

Evidence-based advanced prompt engineering in nursing research: quality analysis of ChatGPT-generated Boolean search query

Oparta na dowodach zaawansowana inżynieria zapytań w badaniach pielęgniarskich:
Analiza jakości zaawansowanej strategii wyszukiwania Boole'a generowanej przez ChatGPT

Joanna Gotlib-Małkowska^{A,B,E-F,I} , Ilona Cieślak^{B,D-E,K} , Mariusz Jaworski^{E,G} ,
Mariusz Panczyk^{A,E,G,L} 

Department of Education and Research in Health Sciences, Faculty of Health Sciences, Medical University of Warsaw, Polska

CORRESPONDING AUTHOR

Ilona Cieślak

Department of Education and Research in Health Sciences, Faculty of Health Sciences, Medical University of Warsaw, Poland
Żwirki i Wigury 61, 02-091 Warsaw, Poland
e-mail: ilona.cieslak@wum.edu.pl

A – Development of the concept and methodology of the study/Opracowanie koncepcji i metodologii badań; B – Query - a review and analysis of the literature/Kwerenda – przegląd i analiza literatury przedmiotu; C – Submission of the application to the appropriate Bioethics Committee/Złożenie wniosku do właściwej Komisji Biotycznej; D – Collection of research material/Gromadzenie materiału badawczego; E – Analysis of the research material/Analiza materiału badawczego; F – Preparation of draft version of manuscript/Przygotowanie roboczej wersji artykułu; G – Critical analysis of manuscript draft version/Analiza krytyczna roboczej wersji artykułu; H – Statistical analysis of the research material/Analiza statystyczna materiału badawczego; I – Interpretation of the performed statistical analysis/Interpretacja dokonanej analizy statystycznej; K – Technical preparation of manuscript in accordance with the journal regulations/Opracowanie techniczne artykułu zgodne z regulaminem czasopisma; L – Supervision of the research and preparation of the manuscript/Nadzór nad przebiegiem badań i przygotowaniem artykułu

STRESZCZENIE

OPARTA NA DOWODACH ZAAWANSOWANA INŻYNIERIA ZAPYTAŃ W BADANIACH PIELĘGNIARSKICH: ANALIZA JAKOŚCI ZAAWANSOWANEJ STRATEGII WYSZUKIWANIA BOOLE'A GENEROWANEJ PRZEZ CHATGPT

Cel pracy. W artykule zbadano możliwość wykorzystania zaawansowanej inżynierii podpowiedzi w badaniach z obszaru pielęgniarstwa, ze szczególnym uwzględnieniem zapytań Boole'a (BSQ) generowanych przez ChatGPT.

Materiał i metody. W badaniu porównano skuteczność różnych modeli ChatGPT: ChatGPT-3.5, ChatGPT-4.0 i ChatGPT-4omni, w generowaniu wysokiej jakości zapytań BSQ dla bazy PUBMED. Analizowane metody podpowiedzi obejmowały Zero-Shot, Automated Chain-Of-Thought, Emotional Stimuli, Role-play i Mixed-Methods prompting.

Wyniki. Badanie wykazało, że ChatGPT-4omni, przy wykorzystaniu podpowiedzi Mixed-Methods, osiągnął najwyższą jakość udzielanych odpowiedzi, podczas gdy ChatGPT-3.5, wykorzystujący podpowiedzi zero-shot, jest najmniej skuteczny. Zaobserwowano znaczną zmienność wyników wyszukiwania w różnych modelach i metodach podpowiadania. Autorzy zalecają ChatGPT-4omni jako najskuteczniejszy model do generowania BSQ.

Wnioski. Badanie podkreśla brak wystandaryzowanych metod inżynierii podpowiedzi w badaniach naukowych, co komplikuje wykorzystanie dużych modeli językowych, takich jak ChatGPT oraz wskazuje potencjał ChatGPT do automatyzacji przygotowywania przeglądów systematycznych i opracowywania strategii wyszukiwania w badaniach z obszaru pielęgniarstwa. Pomimo, że ChatGPT okazał się cenny w generowaniu terminów i synonimów, często ma trudności z tworzeniem w pełni dokładnych BSQ. Autorzy argumentują za wykorzystaniem najnowszych modeli ChatGPT, wraz z zaawansowanymi technikami inżynierii podpowiedzi, do zadań naukowych. Zaleca się także prowadzenie dalszych badań w celu udoskonalenia i standaryzacji metod inżynierii podpowiedzi w badaniach z obszaru pielęgniarstwa.

Słowa kluczowe: sztuczna inteligencja, ChatGPT, badania naukowe w pielęgniarstwie, Duże Modele Językowe, strategia wyszukiwania Boole'a

ABSTRACT

EVIDENCE-BASED ADVANCED PROMPT ENGINEERING IN NURSING RESEARCH: QUALITY ANALYSIS OF CHATGPT-GENERATED BOOLEAN SEARCH QUERY

Aim. This article explores the use of advanced prompt engineering in nursing research, with a focus on ChatGPT-generated Boolean search queries (BSQs).

Material and methods. The study compares the effectiveness of different models of ChatGPT: ChatGPT-3.5, ChatGPT-4.0, and ChatGPT-4omni, in generating high-quality BSQs for PUBMED. The prompting methods analysed involved Zero-Shot, Automated Chain-Of-Thought, Emotional Stimuli, Role-play, and Mixed-Methods prompting.

Results. The study found that ChatGPT-4omni, using Mixed-Methods prompting, achieved the highest quality scores, whereas ChatGPT-3.5, using zero-shot prompting, is the least effective. Significant variability in search outcomes was observed across different models and methods of prompting. The authors recommend ChatGPT-4omni as the most effective model for generating BSQs.

Conclusions. The study highlights the lack of standardized methods for prompt engineering in scientific research, complicating the use of large language models such as ChatGPT and underline the potential of ChatGPT to automate the preparation of systematic reviews and the development of search strategies. While ChatGPT proved valuable for generating search terms and synonyms, it often struggles to produce fully accurate BSQs. The article argues for the use of the latest ChatGPT models, along with advanced prompt engineering techniques, for scientific tasks. It also calls for further research to refine and standardise prompt engineering methods in nursing research.

Key words: nursing research, artificial intelligence (AI), ChatGPT, Large Language Model (LLM), Boolean search query

INTRODUCTION

Since the release of the world's most advanced Large Language Model (LLM), called ChatGPT, by OpenAI, in November 2022, there has been a growing interest among researchers in the effective use of this technology to support research and automate the preparation of publications in various fields of science. With its potential, ChatGPT also became of immediate interest to medical and health sciences researchers worldwide [1-4], but insufficiently in the area of nursing research, as highlighted in international literature [5]. This paper addresses the need for a more empirical and qualitative approach to bridging the gap in the literature on how LLMs can support nursing research.

There are more and more papers reporting on experiments with ChatGPT for specific tasks, such as to automate the preparation of systematic reviews [3,6,7], to generate literature searches [8-10], Boolean search queries [8,11] or answers to clinical questions in PICOT format [12,13]. There is also a growing body of literature on the effectiveness of ChatGPT in producing publication abstracts and how they compare with human-written abstracts, with a particular focus on the issues of plagiarism and AI-detected content [6,7,14,15].

The majority of these publications address the use of one or two ChatGPT models (mostly ChatGPT-3.5 or ChatGPT-4.0 due to their current availability) and/or are limited to a single prompting method, without advanced prompt engineering [16-19]. According to the available research results, the way the question is posed to the LLM, i.e. prompting, is of considerable importance for the quality of the output, not only for general, but also for a scientific output of ChatGPT, where knowledge of advanced prompt engineering is required [16,17,19]. To date, there are no methodological guidelines for prompting in the context of scientific research, which is a substantial obstacle for researchers. This lack of uniform standards makes it difficult to systematise the results and reduces the comparability and reproducibility of studies using generative language models such as ChatGPT. The implementation of such guidelines would be essential to standardise the approach to prompting and ensure consistency and high quality scientific applicability of this technology.

There is a notable lack of research findings on the use of ChatGPT to support research and the analysis of its effectiveness in supporting nursing research [5]. The results presented fill an important gap in knowledge about the evidence-based effectiveness of using different ChatGPT models and various methods of advanced prompt engineering.

They may also provide a set of practical guidelines for other authors to further plan research using ChatGPT in nursing research.

For researchers, the lack of a well-established methodology and the dynamically changing conditions due to rapid advances in LLM technology pose a major challenge. The continuous development of language models such as ChatGPT, together with the lack of unified methodological standards, make it difficult not only to approach research systematically, but also to compare and reproduce results. This situation calls for the development of flexible but consistent guidelines that enable researchers to use LLM technology effectively and reliably, despite its rapidly evolving nature.

AIM

The primary aim of this study was to assess the quality of the scientific output generated by ChatGPT in the form of a Boolean search query to the PUBMED database, aimed at finding literature on measuring the level of cultural competence of nursing students. The topic of the measurement of nursing students' cultural competence was selected due to the author's previous expertise in this area [20-22].

The specific study objectives were:

1. Evaluation and comparison of the quality of ChatGPT-generated Boolean search query depending on the ChatGPT model used: 3.5; 4.0 or 4omni.
2. Evaluation and comparison of the quality of the ChatGPT-generated Boolean search query depending on the prompting method used: Zero-Shot, automated Chain-of-Thought, Role-Playing, and Mixed-Methods Prompting.

METHODS

Study design

The study design was in line with the METRICS model for studies with LLMs proposed by Sallam et al.[23].

Theoretical framework

ChatGPT is based on certain basic GPT models that have been fine-tuned for conversational use. The fine-tuning process makes use of supervised learning and reinforcement learning from human feedback (RLHF). At the time of the study (in July 2024), ChatGPT was available from OpenAI in three models: ChatGPT-3.5; ChatGPT-4.0 and ChatGPT-4omni. The models differ

significantly from each other. There are two main differences. The first one is the neural network architecture used in each model, i.e. the number of parameters and the inter-parameter linkages. The models differ in their training data which is OpenAI's trade secret and not publicly available. The second one is the availability of ChatGPT's free-of-charge models and/or the paid access to each model's throughput.

Research hypotheses

H1: ChatGPT-generated Boolean Search Queries will vary in quality depending on the model used and the sophistication of each model.

H2: ChatGPT-4omni will generate the highest quality Boolean Search Queries due to its advanced neural network architecture and training data.

H3: ChatGPT-3.5 will generate the lowest quality Boolean Search Queries due to its comparatively simpler architecture and limited capabilities.

H4: ChatGPT-generated Boolean Search Queries will vary in quality depending on the prompting method used.

H5: Mixed-Methods prompting will result in the highest quality Boolean Search Queries, reflecting its comprehensive and optimized approach.

Prompting procedure

The ChatGPT platform was accessed via the web-based interface [24]. At the time of the analysis (July, 2024), ChatGPT was available in three models: ChatGPT-3.5, ChatGPT-4.0, and ChatGPT-4omni. All the models were used in the research presented.

While prompting simply refers to the provision of input or commands to guide LLMs, advanced prompt engineering methods involve the systematic design and optimization of prompts to improve model performance [16,19,25].

Using advanced prompting from international research [17], we first generated a prompt directly related to the aim of the study with the zero-shot prompting method [26-30]. Next this prompt was employed uniformly across four remaining prompting methods: Automated Chain-of-Thought, Emotional Stimuli, Role-play, and Mixed-Methods prompting, and tested on three ChatGPT models: ChatGPT-3.5, ChatGPT-4.0, and ChatGPT-4omni.

Zero-shot prompting [26-30]

Zero-shot prompting is like being asked to solve a problem or perform a task without being briefed or given specific examples for that task [26,27,30]. The model generates answers without prior training with particular examples. This technique saves time and promotes innovation and efficiency when researchers need quick, comprehensive answers to different questions without having to prepare specific training examples [26,28-30].

The prompt was as follows: Create a Boolean search query for PUBMED to identify all papers relevant to measuring the cultural competence of nursing students.

Advanced Prompt Engineering Methods (APEM)

Drawing on existing research, the authors then selected evidence-based advanced prompting methods and modified the original prompt using advanced prompt engineering methods (APEM) applied sequentially: Automated Chain-of-Thought Prompting, Emotional Stimuli Prompting, Role-play Prompting [27,31-33].

Automated Chain-of-Thought [34,35]

LLMs can perform complex reasoning by generating intermediate reasoning steps. Providing these steps for demonstration prompts is called Chain-of-Thought (CoT) prompting [34,35]. Chain of Thought prompting is characterized by a series of intermediate reasoning steps that significantly improve the ability of LLMs to perform complex reasoning [35]. CoT prompting has two main paradigms.

One paradigm uses a series of sequential manual demonstrations, each consisting of a question and a chain of reasoning leading to an answer. The superior performance of the other paradigm depends on manual generation of task-specific demonstration prompts one by one [35].

It uses a simple prompt such as "Let's think step by step" to facilitate step-by-step thinking before answering a question [34]. This method is called Automated Chain-of-Thought.

The Automated Chain-of-Thought prompt, modified as intended, was as follows: "Create a Boolean search query for PUBMED to identify all papers relevant to measuring the cultural competence of nursing students. Act step-by-step".

Emotional Prompting [36]

Emotion Prompt is a technique described as a simple yet effective approach to exploring the emotional intelligence of LLMs. The Emotion Prompt technique involves the use of 11 sentences that act as emotional stimuli. These prompts are designed to be added to the initial prompt, thereby influencing the LLM's responses. The prompts range from direct queries about the LLM's confidence in its answer to more emotionally charged statements such as "This is very important to my career. You'd better be sure" [36].

The Emotion Prompt, modified as intended, was as follows: "Create a Boolean search query for PUBMED to identify all papers relevant to measuring the cultural competence of nursing students. This is very important to my career".

Role-play prompting [27,31-33]

The conventional practice of Role-playing prompting (also: impersonation or persona) is simply to combine the role assignment with the reasoning question into a single prompt to query the LLM, forming a single-turn interaction, such as: "Act as a teacher" [27,31]. The Role-play Prompt, modified as intended, was as follows: "Act as a nursing professor with many years of experience in research in the field of nursing students' cultural competences and author of many publications. Create a Boolean

search query for PUBMED to identify all papers relevant to measuring the cultural competence in nursing”.

Mixed-methods of Advanced Prompt Engineering

Mixed-methods of advanced prompt engineering involve the use of different methods for the optimisation and effective generation of prompts for language models. The concept of mixing different prompt engineering methods to improve the performance of ChatGPT is discussed in various papers, although primary sources on this exact topic are scarce. While there is a consensus that employing a variety of prompt engineering methods can improve the performance of ChatGPT, the specific impact of mixing these methods is often implied rather than explicitly investigated.

The Mixed-methods Prompt, modified as intended, was as follows: “Act as a nursing professor with many years of experience in research in the field of nursing students’ cultural competence and author of many research papers. Create a Boolean search query for PUBMED to identify all papers relevant to measuring the cultural competence in nursing”.

Each of the five prompts described was entered into the ChatGPT models available at the time of the study: 3.5, 4.0, and 4omni. A total of 15 ChatGPT-generated scientific results were obtained. These were Boolean search queries.

During our experiments using ChatGPT, we deliberately disabled the memory function. This means that the model did not retain any information from previous interactions and did not have access to context beyond the current session. This decision was made to ensure the highest level of data privacy and to avoid potential biases arising from the long-term storage of information. Disabling the memory function also aimed to increase the reproducibility of our experiments by ensuring that each query to the model was treated as independent and did not influence subsequent responses. This allowed us to precisely control the experimental conditions and better interpret the results of our research. In this study, careful attention was paid to the potential influence of model memory linked to specific accounts. Although OpenAI has stated that ChatGPT does not retain memory across separate conversations unless explicitly enabled by the user, the possibility of subtle account-specific factors influencing model outputs cannot be entirely ruled out. Such factors might include implicit user profiling or adaptive optimizations based on repeated interactions, even if they do not involve persistent memory in the traditional sense.

To mitigate any potential confounding effects, all prompts were conducted in separate ChatGPT conversations. Each new query was entered in a new conversation to ensure that prior interactions did not influence subsequent responses. This approach was critical in maintaining consistency and reproducibility of the generated outputs across the various models and prompting methods tested. All activities using ChatGPT were performed by the same author (IC) on the same day (11 July 2024), with the same computer and IP number.

While the same user account was utilized for all interactions in this study, the influence of account-linked parameters such as IP address, usage patterns, or regional settings was not explicitly tested. The decision not to create new accounts for this research was based on the assumption that OpenAI’s publicly available documentation accurately represents the behavior of its models, specifically regarding the absence of session carryover effects when starting a new conversation.

Methods for evaluating the quality of ChatGPT-generated Boolean search queries

The quality assessment of the ChatGPT-generated BSQs was conducted by a team of four Experts, authors of the publication, including two professors in the field of health sciences (JGM; MP) and two PhDs in the field of health sciences (IC; JM) with a recognized background in measuring the level of cultural competence of nursing students [20,21,22]. Each of the 15 BSQs generated by ChatGPT was independently evaluated by a Team of Experts.

In order to analyse the quality of the ChatGPT-generated BSQs, the Experts Team developed a scoring system following general guidelines [23] (Tab. 1). For each BSQ, the Experts assigned scores according to the criteria, and the individual scores were then averaged to obtain a final score for each query. This average score represents the collective assessment of the experts and was used for comparative analysis in the study.

Neither a minimum nor a maximum number of points to be obtained, nor a scale for positive or negative evaluation of the ChatGPT-generated BSQs was established, as it was impossible to predict the responses ChatGPT would generate, which is an inherent characteristic of LLM models. An ex-post evaluation was performed of ChatGPT.

■ Tab. 1. Assessment areas and indicators of the quality of ChatGPT-generated Boolean search queries

Assessment area	Indicators	Scoring
1. ChatGPT correctly identified synonyms or related terms	The number of synonyms or related terms	One point for each term
2. ChatGPT combined terms with Boolean operators: AND, OR, NOT	The number of Boolean operators used	One point for each operator
3. ChatGPT grouped synonyms and related terms with parentheses	The number of groups of synonyms	One point for each group of synonyms
4. ChatGPT used quotation marks for exact phrases	The number of quoted phrases	One point for each phrase with quotation marks
5. ChatGPT used field tags to search within specific fields	The number of words with field tags	One point for each word used in field tags
6. ChatGPT used truncation/an asterisk (*) to find variations of a word	The number of asterisk (*) words used	One point for each word with an asterisk (*) used

RESULTS

In line with international publication practice for ChatGPT-generated content [25], the unmodified ChatGPT output is available in the Supplementary file.

ChatGPT-generated output typically contains not only the anticipated response, but also additional randomly

generated content. However, the authors of this paper did not analyse the additional ChatGPT-generated content in a systematic way, but focused only on the analysis of Boolean search queries, which was the aim of this paper.

A total of 15 Boolean search queries generated by ChatGPT were analysed (Tab. 2).

■ Tab. 2. Characteristics of the group of respondents in terms of a year of starting the studies and their profession

Prompt engineering method	ChatGPT model	ChatGPT-generated Boolean search query	Total number of synonyms	Boolean operators: AND, OR, NOT/No. of points	Grouped terms /No. of points	Quotation Marks /No. of points	Field tags /No. of points	An asterisk* /No. of points	Total number of points	The total number of papers in PUBMED
Zero-shot Prompting	ChatGPT 3.5	(„cultural competence” OR „cultural competency”) AND („nursing students” OR „student nurses”) AND („assessment” OR „measurement” OR „evaluation”)	7	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/7	No/0	No/0	19	222
	ChatGPT 4.0	(„cultural competence” OR „cultural competency” OR „cultural sensitivity” OR „cultural awareness”) AND („nursing students” OR „student nurses” OR „nurse trainees”) AND (measurement OR assessment OR evaluation OR scale OR instrument)	12	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/7	No/0	No/0	24	329
	ChatGPT 4omni	(„cultural competence” OR „cultural sensitivity” OR „cultural awareness” OR „transcultural nursing” OR „intercultural competence”) AND („nursing students” OR „student nurses” OR „undergraduate nursing” OR „nursing education” OR „nursing training”) AND („measurement” OR „assessment” OR „evaluation” OR „scale” OR „instrument” OR „survey”)	16	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/16	No/0	No/0	37	599
Automated Chain-of-thought Prompting	ChatGPT 3.5	(cultural competence OR cultural competency OR cross-cultural competence OR intercultural competence) AND (nursing student* OR nursing education OR student nurse* OR undergraduate nursing) AND (measure* OR assess* OR evaluation OR instrument OR tool OR scale)	14	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	No/0	No/0	Yes/4	23	1468
	ChatGPT 4.0	(cultural competence OR cultural competency OR cultural awareness OR cultural sensitivity) AND (nursing students OR nursing learners OR nursing undergraduates OR student nurses) AND (measurement OR assessment OR evaluation)	11	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	No/0	No/0	No/0	16	936
	ChatGPT 4omni	(„cultural competence” OR „cultural sensitivity” OR „cultural awareness” OR „intercultural competence” OR „transcultural nursing” OR „cultural proficiency”) AND („nursing students” OR „student nurses” OR „nursing education” OR „nursing trainees”) AND (measurement OR assessment OR evaluation OR „measurement tools” OR „measurement instruments” OR scales)	16	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/12	No/0	No/0	33	677

■ cont. Tab. 2. Characteristics of the group of respondents in terms of a year of starting the studies and their profession

Prompt engineering method	ChatGPT model	ChatGPT-generated Boolean search query	Total number of synonyms	Boolean operators: AND, OR, NOT/No. of points	Grouped terms /No. of points	Quotation Marks /No. of points	Field tags /No. of points	An asterisk* /No. of points	Total number of points	The total number of papers in PUBMED
Emotional Stimuli Prompting	ChatGPT 3.5	(„cultural competence” OR „cultural competency”) AND („nursing students” OR „nursing education”) AND („assessment” OR „evaluation” OR „measurement”)	7	AND: yes/1 OR: yes/1	Yes/3	Yes/7	No/0	No/0	19	373
	ChatGPT 4.0	(„cultural competence” OR „cultural awareness” OR „intercultural competence”) AND („nursing students” OR „nursing education” OR „nurse trainees”)	7	AND: yes/1 OR: yes/1 NOT: no/0	Yes/2	Yes/7	No/0	No/0	18	1209
	ChatGPT 4omni	(„cultural competence” OR „cultural sensitivity” OR „cultural awareness” OR „cultural intelligence” OR „intercultural competence”) AND („nursing students” OR „student nurses” OR „nurse trainees” OR „nursing education” OR „nurse training”) AND („assessment” OR „measurement” OR „evaluation” OR „survey” OR „instrument” OR „tool” OR „questionnaire”)	17	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/17	No/0	No/0	39	365
Role-playing Prompting	ChatGPT 3.5	(„cultural competence” OR „cross-cultural competence” OR „cultural competency”) AND („nursing students” OR „student nurses”) AND („measurement” OR „assessment” OR „evaluation”)	8	AND: yes/1 OR: yes/1	Yes/3	Yes/8	No/0	No/0	21	222
	ChatGPT 4.0	(„cultural competence” OR „cultural competency” OR „intercultural competence” OR „cultural sensitivity”) AND („nursing students” OR „nursing education”) AND (measurement OR „assessment tool” OR scale OR survey)	10	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/7	No/0	No/0	22	391
	ChatGPT 4omni	(„cultural competence” OR „cultural sensitivity” OR „cultural awareness” OR „cultural diversity” OR „transcultural nursing”) AND („nursing students” OR „student nurses” OR „nursing education” OR „nursing curriculum”) AND („measurement” OR „assessment” OR „evaluation” OR „scale” OR „instrument” OR „questionnaire”)	15	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/15	No/0	No/0	35	823
Mixed-methods	ChatGPT 3.5	(„cultural competence” OR „cultural competency” OR „cross-cultural competence” OR „intercultural competence”) AND („nursing students” OR „nursing education” OR „student nurses”) AND (assessment OR measurement OR tools OR instruments OR evaluation OR scale)	13	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/7	No/0	No/0	25	556
	ChatGPT 4.0	(„cultural competence” OR „cultural proficiency” OR „cultural sensitivity” OR „intercultural competence”) AND („nursing students” OR „nursing undergraduates” OR „student nurses”) AND („measurement” OR „assessment” OR „evaluation”)	10	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/10	No/0	No/0	25	153
	ChatGPT 4omni	(„Cultural competence” OR „Cultural sensitivity” OR „Cultural awareness” OR „Cultural knowledge” OR „Cross-cultural competence” OR „Intercultural competence”) AND („Nursing students” OR „Nursing education” OR „Nursing trainees” OR „Student nurses” OR „Nursing undergraduates”) AND („Measurement” OR „Assessment” OR „Evaluation” OR „Measurement tool” OR „Measurement scale” OR „Measurement instrument” OR „Survey”)	18	AND: yes/1 OR: yes/1 NOT: no/0	Yes/3	Yes/18	No/0	No/0	41	325

Based on the indicators in Tab. 1, the Boolean search queries generated by ChatGPT 4omni with the Mixed-Methods approach obtained the highest score (41 points) and the queries generated by ChatGPT 4.0 with the automated Chain-of-Thought approach obtained the lowest score (16 points).

The analysis of the number of synonyms or related terms was the first criterion of the BSQ quality assessment. The highest number of synonyms, i.e. 18 terms, was generated by ChatGPT-4omni with the Mixed-Methods approach, while the lowest number, i.e. 7 terms, was generated by ChatGPT-3.5 using Zero-Shot Prompting and ChatGPT-3.5 and ChatGPT-4.0 using Emotional Stimuli Prompting.

The second criterion was the use of Boolean operators: AND, OR and NOT. Regardless of the ChatGPT model or prompting method, all generated BSQs had AND and OR operators, while none had the NOT operator. Also, regardless of the ChatGPT model or prompting method, all generated BSQs had synonyms grouped by parentheses.

The vast majority of the ChatGPT-generated BSQs featured quotation marks for exact phrases. There were two cases where ChatGPT-3.5 and ChatGPT-4.0 with Automated Chain-of-Thought did not generate quotation marks. There were four cases where the use of quotation marks was ‚mixed‘ (some query phrases used quotation marks and some did not). Regardless of the prompting method, none of the ChatGPT models used field tags to search within specific fields. The analysis of the use of truncation/ an asterisk* to find variations of a word showed that the only case where truncation (an asterisk*) was used was in ChatGPT-3.5 and prompting with Automated Chain-of-Thought (two synonyms tagged with an asterisk (*)).

A manual literature search of the PUBMED database using the ChatGPT-generated BSQs yielded highly variable results, ranging from 153 to 1,468 publications. Selectivity and relevance of the retrieved publications were not analysed (see Discussion).

In conclusion, the ChatGPT-4omni model with Mixed-Methods prompting was the most effective in analysing ChatGPT-generated synonyms. For the other criteria there was no clear tendency for any of the ChatGPT models or any of the advanced prompting methods used to generate correct output.

DISCUSSION

A preliminary analysis of literature databases showed that the number of publications on ChatGPT and nursing is very limited, amounting to 185 publications in Web of Science, 140 publications in PUBMED and 108 publications in SCOPUS (search performed on 23 July 2024, duplicate publications were not analysed). Even fewer papers in nursing research address the key issue related to the use of ChatGPT, i.e. knowledge of prompt engineering principles [37,38].

The vast majority of the available publications [5] are primarily concerned with analysing the potential of

ChatGPT in undergraduate [37,39] and postgraduate nursing education [40], and the use of ChatGPT for specific tasks carried out in daily nursing practice e.g. such as preparing care plans [41], patient education or administrative tasks [42-45]. A number of publications also address the ethical use of GPT technology, particularly by students [46].

This topic has also been described in the form of a detailed guide and strategies specifically designed for nursing education to improve teaching effectiveness, work efficiency and student learning outcomes [37,38].

No specific example of the use of ChatGPT as a research support tool in nursing was identified: no publication was found on the automation of specific tasks in the preparation of a systematic review (SR) in nursing, e.g. the generation of a clinical question in the PICOT format or the generation of a search query for literature databases. There are a few publications in the international literature describing the above topics, but not in nursing [8,47-49].

Therefore, the present analysis of the effectiveness of the scientific application of available ChatGPT models and selected advanced prompt engineering methods (APEM) in nursing is a novel one. The results may be useful for planning further research experiments with ChatGPT in nursing research.

The tasks described in the literature that could be automated in the development of a systematic review (SR) involve the generation of search queries (SQs) for literature databases. Several publications present the results of such an attempt, but none in nursing [8,47-49].

Our study assessed the effectiveness of ChatGPT in generating a Boolean search query for the PUBMED database to find publications on measuring cultural competence in nursing students. As we know, the ChatGPT model operates as a black box, i.e. the process by which it arrives at its answers is unknown, and the results obtained must be analysed with great caution and awareness of the limitations of the ChatGPT technology [50].

In all cases, apart from generating the expected Boolean query, ChatGPT provided additional explanations of its actions, with varying degrees of detail (see: Supplementary), typical of a language model such as ChatGPT. However, it should be noted that in none of the queries did the authors specify that ChatGPT only generated BSQs without additional comments, hence this was not surprising. In one case (ChatGPT-3.5 model), the response began by addressing the author of the query with a male Polish name: Jerzy (see Supplementary), which was neither anticipated nor accurate.

The authors only analysed the ChatGPT-generated Boolean search queries, not the entire content of the output.

Our analysis of the results and evaluation of the quality of the ChatGPT-generated BSQ to the prompt: Create a Boolean search query for PUBMED to identify all papers relevant to measuring the cultural competence of nursing students demonstrated that none of the currently available ChatGPT models, nor any of the advanced prompt engineering methods described in the existing research,

produced a fully correct ChatGPT-generated BSQ. However, as a language model, ChatGPT can be very useful for generating search terms, related terms and their synonyms. It seems that ChatGPT-4omni and the Mixed-Methods prompting produce the most accurate BSQ results, yet this conclusion requires further research and in-depth analysis.

The analysis of the variation of ChatGPT-generated search terms proved ChatGPT-3.5 with Zero-Shot prompting to be the least effective and ChatGPT-4omni with Mixed-Methods prompting to be the most effective.

The diversity of ChatGPT-generated Boolean search queries (BSQs) significantly affected the literature search results. The authors' manual literature search of the PUBMED database with the ChatGPT-generated BSQs yielded search results ranging from 153 to 1,468 publications, but no detailed analysis of the relevance and selectivity of the retrieved publications was performed. The analysis of indicators such as specificity and sensitivity of the manual search in the PUBMED database was, in our opinion, not justified in this case as none of the ChatGPT-generated BSQs fully met the accuracy criteria set out by the authors in Tab. 1.

In conclusion, the use of the ChatGPT-3.5 model for Boolean search queries is not recommended. The ChatGPT-4omni model seems to be far more efficient, with more parameters and their connections. Furthermore, Zero-Shot prompting, is not effective in generating BSQs, even with advanced prompt engineering methods. Therefore, the use of BSQ examples is recommended before generating BSQs with ChatGPT.

In other studies, the quality of the ChatGPT-generated Boolean search query strategies depended heavily on the prompting method [8,9,47,49], reflecting the general principles and assumptions of LLM operation.

In a study by Guimarães, the authors obtained unsatisfactory results for a BSQ generated with Zero-Shot prompting, without successive iterations or instructions to generate a search query under Boolean logic, by entering only a simple prompt in the ChatGPT conversation window: Create a search strategy in MEDLINE that reflects the following central question: When does weight regain occur in obese individuals after bariatric surgery? The output obtained by Guimarães contained many errors, in particular imaginary MeSH terms (ChatGPT used non-existent medical subject headings as part of the search) and search filters added to the BSQ against the authors' instructions (e.g. publication date and/or type of publication) [49].

In contrast, research clearly shows generating BSQs to be more effective with few-shot prompts [28].

Wang et al. have presented a strategy for generating BSQ for SR using PubMed syntax, with few-shot prompting and Chain-of-Thought prompting, performed in multiple iterations. Their method requires four prompts that provide step-by-step instructions for ChatGPT and a title and quotes from a source publication (ideally a previously published systematic review) describing the searchable PICOT items in the search question, i.e. P (Population/Patient/Problem), I (Intervention/Indicator), C (Comparison),

O (Outcome); T (Time/Type of Study). According to Wang et al., ChatGPT could still generate satisfactory BSQs with high precision, but with lower recall, compared to other state-of-the-art automated search strategy generators and human-generated BSQs [28].

In contrast, Alaniz et al. obtained satisfactory results in generating a Boolean search query in line with author-generated strategies. They used a few-shot prompting strategy to generate a BSQ, instructing ChatGPT in four successive iterations on the next steps for generating a literature search strategy in PUBMED, as well as translating search strings from PubMed to other databases [8]. Although the process was successful, given ChatGPT-4's training cut-off date of 2021, the researchers still needed to match MeSH terms to the database. Similar results were found by Qureshi et al., who only generated a PubMed query for a systematic review in the third iteration, preceded by activities related to the PICO framework and eligibility criteria [48]. However, the proposed search strategy was unusable due to several problems, including fabricating controlled vocabulary by ChatGPT, which would not be apparent without expertise in search construction [48]. Qureshi et al. also emphasise that great caution should be taken by non-content experts when using these tools, as though much of the output appears to be at a high level, or to be valid, in fact, much of it is flawed erroneous and requires active vetting [48].

Kurian et al. have reported satisfactory results in generating BSQs, although in a letter to the editor published in the British Dental Journal they state that ChatGPT generated BSQs that were correct yet required further revision. However, the authors did not specify the content of the prompt or any details about the quality of the output obtained [51].

These findings are not surprising given that the research emphasises that LLMs work more effectively and efficiently as few-shot learners, i.e. after the user has provided examples of the expected output [52].

None of the analysed international publications used advanced prompt engineering techniques such as Automated Chain-of-Thoughts [34,38], Emotional Stimuli [36], or Role-playing [27,32,33], therefore, research using these methods, including in nursing research, should continue.

The generation of a valid Boolean query by a researcher, regardless of the technology used, requires extensive knowledge, relevant experience and the ability to construct complex Boolean queries according to predefined rules. ChatGPT's support for Boolean query design, with a particular focus on few-shot prompting, appears to be a new way of speeding up and streamlining query generation. However, without a good knowledge of the methodology of how to construct a proper Boolean query, it is impossible to use ChatGPT effectively and successfully for this task.

As for now, the process of efficient and effective BSQ generation with ChatGPT requires validation, verification, researcher oversight and further study.

Strengths

This comparison of the quality of ChatGPT-generated BSQs using the three available ChatGPT models and five different advanced prompting methods is novel. Furthermore, there is no limitation regarding the use of the national language (Polish) in conversation with ChatGPT. Although the authors are not native speakers of English, the analyses presented relate to prompts formulated in English, which increases their practical usefulness for an international audience.

Limitations

Limitations of the Technology

Due to the architecture of LLM models such as ChatGPT, it is not possible to obtain reproducible output results, even when using exactly the same prompt, and the chance of obtaining the same result when entering the same prompt is estimated to be one in a million. Therefore, any subsequent attempt to use the prompting methods presented in the existing literature, including those presented in this publication, will produce different results that are difficult to compare.

It is also important to highlight the rapid variability of the available GPT technologies. At the time of the study (11 July 2024), ChatGPT was available with ChatGPT-3.5; at the time of preparing the publication for print (23 July 2024), this model was no longer available and had been replaced by newer versions.

Limitations of the Research Project

The analysis of the output generated after the first prompt was entered, with no subsequent iterations is a limitation of the findings. As LLM models are iterative models, researchers emphasise the importance of the impact of subsequent iterations on increasing the quality of ChatGPT-generated output. Therefore, in future studies we plan to analyse the quality of ChatGPT-generated scientific output taking into account multiple iterations in a conversation with ChatGPT, few-shot prompting methods, as well as the use of other advanced prompt engineering methods.

Suggestions for future research

The rapid evolution of large language models (LLMs) and their application in nursing research opens numerous opportunities for future studies. One key area for exploration is the impact of iterative prompting strategies on the quality and accuracy of LLM-generated outputs. Investigating these techniques could provide deeper insights into optimizing the use of LLMs for nursing-specific research tasks.

Furthermore, comparative research across disciplines would be valuable in assessing how advanced prompt engineering techniques perform in nursing relative to fields like medicine, public health, and social sciences. Such studies could help establish domain-specific guidelines, ensuring that LLMs are utilized effectively across various research contexts. Given the continuous development of newer LLM versions, future research should also focus on validating findings with updated models to

ensure that recommendations remain relevant and adaptable to technological advancements.

To facilitate such advancements, there is a clear need for standardized evaluation frameworks. Developing universal metrics for assessing the quality and utility of LLM-generated outputs would improve comparability and reproducibility across studies. Ethical considerations also deserve attention, particularly regarding data privacy, biases in model outputs, and equitable access to advanced AI tools for researchers in resource-constrained environments.

Finally, integrating prompt engineering techniques into nursing education could empower students and professionals to harness LLMs effectively. This could bridge the gap between theoretical knowledge and practical application, fostering innovation and efficiency in both research and clinical practice.

CONCLUSIONS

The usefulness of ChatGPT for generating a perfectly valid Boolean search query in nursing research, regardless of the ChatGPT model and prompting method, is currently limited and requires caution and awareness of the limitations of GPT technology. The correctness of ChatGPT-generated BSQs should always be ultimately verified by a researcher familiar with search queries and able to critically evaluate ChatGPT-generated output.



At present, ChatGPT-4o appears to be the most effective model for generating BSQs. Given the architecture of the ever-evolving LLM models, it seems reasonable to recommend the use of only the latest available version of the chosen LLM model for research tasks and to use advanced prompt engineering techniques, e.g. few-shot prompting or mixed-methods prompting. The use of less advanced LLM models for scientific purposes, e.g. ChatGPT-3.5, may not only be considered unhelpful in terms of the overall usefulness of the scientific output obtained, but also as too energy consuming with such unsatisfactory results [53,54]. However, the ethical issues of equal access for all, also in science, to state-of-the-art technologies, including ChatGPT models, must also be considered. The use of the most technologically advanced models, often in paid versions, does not meet the conditions of equal access to tools and may further exacerbate inequalities among economically diverse groups of scientists [55].

Research into the effectiveness of various advanced methods of prompt engineering and their application in nursing research should continue.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT for editorial support. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

ORCID

Joanna Gotlib-Małkowska  <https://orcid.org/0000-0002-2717-7741>Ilona Cieślak  <https://orcid.org/0000-0001-7752-6527>Mariusz Jaworski  <https://orcid.org/0000-0002-5207-8323>Mariusz Panczyk  <https://orcid.org/0000-0003-1830-211>

REFERENCES

- Imran M, Almusharraf N. Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Cont. Ed. Technology*. 2023; 15(4): ep464. <https://doi.org/10.30935/cedtech/13605>.
- Chen Q, Sun H, Liu H, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics*. 2023; 39(9): btad557. <https://doi.org/10.1093%2Fbioinformatics%2Fbtad557>
- Islam I, Islam MN, et al. Opportunities and Challenges of ChatGPT in Academia: A Conceptual Analysis. *Authorea* (preprint). 2023. <https://doi.org/10.22541/au.167712329.97543109/v1>
- Biswas SS. Role of Chat GPT in Public Health. *Ann. Bomed. Eng.* 2023; 51(5): 868-869. <https://doi.org/10.1007/s10439-023-03172-7>
- Hobensack M, von Gerich H, Vyas P, et al. A rapid review on current and potential uses of large language models in nursing. *Int. J. Nurs. Stud.* 2024; 154: 104753. <https://doi.org/10.1016/j.ijnurstu.2024.104753>
- Howard FM, Li A, Riffon M, et al. Artificial intelligence (AI) content detection in ASCO scientific abstracts from 2021 to 2023. *J. Clin. Oncol.* 2024; 42(16_suppl). https://doi.org/10.1200/JCO.2024.42.16_suppl.1565
- Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit. Med.* 2023; 6(1): 75. <https://doi.org/10.1038/s41746-023-00819-6>
- Alaniz L, Vu C, Pfaff MJ. The Utility of artificial intelligence for systematic reviews and Boolean query formulation and translation. *Plast. Reonstr. Surg. Glob. Open.* 2023; 11(10): e5339. <https://doi.org/10.1097%2F6GOX.0000000000005339>
- Wang S, Scells H, Koopman B, et al. Generating natural language queries for more effective systematic review screening prioritisation. In: *SIGIR-AP 2023 – Proceedings of the Annual International ACM SIGIR conference on research and development in information retrieval in the Asia Pacific Region*; 2023, p. 73-83. <https://doi.org/10.1145/3624918.3625322>
- Khraisha Q, Put S, Kappenberg J, et al. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res. Synth. Methods*. 2024; 15(4): 616-626. <https://doi.org/10.1002/jrsm.1715>
- Haque S, Eberhart Z, Bansal A, et al. Semantic Similarity Metrics for Evaluating Source Code Summarization. *ICPS '22: Proceedings of the 30th IEEE/AMC International Conference on Program Comprehension*. 2022: 36-47. <https://doi.org/10.1145/3524610.3527909>
- Branum C, Schiavenato M. Can ChatGPT accurately answer a PICOT question? Assessing AI response to a clinical question. *Nurse Educ.* 2023; 48(5): 231-233. <https://doi.org/10.1097/nne.0000000000001436>
- Boudin F, Nie JY, Dawes M. Human clinical information retrieval using document and PICO structure. [In:] *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California; 2010, p. 822-830. <https://aclanthology.org/N10-1124>
- Levin G, Pareja R, Viveros-Carreño D, et al. Association of reviewer experience with discriminating human-written versus ChatGPT-written abstracts. *Int. J. Gynecol. Cancer.* 2024; 34(5): 669-674. <https://doi.org/10.1136/ijgc-2023-005162>
- Makiev KG. A Study on distinguishing ChatGPT-Generated and human-written orthopaedic abstracts by reviewers: decoding the discrepancies. *Cureus*. 2023; 15(11): e9166. <https://doi.org/10.7759/cureus.49166>
- Meskó B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *J. Med. Internet Res.* 2023; 25: e50638. <https://doi.org/10.2196/50638>
- Sahoo P, Singh AK, Saha S, et al. A systematic survey of prompt engineering in large language models: techniques and applications. *arXivLabs* (preprint). 2024. <https://doi.org/10.48550/arXiv.2402.07927>
- Rahman M, Terano HJR, Rahman N, et al. ChatGPT and academic research: a review and recommendations based on practical examples. *Journal of Education, Management and Development Studies*. 2023; 3(1): 1-12. <https://doi.org/10.52631/jemds.v3i1.175>
- Giray L. Prompt engineering with ChatGPT: A guide for academic writers. *Ann. Biomed. Eng.* 2023; 51(12): 2629-2633. <https://doi.org/10.1007/s10439-023-03272-4>
- Cieślak I, Panczyk M, Jaworski M, et al. Access to information on the requirements to work as a nurse in Poland, provided to Ukrainian refugee background nurses by nursing self-government institutions. *Word Wide Web Content Analysis. Pielęgniarstwo XXI Wieku.* 2023; 22(3):132-138. <https://doi.org/10.2478/pielxw-2023-0023>
- Cieślak I, Jaworski M, Panczyk M, et al. Multicultural personality profiles and nursing student attitudes towards refugee healthcare workers: A national, multi-institutional cross-sectional study. *Nurse Educ. Today.* 2024; 134: 106094. <https://doi.org/10.1016/j.nedt.2024.106094>
- Gotlib J, Cieślak I, Wawrzuta D, et al. Challenges in job seeking and the integration of Ukrainian War refugee healthcare workers into the Polish Healthcare System: Facebook content analysis. *Int. J. Public Health.* 2023; 68: 1606139. <https://doi.org/10.3389/ijph.2023.1606139>
- Sallam M, Barakat M, et al. A Preliminary Checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: development study involving a literature review. *Interact J. Med. Res.* 2024; 13: e54704. <https://doi.org/10.2196/54704>
- OpenAI. (2023). ChatGPT [Large language model]. <https://www.openai.com/chatgpt>
- Korzynski P, Mazurek G, Krzykowska P, et al. Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review.* 2023; 11(3): 25-37. <http://dx.doi.org/10.15678/EBER.2023.110302>
- Xie T, Li Q, Zhang J, et al. Empirical study of Zero-Shot NER with ChatGPT. [In:] *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, p. 7935-7956., Association for Computational Linguistics. 2023. <https://doi.org/10.18653/v1/2023.emnlp-main.493>
- Kong A, Zhao S, Chen H, et al. Better Zero-Shot reasoning with role-play Prompting. *arXivLabs* (preprint). 2023. <https://doi.org/10.48550/arXiv.2308.07702>
- Li G, Wang P, Ke W. Revisiting large language models as Zero-shot relation extractors. *arXivLabs* (preprint). 2023. <https://doi.org/10.48550/arXiv.2310.05028>
- Zhu Z, Cheng X, An H, et al. Zero-Shot Spoken language understanding via large language models: a preliminary study. [In:] *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia. ELRA and ICCL; 2024, p. 17877-17883.
- Wei X, Cui X, Cheng N, et al. ChatIE: Zero-Shot information extraction via chatting with ChatGPT. *arXivLabs* (preprint). 2023. <https://doi.org/10.48550/arXiv.2302.10205>
- Salewski L, Alaniz S, Rio-Torto I, et al. In-context impersonation reveals large language models' strengths and biases. *arXivLabs* (preprint). 2023. <https://doi.org/10.48550/arXiv.2305.14930>
- Chen J, Wang X, Xu R, et al. From persona to personalization: a survey on role-playing language agents. *arXivLabs* (preprint). 2024. <https://doi.org/10.48550/arXiv.2404.18231>
- Zheng M, Pei J, Jurgens D. Is "A Helpful Assistant" the best role for large language models? A systematic evaluation of social roles in system prompts. *arXivLabs* (preprint). 2023. https://ui.adsabs.harvard.edu/link_gateway/2023arXiv231110054Z/doi:10.48550/arXiv.2311.10054
- Zhang Z, Zhang A, Li M, et al. Automatic chain of thought prompting in large language models. *arXivLabs* (preprint). 2022. <https://doi.org/10.48550/arXiv.2210.03493>
- Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc. Red Hook, NY, USA; Article 1800, 24824-24837.
- Li C, Wang J, Zhang Y, et al. Large language models understand and can be enhanced by emotional stimuli. *arXivLabs* (preprint). 2023. <https://doi.org/10.48550/arXiv.2307.11760>
- O'Connor S, Peltonen LM, Topaz M, et al. Prompt engineering when using generative AI in nursing education. *Nurse Educ. Pract.* 2024; 74: 103825. <https://doi.org/10.1016/j.nepr.2023.103825>
- Sun GH. Prompt engineering for nurse educators. *Nurse Educ.* 2024. <https://doi.org/10.1097/nne.0000000000001705>
- Labrague LJ, Aguilar-Rosales R, Yboa BC, et al. Student nurses' attitudes, perceived utilization, and intention to adopt artificial intelligence (AI) technology in nursing practice: A cross-sectional study. *Nurse Educ. Pract.* 2023; 73: 103815. <https://doi.org/10.1016/j.nepr.2023.103815>
- Lin HL, Liao LL, Wang YN, et al. Attitude and utilization of ChatGPT among registered nurses: A cross-sectional study. *Int. Nurs. Rev.* 2024; 1-10. <https://doi.org/10.1111/inr.13012>
- Woodnutt S, Allen C, Snowden J, et al. Could artificial intelligence write mental health nursing care plans? *J. Psychiatr. Ment. Health Nurs.* 2024; 31(1): 79-86. <https://doi.org/10.1111/jpm.12965>

42. Hara K, Tachibana R, Kumashiro R, et al. Sentiment analysis of operating room nurses in acute care hospitals in Japan: unveiling passion for perioperative nursing using ChatGPT. *Research Square* (preprint). 2024; <https://doi.org/10.21203/rs.3.rs-4505331/v1>
43. Wang T, Mu J, Chen J, et al. Comparing ChatGPT and clinical nurses' performances on tracheostomy care: A cross-sectional study. *Int. J. Nurs. Stud. Adv.* 2024; 6: 100181. <https://doi.org/10.1016/j.ijnsa.2024.100181>
44. Levin C, Kagan T, Rosen S, et al. An evaluation of the capabilities of language models and nurses in providing neonatal clinical decision support. *Int. J. Nurs. Stud.* 2024; 155: 104771. <https://doi.org/10.1016/j.ijnurstu.2024.104771>
45. Nashwan AJ, Bani Hani S. Enhancing oncology nursing care planning for patients with cancer through Harnessing large language models. *Asia Pac. J. Oncol. Nurs.* 2023; 10(9): 100277. <https://doi.org/10.1016%2Fj.apjon.2023.100277>
46. He FX, Fanaian M, Zhang NM, et al. Academic dishonesty in university nursing students: A scoping review. *Int. J. Nurs. Stud.* 2024; 154: 104752. <https://doi.org/10.1016/j.ijnurstu.2024.104752>
47. Alshami A, Elsayed M, Ali E, et al. Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. *Systems.* 2023; 11(7): 351. <https://doi.org/10.3390/systems11070351>
48. Qureshi R, Shaughnessy D, Gill KAR, et al. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews.* 2023; 12: 72. <http://dx.doi.org/10.1186/s13643-023-02243-z>
49. Sernizon Guimarães N, Joviano-Santos JV, Reis MG, et al. Development of search strategies for systematic reviews in health using ChatGPT: a critical analysis. *J. Transl. Med.* 2024; 22: 1. <https://doi.org/10.1186%2Fs12967-023-04371-5>
50. Dwivedi YK, Kshetri N, Hughes L, et al. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manage.* 2023; 71: 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
51. Kurian N, Cherian JM, Cherian KK, et al. AI-assisted Boolean search. *Br. Dent J.* 2023; 235(6): 363. <https://doi.org/10.1038/s41415-023-6345-0>
52. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *arXivLabs* (preprint). 2020. <https://doi.org/10.48550/arXiv.2005.14165>
53. An J, Ding WD, Lin C. ChatGPT tackle the growing carbon footprint of generative AI. *Nature.* 2023; 615(7953): 586-586. <http://dx.doi.org/10.1038/d41586-023-00843-2>
54. Huisin MO, Aron AR. US universities must tackle their huge carbon footprints. *Nature.* 2023; 623(7985): 32. <https://doi.org/10.1038/d41586-023-03348-0>
55. Khan IA, Paliwal NW. "ChatGPT and Digital Inequality: A Rising Concern." *Scholars Journal of Applied Medical Sciences.* 2023; 11(09): 1646-1647. <http://dx.doi.org/10.36347/sjams.2023.v11i09.010>

Manuscript received: 10.11.2024

Manuscript accepted: 17.12.2024

Translation: Maria Chojnacka