J Korean Med Sci. 2024 Apr 29;39(16):e148
https://doi.org/10.3346/jkms.2024.39.e148
eISSN 1598-6357·pISSN 1011-8934

JKMS

Original Article
Medical Informatics

Check for updates

Patient-Friendly Discharge Summaries in Korea Based on ChatGPT: Software Development and Validation

Hanjae Kim,[1] **Hee Min Jin**,[2] **Yoon Bin Jung**,[3] **and Seng Chan You**[2]

[1]College of Nursing, Yonsei University, Seoul, Korea
[2]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea
[3]Department of Surgery, Yonsei University College of Medicine, Seoul, Korea

OPEN ACCESS

Received: Oct 9, 2023
Accepted: Mar 31, 2024
Published online: Apr 17, 2024

Address for Correspondence:
Seng Chan You, MD, PhD
Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.
Email: Chandryou@yuhs.ac

ORCID iDs
Hanjae Kim
https://orcid.org/0009-0000-4561-8400
Hee Min Jin
https://orcid.org/0009-0001-7550-2350
Yoon Bin Jung
https://orcid.org/0000-0001-9829-1931
Seng Chan You
https://orcid.org/0000-0002-5052-6399

Funding
This study was supported by a faculty research grant of Yonsei University College of Medicine (6-2023-0067).

ABSTRACT

Background: Although discharge summaries in patient-friendly language can enhance patient comprehension and satisfaction, they can also increase medical staff workload. Using a large language model, we developed and validated software that generates a patient-friendly discharge summary.
Methods: We developed and tested the software using 100 discharge summary documents, 50 for patients with myocardial infarction and 50 for patients treated in the Department of General Surgery. For each document, three new summaries were generated using three different prompting methods (Zero-shot, One-shot, and Few-shot) and graded using a 5-point Likert Scale regarding factuality, comprehensiveness, usability, ease, and fluency. We compared the effects of different prompting methods and assessed the relationship between input length and output quality.
Results: The mean overall scores differed across prompting methods (4.19 ± 0.36 in Few-shot, 4.11 ± 0.36 in One-shot, and 3.73 ± 0.44 in Zero-shot; $P < 0.001$). Post-hoc analysis indicated that the scores were higher with Few-shot and One-shot prompts than in zero-shot prompts, whereas there was no significant difference between Few-shot and One-shot prompts. The overall proportion of outputs that scored ≥ 4 was 77.0% (95% confidence interval: 68.8–85.3%), 70.0% (95% confidence interval [CI], 61.0–79.0%), and 32.0% (95% CI, 22.9–41.1%) with Few-shot, One-shot, and Zero-shot prompts, respectively. The mean factuality score was 4.19 ± 0.60 with Few-shot, 4.20 ± 0.55 with One-shot, and 3.82 ± 0.57 with Zero-shot prompts. Input length and the overall score showed negative correlations in the Zero-shot ($r = -0.437$, $P < 0.001$) and One-shot ($r = -0.327$, $P < 0.001$) tests but not in the Few-shot ($r = -0.050$, $P = 0.625$) tests.
Conclusion: Large-language models utilizing Few-shot prompts generally produce acceptable discharge summaries without significant misinformation. Our research highlights the potential of such models in creating patient-friendly discharge summaries for Korean patients to support patient-centered care.

Keywords: ChatGPT; Artificial Intelligence; Large Language Model; Patient Discharge Summaries; Patient-Centered Care; Documentation

Disclosure

You SC reports being a chief technology officer of the PHI Digital Healthcare. Other authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: You SC. Data curation: Kim H, Jin HM. Formal analysis: Kim H, Jin HM. Investigation: Kim H, Jung YB, You SC. Methodology: Kim H, Jin HM, You SC. Project administration: You SC. Resources: Jung YB, You SC. Software: Kim H. Supervision: You SC. Validation: You SC. Visualization: Kim H, Jin HM. Writing - original draft: Kim H. Writing - review & editing: Jin HM, Jung YB, You SC.

INTRODUCTION

Discharge summaries serve as a communication medium between hospitals and primary care providers.[1] Although no fixed format exists, discharge summaries typically comprise details of significant findings and provided treatments.[1] A discharge summary can also be provided to patients upon request; however, English medical terminology makes it difficult for patients to understand the document,[2] especially for those unfamiliar with English medical terminology.

Medical record documentation contributes to a high workload for clinicians. In a study by Gaffney et al.,[3] 58.1% of US office-based physicians disagreed that the time spent documenting electronic health records (EHRs) was appropriate. In a study by Tajirian et al.,[4] 74.5% of physicians and trainees (residents or fellows) in Canadian hospitals identified EHR as a contributor to burnout. As most hospitals in South Korea have adopted EHRs (97.3% of tertiary teaching hospitals and 91.4% of general hospitals[5]), clinicians in South Korea are expected to experience a significant burden of documentation.

Recently, 'ChatGPT' (GPT-3.5),[6] an artificial intelligence-based large language model (LLM) of 'OpenAI' (San Francisco, CA, USA), has gained considerable interest from medical researchers. Three weeks after its release, ChatGPT demonstrated its ability in the medical field by passing 2 out of the 3 steps of the United States Medical Licensing Examination.[7] In the study by Sarraju et al.,[8] 84% of the ChatGPT responses to cardiovascular disease prevention questions were graded as appropriate unanimously by clinicians. In another study evaluating ChatGPT responses to patient questions,[9] 78.6% of healthcare professionals preferred ChatGPT responses to physician responses and rated ChatGPT responses as of higher quality than those of physicians. ChatGPT has also shown potential for medical documentation tasks. In the study by Nayak et al.,[10] the history of present illness summaries generated by ChatGPT were graded similarly to those written by senior residents.

Providing a discharge summary written in patient-friendly language can enhance patients' understanding and satisfaction,[2,11] and inappropriate understanding may lead to various adverse outcomes.[12] However, writing an additional detailed document increases the burden on clinicians. Given its potential as an assistant in the medical field, as suggested by Patel and Lam,[13] ChatGPT could be used to help clinicians write discharge summaries. Therefore, this study aimed to develop a 'Patient-Friendly Discharge Summary-Generating Software' using ChatGPT and assess the feasibility of the software with actual clinical data.

METHODS

Development of the software

We developed a 'Patient-Friendly Discharge Summary-Generating Software' (**Fig. 1**), which generates new summaries in plain Korean with minimal medical jargon for patients based on the submitted original discharge summary. The software uses the OpenAI API to access ChatGPT. The software is built as a graphical user interface for user convenience so that users can obtain outputs by simply clicking buttons without knowledge of coding or prompting. This software has 2 key functions: summarization and replacement. The 'Summarization' function summarizes input text using simple terminologies, while the 'Replacement' function replaces medical terminologies of input text with simple terminologies while retaining the original format. The 'Replacement' function is useful when medical jargon remains in summarization outputs.

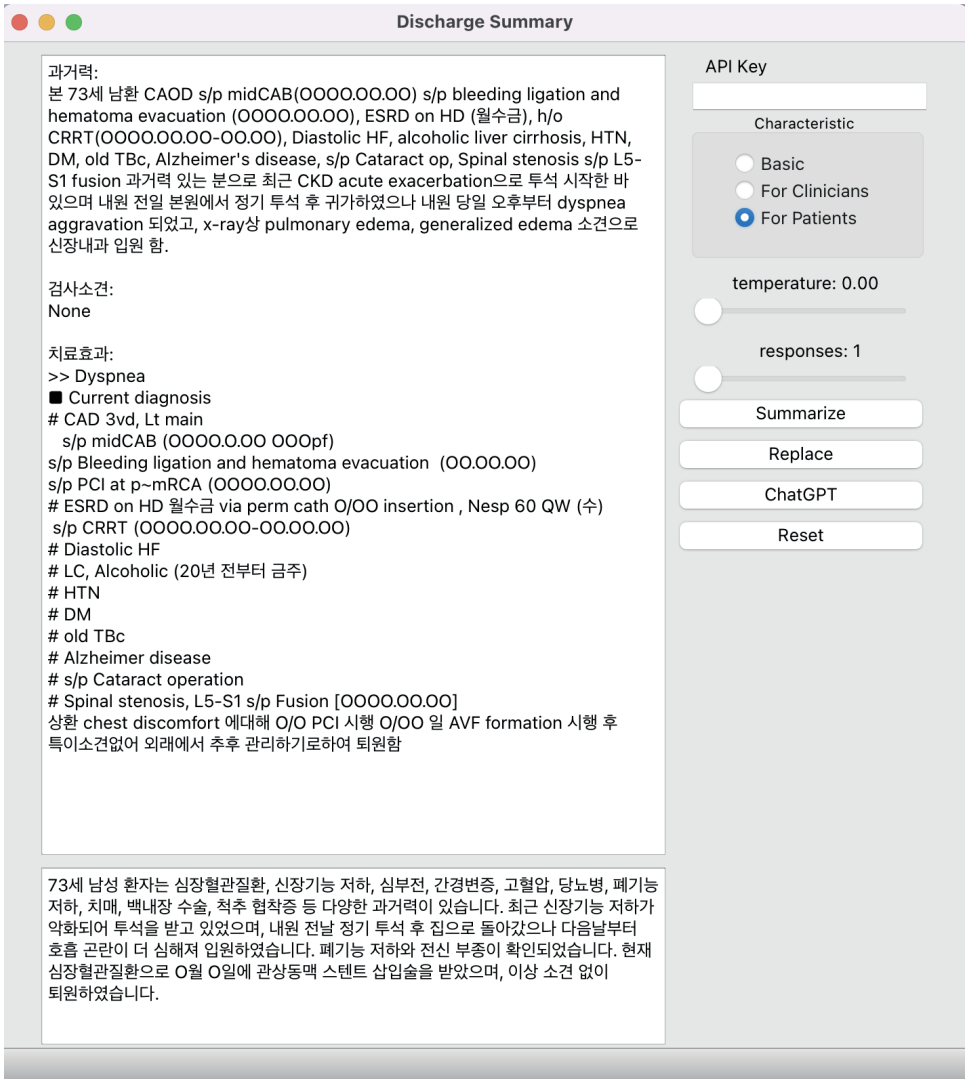

Fig. 1. Example of 'Patient-Friendly Discharge Summary-Generating Software.'

Prompt engineering

Since the software uses the OpenAI API, the experiments for this study were also conducted using the API. In addition, using the API eliminated the bias that could result from ChatGPT's ability to reference previous requests. We used the *gpt-3.5-turbo*,[14] the most capable GPT-3.5 model powering ChatGPT as of June 11, 2023 (the date of our experiment). The instructions for summarization and replacement were written in Korean. To minimize the randomness of the outputs, the temperature[15] was set as 0.

There are 3 ways to prompt engineering: Zero-shot, One-shot, and Few-shot.[16] In the Zero-shot prompting method, the model is given natural language instructions without examples or demonstrations. In the One-shot prompting method, the model is provided instructions using a single example. The model is provided with instructions and multiple examples in the Few-shot prompting method. In this study, the model was provided with 2 examples in the Few-shot prompting method. Details of the instructions and examples used in this study are provided in **Supplementary Table 1**.

Data collection

Our study's target sample size was 246, calculated by G*Power 3.1.9.6, aiming for 80% power at a 5% significance level to detect an effect size of .20 among Zero-shot, One-shot, and Few-shot prompts using one-way analysis of variance (ANOVA). The discharge summary documents of patients diagnosed with myocardial infarction (MI) and those treated at the Department of General Surgery (GS) were collected from Severance Hospital (Seoul, Korea). From 11,698 documents generated between March 1, 2022, and February 28, 2023, 100 documents (50 each from MI and GS) were randomly sampled. All documents were de-identified by masking personal information to protect privacy. Each document comprised past medical history, examination findings, and treatment outcomes (**Supplementary Table 2**) and underwent consecutive steps of summarization and replacement (**Fig. 2**). Example outputs of each step are provided in **Supplementary Table 3**. Three responses were generated for each document using 3 different prompting methods (Zero-shot, One-shot, and Few-shot), resulting in 300 output samples (a new summary). All responses were generated on June 11, 2023.

Evaluation design

The outputs were evaluated by 4 medical (2 doctors and 2 nurses) and 2 non-medical personnel. Medical personnel were provided with original documents and 3 outputs (with Zero-shot, One-shot, and Few-shot prompts), while non-medical personnel were provided only with outputs. The outputs for each document were randomly ordered to prevent the evaluators from identifying the prompting method used. Medical personnel were instructed

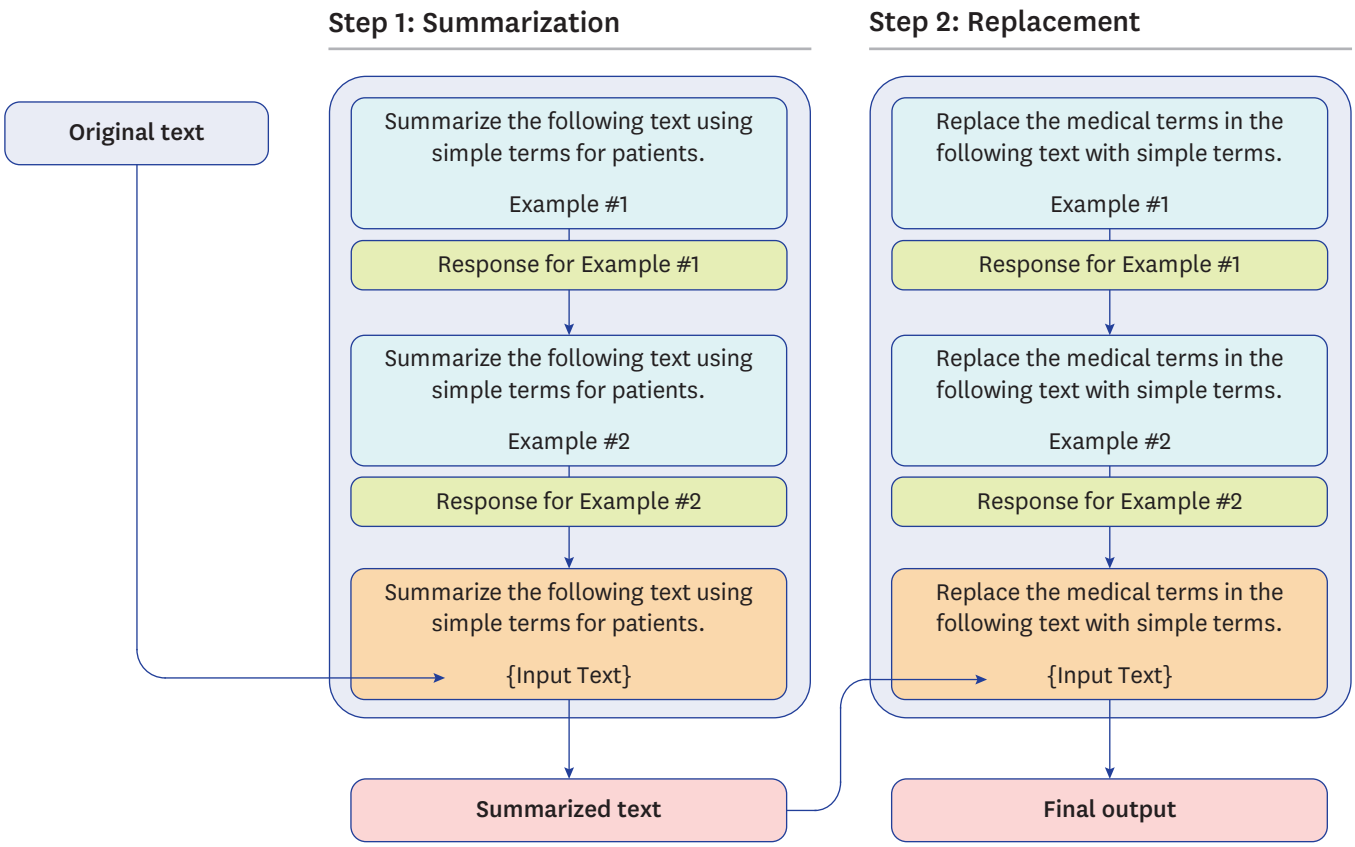

Fig. 2. Prompt design. The actual prompts were typed in Korean, as instructions in English resulted in English output in some cases. **Supplementary Table 1** shows details of the instructions and examples of each step.

to grade the outputs in terms of factuality (whether the original document supported the contents), comprehensiveness (whether the essential contents were well included), and usability (whether the output was good enough to be provided to the patient). Non-medical personnel graded the outputs regarding ease (whether non-medical people could easily understand the output) and fluency (whether the sentences were linguistically acceptable). All metrics were graded using 5-point Likert Scales, with higher scores indicating better output quality. The specific criteria for each metric are listed in **Supplementary Table 4**. Intraclass correlation coefficient (ICC) estimates of medical personnel evaluations were calculated using SPSS statistical package version 27 (IBM Corp., Armonk, NY, USA) based on a mean-rating (k = 4), consistency, 2-way random-effects model. Additionally, all evaluators were instructed to choose their preferred one among the 3 outputs of each original document.

Statistical analysis

Factuality, comprehensiveness, and usability scores were averaged across the evaluations by medical personnel, and ease and fluency scores were averaged across the evaluations by non-medical personnel. The overall score was calculated by averaging the scores of all 5 metrics. A one-way ANOVA and post-hoc analysis (Tukey's test) were performed on each metric to compare the effects of the different prompting methods on output quality.

Moreover, because some studies assert that a longer input leads to a lower performance of ChatGPT on the summarization task,[17,18] we computed the Pearson correlation coefficient to assess the relationship between the token count of the original document and the scores for each metric. Tokens are chunks of characters broken down from the input text so language models can process the prompts.[19] The token count of each discharge summary was computed using tiktoken (Python package) version 0.3.0. One outlier was removed, and a square root transformation was applied to normalize the distribution of the token counts.

Finally, we compared the mean scores of the MI and GS outputs using two-tailed t-tests. Since MI and GS were randomly selected from various medical departments, we did not anticipate significant differences between them. Furthermore, as all the examples with the One-shot and Few-shot prompts were data from patients with MI, we wanted to examine whether they biased the results. All statistical analyses except ICC calculation were performed using Python version 3.10.5 with Numpy version 1.24.2, Pandas version 1.5.3, SciPy version 1.11.1, Statsmodels version 0.14.0, and Pingouin version 0.5.3.

Ethics statement

This study was approved by the Institutional Review Board (IRB) of Severance Hospital (IRB approval number: 4-2023-0441) on June 8, 2023. The requirement to obtain written consent was waived because this study used de-identified patient data.

RESULTS

The total number of output samples was 300, including 50 MI and 50 GS discharge summaries, which were rewritten using Zero-, One-, and Few-shot prompts. For each output, 4 medical personnel graded factuality, comprehensiveness, and usability, and 2 non-medical personnel graded ease of use and fluency. There were no missing data. ICC estimates of medical personnel evaluations were 0.565 (95% confidence interval [CI], 0.479–0.640; $P < 0.001$) in factuality, 0.694 (95% CI, 0.634–0.747; $P < 0.001$) in comprehensiveness,

0.693 (95% CI, 0.632–0.746; $P < 0.001$) in usability, and 0.709 (95% CI, 0.651–0.759; $P < 0.001$) for averaged score of the three metrics. In addition, all 6 evaluators chose 1 of 3 different outputs for each discharge summary, resulting in 600 of the most preferred outputs.

Evaluation of software

The mean overall score for outputs was 3.73 ± 0.44 with Zero-shot prompts (Factuality: 3.82 ± 0.57; Comprehensiveness: 3.68 ± 0.70; Usability: 3.36 ± 0.65; Ease: 4.04 ± 0.58; Fluency: 3.77 ± 0.73), 4.11 ± 0.36 with One-shot prompts (Factuality: 4.20 ± 0.55; Comprehensiveness: 4.08 ± 0.64; Usability: 3.93 ± 0.59; Ease: 4.25 ± 0.49; Fluency: 4.11 ± 0.51), and 4.19 ± 0.36 with Few-shot prompts (Factuality: 4.19 ± 0.60; Comprehensiveness: 4.18 ± 0.59; Usability: 3.97 ± 0.59; Ease: 4.39 ± 0.45; Fluency: 4.22 ± 0.58) (**Table 1**). The output with the highest overall score was generated from Few-shot prompts, scoring 4.90, and the output with the lowest overall score was generated from Zero-shot prompts, scoring 2.70 (**Supplementary Table 5**).

One-way ANOVA for method comparison

One-way ANOVA revealed a statistically significant difference in output scores among the different prompting methods for all metrics (Factuality: F = 14.56, $P < 0.001$; Comprehensiveness: F = 16.43, $P < 0.001$; Usability: F = 31.38, $P < 0.001$; Ease: F = 11.58, $P < 0.001$; Fluency: F = 14.57, $P < 0.001$; Overall: F = 39.38, $P < 0.001$). Post-hoc comparisons using Tukey's test indicated that the outputs from Few-shot and One-shot prompts were graded higher than those from Zero-shot prompts in all metrics. Output scores were not significantly different between the Few-shot and One-shot prompts (**Table 1**).

Proportions of score intervals and preferred methods

The proportion of each score interval was calculated to assess the general performance of the model (**Fig. 3A**). The overall proportion of outputs that scored ≥ 4 was 77.0% (95% CI, 68.8–85.3%) with the Few-shot, 70.0% (95% CI, 61.0–79.0%) with the One-shot, and 32.0% (95% CI, 22.9–41.1%) with the Zero-shot prompts. Meanwhile, the proportion of outputs that scored < 3 overall was 0% with both Few-shot and One-shot and 2.0% (95% CI, −0.7–4.7%)

Table 1. Comparison of output qualities among different prompting methods (N = 300)

Metrics	Methods	Mean ± SD	95% CI	F	*P* value	*Post-hoc* Tukey's test
Factuality	Zero-shot[a]	3.82 ± 0.57	3.70–3.93	14.56	< 0.001***	b > a***
	One-shot[b]	4.20 ± 0.55	4.09–4.31			c > a***
	Few-shot[c]	4.19 ± 0.60	4.07–4.31			
Comprehensiveness	Zero-shot[a]	3.68 ± 0.70	3.54–3.82	16.43	< 0.001***	b > a***
	One-shot[b]	4.08 ± 0.64	3.95–4.21			c > a***
	Few-shot[c]	4.18 ± 0.59	4.06–4.29			
Usability	Zero-shot[a]	3.36 ± 0.65	3.23–3.49	31.38	< 0.001***	b > a***
	One-shot[b]	3.93 ± 0.59	3.81–4.04			c > a***
	Few-shot[c]	3.97 ± 0.59	3.86–4.09			
Ease	Zero-shot[a]	4.04 ± 0.58	3.92–4.16	11.58	< 0.001***	b > a*
	One-shot[b]	4.25 ± 0.49	4.15–4.35			c > a***
	Few-shot[c]	4.39 ± 0.45	4.30–4.47			
Fluency	Zero-shot[a]	3.77 ± 0.73	3.62–3.92	14.57	< 0.001***	b > a***
	One-shot[b]	4.11 ± 0.51	4.01–4.21			c > a***
	Few-shot[c]	4.22 ± 0.58	4.10–4.34			
Overall	Zero-shot[a]	3.73 ± 0.44	3.65–3.82	39.38	< 0.001***	b > a***
	One-shot[b]	4.11 ± 0.36	4.04–4.19			c > a***
	Few-shot[c]	4.19 ± 0.36	4.12–4.26			

The overall score was calculated by averaging the factuality, comprehensiveness, usability, ease, and fluency scores.
SD = standard deviation, CI = confidence interval.
*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

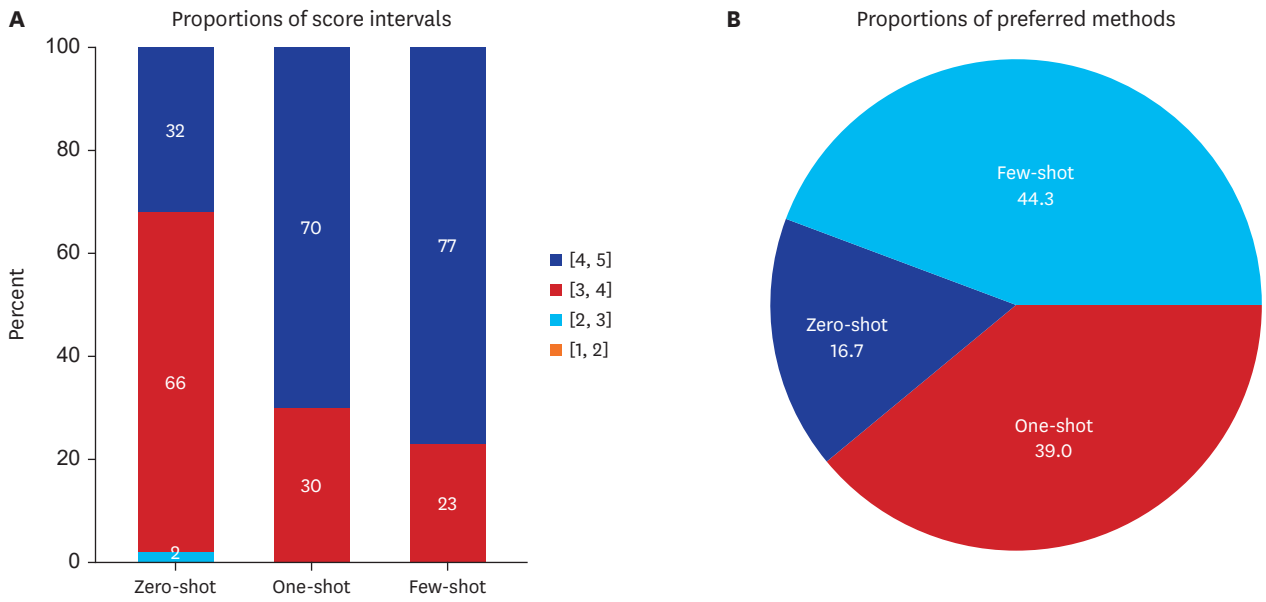

Fig. 3. Proportions of score intervals and preferred methods. (**A**) Proportions of score intervals. (**B**) Proportions of preferred methods.

with Zero-shot prompts. As a result of the evaluators choosing the output they preferred, Few-shot prompting method was chosen as the most preferred output in 44.3% (95% CI, 40.4–48.3%) of the 600 evaluations. The preference for One-shot prompting method was 39.0% (95% CI, 35.1–42.9%), and that for Zero-shot prompting method was 16.7% (95% CI, 13.7–19.6%) (**Fig. 3B**).

Effect of input length

The tokens for each original discharge summary were calculated as text length indicators. The mean token count was 375 ± 245. After normalizing by removing an outlier and applying square root transformation, the new mean value was 18 ± 5. Pearson's correlation coefficient showed a negative correlation between normalized token count and overall output score with Zero-shot ($r = -0.437$, $P < 0.001$) and One-shot ($r = -0.327$, $P < 0.001$). There was no significant correlation between token counts and output quality with Few-shot prompts ($r = -0.050$, $P = 0.625$). More specifically, scores of factuality, comprehensiveness, usability, and ease with Zero-shot prompts and comprehensiveness, usability, and fluency with One-shot prompts showed negative correlations with the normalized token count. None of the metrics in the Few-shot test showed a significant correlation with normalized token count (**Table 2**).

Table 2. Correlation between the token count of input text and output quality (N = 297)

Metrics	Tokens		
	Zero-shot	One-shot	Few-shot
Factuality	−0.198*	−0.096	0.095
Comprehensiveness	−0.457***	−0.307**	−0.074
Usability	−0.451***	−0.278**	−0.096
Ease	−0.377***	−0.153	−0.037
Fluency	−0.022	−0.203*	−0.056
Overall	−0.437***	−0.327***	−0.050

The overall score was calculated by averaging the factuality, comprehensiveness, usability, ease, and fluency scores.
*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

Table 3. Comparison of the mean overall scores of outputs between patient groups (N = 300)

Methods	MI (n = 50), Mean ± SD	GS (n = 50), Mean ± SD	t	*P* value
Zero-shot	3.65 ± 0.43	3.82 ± 0.44	−1.95	0.054
One-shot	4.07 ± 0.38	4.16 ± 0.34	−1.30	0.198
Few-shot	4.26 ± 0.33	4.12 ± 0.38	1.89	0.062

The overall score was calculated by averaging the factuality, comprehensiveness, usability, ease, and fluency scores.
MI = myocardial infarction, GS = general surgery, SD = standard deviation.
$^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Comparison of patient groups

A two-tailed t-test was conducted to compare the mean overall scores of the MI and GS groups for each prompting method. There was no significant difference in the mean overall scores between MI and GS with Few-shot (MI = 4.26 ± 0.33, GS = 4.12 ± 0.38; t = 1.89, $P = 0.062$), One-shot (MI = 4.07 ± 0.38, GS = 4.16 ± 0.34; t = −1.30, $P = 0.198$), and Zero-shot prompts (MI = 3.65 ± 0.43, GS = 3.82 ± 0.44; t = −1.95, $P = 0.054$) (**Table 3**).

DISCUSSION

In this study, we developed a software that generates a patient-friendly discharge summary using the GPT-3.5 API and evaluated its performance. The output qualities of the Few-shot and One-shot methods were acceptable, achieving mean overall scores higher than 4 out of 5. Few-shot and One-shot prompting methods generated significantly better outputs than Zero-shot prompting method, and none of the outputs with Few-shot or One-shot prompts was graded below 3 for the overall score, indicating a minimal probability of unacceptably poor quality. This demonstrates that LLMs, such as ChatGPT, with adequate prompt engineering, can assist clinicians by generating patient-friendly discharge summaries. There was no notable difference in the performance between patients with MI and those treated in GS, which implies the generalizability of our system under various conditions.

This study is pioneering in its exploration of the potential of a LLM for creating patient-friendly discharge summaries in Korean, marking the first investigation into its application in routine clinical care. It is crucial to provide information to patients and help them understand their health in patient-centered care.[20,21] With our software, clinicians can easily rewrite new discharge summaries in patient-friendly language. The mean ease score for outputs with the Few-shot prompts was 4.39 ± 0.45, indicating that new summaries are easily understandable for people without knowledge of English medical terminology.

Another significance of this study is the utilization of One-shot and Few-shot prompting methods. To the best of our knowledge, this is the first study in Korean medical text to compare the differences among Zero-shot, One-shot, and Few-shot prompting methods through a statistical analysis. In this study, outputs with Few-shot and One-shot prompts were graded significantly higher than those with Zero-shot prompts in all metrics, while there was no significant difference in the output scores between Few-shot and One-shot. As both methods showed similar performances, the One-shot could be considered more reasonable because it requires only one example. However, the overall score showed a negative correlation with the token count of the original text in the One-shot, whereas there was no significant correlation between the 2 variables in the Few-shot. This indicates that a longer input text can negatively affect the performance of the software with the One-shot prompts, whereas a Few-shot prompting method shows consistent performance regardless of input length.

Hallucination, especially extrinsic hallucination, where statements cannot be verified from the input text, is a common problem in the summarization and translation task of language models.[22,23] Since there is evidence that automatic metrics (such as ROUGE,[24] BERTScore,[25] METEOR,[26] and BLEU[27]) cannot effectively evaluate the factuality of text summarization,[17,22] we conducted a human evaluation of the factuality score to assess the level of hallucination in our study. The factuality score was ≥ 4 in the majority of cases in Few-shot (67%) and One-shot (70%), indicating no notable hallucination. Only 46% of the Zero-shot cases had a factuality score of ≥ 4. The mean factuality score was also significantly higher in Few-shot (4.19 ± 0.60) and One-shot (4.20 ± 0.55) than in Zero-shot (3.82 ± 0.57), which shows that the risk for hallucination can be reduced with more detailed prompt engineering. Nevertheless, as the factuality score is not always 5, further techniques should be explored to find a way to improve the reliability of our software. For example, more detailed instructions can be provided. In a study by Nayak et al.,[10] errors in the history of present illness generated by ChatGPT were reduced by improving prompt quality. Furthermore, several studies have been conducted on detecting hallucination of LLMs.[28] Implementing a final review step focusing on detecting hallucination could be another solution.

Besides hallucination, LLMs possess more notable weaknesses. Datasets used for training LLMs are normally not domain-specific and lack reliability since unverified internet sources are included.[29-31] This could lead LLMs to generate inaccurate information in medical domain of specific matters.[29,32] To address this issue, LLMs can be fine-tuned with medical data. 'MedPaLM'[33] and 'MedPaLM2'[34] were built upon PaLM[35] and PaLM2,[36] respectively, trained with medical question answering data by a method called 'instruction prompt-tuning,' resulting in improvement compared to basic PaLM models in answering medical questions. Moreover, Li et al.[37] fine-tuned GPT-2 with tissue data from cancer and the resulting model, 'CancerGPT,' achieved the comparable accuracy to GPT-3 in predicting drug pair synergy for different tissue types.

Another problem of LLMs is lack of recent information since they are not trained in real-time.[31,38] Therefore, LLMs wouldn't be able to translate discharge summaries properly if coined terms are included in the text. Connecting external data to LLMs using vector database could be a key to solve this problem, as it enables semantic search of relevant documents through embeddings. Rau et al.[39] connected GPT-3.5 to embeddings created with radiologic imaging guidelines. The resulting model, accGPT, was superior to GPT-3.5 and GPT-4 in providing radiologic imaging recommendation. Using the same method, Russe et al.[40] proposed 'FraCChat' built upon GPT-3.5 and GPT-4 with fracture classification criteria and the model outperformed generic GPT models in fracture classification.

This study had several limitations. First, the Few-shot prompting method generally performs better than the One-shot and Zero-shot.[16] However, in this study, the outputs with Few-shot prompts were similar to those with One-shot prompts across all metrics. We provided only 2 examples of Few-shot prompting that could have contributed to this result. *Gpt-3.5-turbo-16k*,[14] the new GPT-3.5 model released 2 days after our experiment, had a maximum token limit of 16,384. Considering that the mean token count of our input discharge summary was 375 ± 245, we hope that further research will be conducted using more examples. Second, GPT-4,[41] the latest model of OpenAI, or other LLMs such as 'Bard'[42] were not considered in this study. Until now, Bard refused to respond to prompts with medical content on our experimental data. Furthermore, a recent study reported that the performance of GPT-4 declines in various tasks, whereas the performance of GPT-3.5 has been improving.[43]

Third, the evaluation of the model was conducted by human evaluators with Likert Scales, which could possibly be affected by subjective views. Following studies should adopt additional methods to obtain more objectivity or reliability. Fourth, it is uncertain whether similar results would be observed in languages other than Korean, as LLMs perform poorly when handling languages with limited resources.[44] Fifth, although the temperature was set to 0, the reproducibility of this study is not assured as ChatGPT does not always respond identically. Finally, there is a risk of patients' private information being leaked to commercial enterprises if LLMs are used in actual hospitals. In the case of ChatGPT, OpenAI retains any data submitted through its API for a maximum of 30 days.[45] In this study, we used only de-identified discharge summaries under IRB approval. Clinicians should avoid submitting patient information to LLMs until clear guidelines are established. Nonetheless, our study demonstrated the potential of using LLMs to generate patient-friendly discharge summaries.

In conclusion, our software can effectively transform discharge summaries into new texts that patients can easily understand. The software performed best with Few-shot prompts, and examples from 1 specific specialty seemed to function moderately in other specialties. This study demonstrates that LLMs, such as ChatGPT, can be valuable assistants for writing patient-friendly discharge summaries and could be a critical factor in patient-centered care. We anticipate further studies to be conducted with more various models, methods, and sample data to improve the quality of discharge summaries generated by LLMs.

ACKNOWLEDGMENTS

The authors thank following contributors for grading outputs of our software.

Yeonjae Han, RN, Department of Biomedical Systems Informatics, Yonsei University College of Medicine.
Dami Jung, RN, Severance Hospital, Yonsei University Health System.
Junu Kim, BS, Kim Jaechul Graduate School of AI, KAIST.
Sunjun Kweon, MS, Kim Jaechul Graduate School of AI, KAIST.

SUPPLEMENTARY MATERIALS

Supplementary Table 1
Prompt used for Few-shot

Supplementary Table 2
Example of original summary and new summaries

Supplementary Table 3
Example outputs of summarization step and replacement step

Supplementary Table 4
Criteria for evaluation

Supplementary Table 5
Outputs with the highest and the lowest overall score

REFERENCES

1. Sorita A, Robelia PM, Kattel SB, McCoy CP, Keller AS, Almasri J, et al. The ideal hospital discharge summary: a survey of U.S. physicians. *J Patient Saf* 2021;17(7):e637-44. PUBMED | CROSSREF
2. Lin R, Gallagher R, Spinaze M, Najoumian H, Dennis C, Clifton-Bligh R, et al. Effect of a patient-directed discharge letter on patient understanding of their hospitalisation. *Intern Med J* 2014;44(9):851-7. PUBMED | CROSSREF
3. Gaffney A, Woolhandler S, Cai C, Bor D, Himmelstein J, McCormick D, et al. Medical documentation burden among US office-based physicians in 2019: a national study. *JAMA Intern Med* 2022;182(5):564-6. PUBMED | CROSSREF
4. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. *J Med Internet Res* 2020;22(7):e19274. PUBMED | CROSSREF
5. Kim YG, Jung K, Park YT, Shin D, Cho SY, Yoon D, et al. Rate of electronic health record adoption in South Korea: a nation-wide survey. *Int J Med Inform* 2017;101:100-7. PUBMED | CROSSREF
6. OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt. Updated 2022. Accessed August 14, 2023.
7. 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. PUBMED | CROSSREF
8. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023;329(10):842-4. PUBMED | CROSSREF
9. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183(6):589-96. PUBMED | CROSSREF
10. Nayak A, Alkaitis MS, Nayak K, Nikolov M, Weinfurt KP, Schulman K. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. *JAMA Intern Med* 2023;183(9):1026-7. PUBMED | CROSSREF
11. Cook JL, Fioratou E, Davey P, Urquhart L. Improving patient understanding on discharge from the short stay unit: an integrated human factors and quality improvement approach. *BMJ Open Qual* 2022;11(3):e001810. PUBMED | CROSSREF
12. Newnham H, Barker A, Ritchie E, Hitchcock K, Gibbs H, Holton S. Discharge communication practices and healthcare provider and patient preferences, satisfaction and comprehension: a systematic review. *Int J Qual Health Care* 2017;29(6):752-68. PUBMED | CROSSREF
13. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health* 2023;5(3):e107-8. PUBMED | CROSSREF
14. OpenAI platform. Models. https://platform.openai.com/docs/models. Updated 2023. Accessed August 14, 2023.
15. OpenAI platform. Quickstart. https://platform.openai.com/docs/quickstart. Updated 2023. Accessed August 14, 2023.
16. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020;33:1877-901.
17. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. *NPJ Digit Med* 2023;6(1):158. PUBMED | CROSSREF
18. Deroy A, Ghosh K, Ghosh S. How ready are pre-trained abstractive models and LLMs for legal case judgement summarization? *ArXiv*. June 14, 2023. https://doi.org/10.48550/arXiv.2306.01248. CROSSREF
19. OpenAI platform. Introduction. https://platform.openai.com/docs/introduction. Updated 2023. Accessed August 14, 2023.
20. Mead N, Bower P. Patient-centredness: a conceptual framework and review of the empirical literature. *Soc Sci Med* 2000;51(7):1087-110. PUBMED | CROSSREF
21. Epstein RM, Street RL Jr. The values and value of patient-centered care. *Ann Fam Med* 2011;9(2):100-3. PUBMED | CROSSREF
22. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv* 2023;55(12):1-38. CROSSREF
23. 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.* November 28, 2023. https://doi.org/10.18653/v1/2023.ijcnlp-main.45. CROSSREF

24. Lin CY, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics; May 27, 2003-June 1, 2003; Edmonton, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics; 2003, 150-7.
25. Hanna M, Bojar O. A fine-grained analysis of BERTScore. In: Barrault L, Bojar O, Bougares F, Chatterjee R, Costa-jussa MR, Federmann C, et al., editors. *Proceedings of the Sixth Conference on Machine Translation*. Stroudsburg, PA, USA: Association for Computational Linguistics; 2021, 507-17.
26. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; June 2005; Ann Arbor, MI, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2005, 65-72.
27. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; July 6-12, 2002; Philadelphia, PA, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2002, 311-8.
28. Luo J, Li T, Wu D, Jenkin M, Liu S, Dudek G. Hallucination detection and hallucination mitigation: an investigation. *ArXiv.* January 16, 2024. https://doi.org/10.48550/arXiv.2401.08358.
29. Kim S. In the era of ChatGPT, can medical artificial intelligence replace the doctor? *Korean J Med* 2023;98(3):99-101. CROSSREF
30. Doskaliuk B, Zimba O. Beyond the keyboard: academic writing in the era of ChatGPT. *J Korean Med Sci* 2023;38(26):e207. PUBMED | CROSSREF
31. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nat Med* 2023;29(8):1930-40. PUBMED | CROSSREF
32. Preiksaitis C, Sinsky CA, Rose C. ChatGPT is not the solution to physicians' documentation burden. *Nat Med* 2023;29(6):1296-7. PUBMED | CROSSREF
33. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023;620(7972):172-80. PUBMED | CROSSREF
34. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. *ArXiv*. May 16, 2023. https://doi.org/10.48550/arXiv.2305.09617. CROSSREF
35. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: scaling language modeling with pathways. *J Mach Learn Res* 2023;24(240):1-113.
36. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. Palm 2 technical report. *ArXiv*. September 13, 2023. https://doi.org/10.48550/arXiv.2305.10403. CROSSREF
37. Li T, Shetty S, Kamath A, Jaiswal A, Jiang X, Ding Y, et al. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *NPJ Digit Med* 2024;7(1):40. PUBMED | CROSSREF
38. Deik A. Potential benefits and perils of incorporating ChatGPT to the movement disorders clinic. *J Mov Disord* 2023;16(2):158-62. PUBMED | CROSSREF
39. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology* 2023;308(1):e230970. PUBMED | CROSSREF
40. Russe MF, Fink A, Ngo H, Tran H, Bamberg F, Reisert M, et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. *Sci Rep* 2023;13(1):14215. PUBMED | CROSSREF
41. OpenAI. GPT-4. https://openai.com/research/gpt-4. Updated 2023. Accessed August 14, 2023.
42. Google. Bard. https://bard.google.com. Updated 2023. Accessed August 14, 2023.
43. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? *ArXiv*. October 31, 2023. https://doi.org/10.48550/arXiv.2307.09009. CROSSREF
44. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. *ACM Trans Intell Syst Technol* 2024;15(3):39. CROSSREF
45. OpenAI. API data usage policies. https://openai.com/policies/api-data-usage-policies. Updated 2023. Accessed August 14, 2023.