification or explanation for any artificial intelligence (AI)–based recommendations [26,27]. The growing demand for explainable AI technology in health care is consistent with this requirement [28]. ChatGPT is a significantly more advanced model than BERT in this regard [29]. For instance, when presented with the query, "Yesterday, I sprained my back while lifting something heavy. I felt an electric current in my lower back, and when I stretched my lower back, it was a little stiff and my left leg was very numb." While BERT can accurately suggest the most appropriate medical specialty in all cases, it can only offer a rough estimation by identifying the token the model is focusing on through a heatmap, etc (Figure 3). In

contrast, ChatGPT can deduce the fundamental pathophysiology of the patient's primary symptoms and provide a medical specialty recommendation accompanied by an explanation of the rationale, resulting in increased credibility and acceptance of the recommendation from the user's perspective, even if it cannot address all inquiries. This may be one of the biggest advantages of ChatGPT as a conversational language model.

**Figure 3.** Medical specialty recommendation results of ChatGPT and BERT models. Left: output information from the KM-BERT model. The BERT model reliably predicts the medical specialty based on the calculated probability. The heatmap shows the average attention for each token, which can provide insights into the model's decision-making process. The greater the brightness, the more attention. The order of the text under the heatmap has been changed as it was translated from Korean. Right: output from ChatGPT. Based on the input information, the model infers key pathophysiology and keywords from a medical perspective to recommend the appropriate medical specialty. BERT: bidirectional encoder representations from transformers; KM-BERT: Korean Medical bidirectional encoder representations from transformers.




While ChatGPT did not outperform the fine-tuned BERT model in recommending departments for health care services, it displayed numerous advantages as a conversational language model. The advantages of ChatGPT can be useful in the health care industry. First, ChatGPT can be applied to medical consultations to help patients understand medical information. Patients prefer to receive information that is written in plain language, particularly in health care, where there is an unfamiliar amount of terminology [30]. Enhancing the ability of individuals to understand and interpret the meaning of health information needed to make appropriate health decisions can improve the efficiency of the health care system [31-33]. Second, ChatGPT can assist clinicians in evaluating and diagnosing a patient's symptoms. Patients sometimes have difficulty describing their symptoms [34]. By analyzing patients' textual descriptions, ChatGPT can provide a more specific description of their symptoms, which can help clinicians better understand their patients and provide appropriate treatment [35].

The relatively poor performance of ChatGPT in this exploratory study could be attributed to the fact that the data sources used for its development were general data, mainly US-based data, with relatively little medical-specific data [20]. However, OpenAI has recently launched a fine-tuning service for ChatGPT, which is expected to significantly enhance its performance. Fine-tuning will be especially crucial since each country operates a different medical service system. As a result, we can anticipate the emergence of several ChatGPT variants fine-tuned for use in the health care industry in the future.

Finally, while ChatGPT offers incredible possibilities, concerns about the potential for generating untrue statements are growing [19,36]. As a generative model, some inaccuracies are inevitable, but they can be mitigated through fine-tuning with high-quality and reliable data resources [19,37]. It is also essential to develop and implement algorithms that can fact-check ChatGPT's statements [38]. By addressing these limitations, we can continue to explore the exciting potential of ChatGPT, ensuring that it remains a useful tool for the future of health care.

## Limitations

This study has several limitations. First, the training datasets used for the 2 models were entirely distinct. Despite the extensively large corpus upon which ChatGPT is trained, the KM-BERT model, due to its pretraining with a corpus specific to the medial domain, may exhibit superior performance in the task of medical specialty classification. Second, diverse prompting strategies can affect the classification performance of ChatGPT. A recent study revealed a comparative underperformance of contemporary LLMs against smaller, fine-tuned BERT models, particularly in a zero-shot setting [39]. Moreover, the accuracy and $F_1$-scores of LLMs differed significantly, by upwards of 10%, contingent upon the prompting strategy that is adopted. It suggests that the application of advanced prompting methodologies, such as

autogenerate prompting and chain-of-thought prompting, could potentially enhance the performance of ChatGPT in the context of this study's task [40,41]. Third, this study provides insight into the medical inference ability of ChatGPT through the medical specialty classification and a use case scenario. However, it does not extend to a quantitative evaluation of other complementary studies through objective experimentation. Notably, this study used real-world case data, not included in ChatGPT's training phase. The other previous study has also highlighted ChatGPT's capability to deduce medical symptoms, diagnoses, and treatments without explicit medical training [6]. The impact of the additional inferred information generated by ChatGPT on users' decision-making process and behavioral change necessitates further exploration.

## Conclusions

In conclusion, this study highlighted the capabilities of AI models, such as fine-tuned KM-BERT and ChatGPT, in recommending medical specialties based on general user queries. The fine-tuned KM-BERT model performed better in this task, while ChatGPT showed its strengths as a conversational AI model that can provide more context-aware responses. Future studies could aim to leverage the strengths of each model to create a more comprehensive and effective system for recommending medical specialties. This could improve the health care referral process and result in better health outcomes for patients. Moreover, with the availability of fine-tuning services for ChatGPT, we can expect the development of many more specialized AI models, potentially revolutionizing the delivery of health care information to patients.

## Acknowledgments

This research was supported by a grant of the Ministry of Science and ICT (Information and Communication Technology), Republic of Korea, under the ICT Challenge and Advanced Network of HRD program (IITP-2024-RS-2022-00156439) supervised by the Institute of Information and Communications Technology Planning and Evaluation, and a grant of the medical data–driven hospital support project through the Korea Health Information Service, funded by the Ministry of Health and Welfare, Republic of Korea.

## Data Availability

The data that support the findings of this study are available from the corresponding author, HJJ, upon reasonable request.

## Authors' Contributions

HJJ conceptualized the study and contributed to the development of the methodology. EJ and HY conducted the formal analysis. EJ used the software, conducted data curation, and handled the visualization. HJJ and EJ prepared the original draft and edited the manuscript. YMK and SS provided critical feedback and significant suggestions on the initial drafts. HY assisted with the revised drafts. HJJ and JHK provided project administration. HJJ supervised the study. All authors have read and agreed to the published version of the manuscript.

## Conflicts of Interest

None declared.

## References

1. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: A review. Trends in Anaesthesia and Critical Care. Jun 2021;38:4-9. [doi: 10.1016/j.tacc.2021.02.007]
2. Teng F, Liu Y, Li T, Zhang Y, Li S, Zhao Y. A review on deep neural networks for ICD coding. IEEE Trans. Knowl. Data Eng. 2022;35(5):4357-4375. [doi: 10.1109/tkde.2022.3148267]
3. Li J, Zhou Y, Jiang X, Natarajan K, Pakhomov S, Liu H, et al. Are synthetic clinical notes useful for real natural language processing tasks: a case study on clinical entity recognition. J Am Med Inform Assoc. 2021;28(10):2193-2201. [FREE Full text] [doi: 10.1093/jamia/ocab112] [Medline: 34272955]
4. Antikainen E, Linnosmaa J, Umer A, Oksala N, Eskola M, van Gils M, et al. Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records. Sci Rep. Mar 02, 2023;13(1):3517. [FREE Full text] [doi: 10.1038/s41598-023-30657-1] [Medline: 36864069]
5. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. Nov 19, 2022;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]
6. Nath S, Marie A, Ellershaw S, Korot E, Keane PA. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol. 2022;106(7):889-892. [doi: 10.1136/bjophthalmol-2022-321141] [Medline: 35523534]
7. Introducing ChatGPT. OpenAI. 2022. URL: https://openai.com/blog/chatgpt/ [accessed 2023-02-14]
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]

9. Wang F, Miao Q, Li X, Wang X, Lin Y. What does ChatGPT say: the DAO from algorithmic intelligence to linguistic intelligence. IEEE/CAA J. Autom. Sinica. 2023;10(3):575-579. [doi: 10.1109/jas.2023.123486]
10. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. Aesthetic Plast Surg. 2023;47(5):1985-1993. [FREE Full text] [doi: 10.1007/s00266-023-03338-7] [Medline: 37095384]
11. Seth I, Cox A, Xie Y, Bulloch G, Hunter-Smith D, Rozen W, et al. Evaluating Chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. Aesthet Surg J. 2023;43(10):1126-1135. [doi: 10.1093/asj/sjad140] [Medline: 37158147]
12. Tadayon H, Sadeqi Jabali M, Khanmohammadi MT, Rangraz Jeddi F. Information asymmetry between physicians and patients undergoing laparoscopic cholecystectomy: analysis of patients' awareness level. J Am Med Dir Assoc. 2022;23(4):703-704. [doi: 10.1016/j.jamda.2021.12.040] [Medline: 35114112]
13. Fabes J, Avşar T, Spiro J, Fernandez T, Eilers H, Evans S, et al. Health Economics Survey Group. Information asymmetry in hospitals: evidence of the lack of cost awareness in clinicians. Appl Health Econ Health Policy. 2022;20(5):693-706. [FREE Full text] [doi: 10.1007/s40258-022-00736-x] [Medline: 35606636]
14. Tong Y, Wu Y, Han Z, Xue Z, Wei Y, Lai S, et al. Development and validation of the health literacy environment scale for Chinese hospitals from patients' perspective. Front Public Health. 2023;11:1130628. [FREE Full text] [doi: 10.3389/fpubh.2023.1130628] [Medline: 37333562]
15. Brach C, Keller D, Hernandez L, Baur C, Parker R, Dreyer B, et al. Ten attributes of health literate health care organizations. NAM Perspectives. 2012;02(6):1-27. [doi: 10.31478/201206a]
16. Champlin S, Mackert M, Glowacki EM, Donovan EE. Toward a better understanding of patient health literacy: a focus on the skills patients need to find health information. Qual Health Res. 2017;27(8):1160-1176. [doi: 10.1177/1049732316646355] [Medline: 27179023]
17. Kim Y, Kim J, Lee JM, Jang MJ, Yum YJ, Kim S, et al. A pre-trained BERT for Korean medical natural language processing. Sci Rep. 2022;12(1):13847. [FREE Full text] [doi: 10.1038/s41598-022-17806-8] [Medline: 35974113]
18. Kim Y, Kim JH, Kim YM, Song S, Joo HJ. Predicting medical specialty from text based on a domain-specific pre-trained BERT. Int J Med Inform. 2023;170:104956. [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104956] [Medline: 36512987]
19. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv. Preprint published online February 2023. [FREE Full text] [doi: 10.48550/arXiv.2302.04023]
20. Zhou J, Ke P, Qiu X, Huang M, Zhang J. ChatGPT: potential, prospects, and limitations. Front Inform Technol Electron Eng. 2023;25:6-11. [doi: 10.1631/fitee.2300089]
21. de Vries E, Schoonvelde M, Schumacher G. No longer lost in translation: evidence that Google translate works for comparative bag-of-words text applications. Polit. Anal. 2018;26(4):417-430. [doi: 10.1017/pan.2018.26]
22. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health. Feb 17, 2023;13:01003. [FREE Full text] [doi: 10.7189/jogh.13.01003] [Medline: 36798998]
23. Haleem A, Javaid M, Singh RP. An era of ChatGPT as a significant futuristic support tool: a study on features, abilities, and challenges. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. Oct 2022;2(4):100089. [doi: 10.1016/j.tbench.2023.100089]
24. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323. [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
25. Krstinić D, Braović M, Šerić L, Božić-Štulić D. Multi-label classifier performance evaluation with confusion matrix. International Conference on Soft Computing, Artificial Intelligence and Machine Learning (SAIM 2020). 2020:01-14. [FREE Full text] [doi: 10.5121/csit.2020.100801]
26. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput & Applic. 2019;326(24):18069-18083. [doi: 10.1007/s00521-019-04051-w]
27. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310. [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
28. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst. 2021;32(11):4793-4813. [doi: 10.1109/TNNLS.2020.3027314] [Medline: 33079674]
29. Lee JS, Hsiang J. Patent claim generation by fine-tuning OpenAI GPT-2. World Patent Information. 2020;62:101983. [doi: 10.1016/j.wpi.2020.101983]
30. Safeer RS, Keenan J. Health literacy: the gap between physicians and patients. Am Fam Physician. 2005;72(3):463-468. [FREE Full text] [Medline: 16100861]
31. Adams RJ, Stocks NP, Wilson DH, Hill CL, Gravier S, Kickbusch I, et al. Health literacy--a new concept for general practice? Aust Fam Physician. 2009;38(3):144-147. [Medline: 19283256]

32. Kountz DS. Strategies for improving low health literacy. Postgrad Med. 2009;121(5):171-177. [doi: 10.3810/pgm.2009.09.2065] [Medline: 19820287]
33. Hironaka LK, Paasche-Orlow MK. The implications of health literacy on patient-provider communication. Arch Dis Child. 2008;93(5):428-432. [doi: 10.1136/adc.2007.131516] [Medline: 17916588]
34. Talen MR, Grampp K, Tucker A, Schultz J. What physicians want from their patients: Identifying what makes good patient communication. Families, Systems, & Health. 2008;26(1):58-66. [doi: 10.1037/1091-7527.26.1.58]
35. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2023;3(1):100105. [doi: 10.1016/j.tbench.2023.100105]
36. Alkaissi H, McFarlane S. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. 2023;15(2):e35179. [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
37. Jiang Z, Xu FF, Araki J, Neubig G. How can we know what language models know? Transactions of the Association for Computational Linguistics. 2020;8:423-438. [doi: 10.1162/tacl_a_00324]
38. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging. 2023;104(6):269-274. [FREE Full text] [doi: 10.1016/j.diii.2023.02.003] [Medline: 36858933]
39. Mu Y, Wu B, Thorne W, Robinson A, Aletras N, Scarton C. Navigating prompt complexity for zero-shot classification: a study of large language models in computational social science. ArXiv. 2305.14310:1-14. Preprint posted online on March 24, 2024. [doi: 10.48550/arXiv.2305.14310]
40. Arora S, Narayan A, Chen MF, Orr LJ, Guha N, Bhatia KS. Ask me anything: a simple strategy for prompting language models. ArXiv. 2210.02441:1-72. Preprint posted online on November 20, 2022. [doi: 10.48550/arXiv.2210.02441]
41. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems. 2022;35:24824-24837. [FREE Full text]


## Abbreviations

**AI:** artificial intelligence
**BERT:** bidirectional encoder representations from transformers
**KM-BERT:** Korean Medical bidirectional encoder representations from transformers
**LLM:** large language model


*Edited by A Mavragani; submitted 02.04.23; peer-reviewed by W-F Khaw, M Rodrigues, A Teles; comments to author 17.07.23; revised version received 03.08.23; accepted 13.08.24; published 18.10.24*

*Please cite as:*
*Jo E, Yoo H, Kim J-H, Kim Y-M, Song S, Joo HJ*
*Fine-Tuned Bidirectional Encoder Representations From Transformers Versus ChatGPT for Text-Based Outpatient Department Recommendation: Comparative Study*
*JMIR Form Res 2024;8:e47814*
*URL: https://formative.jmir.org/2024/1/e47814*
*doi: 10.2196/47814*
*PMID: 39423004*