Developing and Evaluating the Use of ChatGPT as a Screening Tool for Nurses Conducting Structured Literature Reviews: Proof of Concept Study Results

MUDD, Alexandra <http://orcid.org/0000-0002-6455-5211>, CONROY, Tiffany <http://orcid.org/0000-0003-0653-7960>, VOLDBJERG, Siri Lygum <http://orcid.org/0000-0002-2622-5481>, GOLDSCHMIED, Anita <http://orcid.org/0000-0003-3819-3728>, FEO, Rebecca <http://orcid.org/0000-0001-9414-2242> and SCHUWIRTH, Lambert <http://orcid.org/0000-0002-6279-5158>

Journal of Clinical Nursing

WILEY

Journal of Clinical Nursing

SPECIAL ISSUE Generative AI in nursing and healthcare

RESEARCH METHODOLOGY: DISCUSSION PAPER - METHODOLOGY OPEN ACCESS

Developing and Evaluating the Use of ChatGPT as a Screening Tool for Nurses Conducting Structured Literature Reviews: Proof of Concept Study Results

Alexandra Mudd[1,2] | Tiffany Conroy[1] | Siri Lygum Voldbjerg[3,4,5] | Anita Goldschmied[2] | Rebecca Feo[1] | Lambert Schuwirth[1]

[1]Flinders University, Adelaide, Australia | [2]Sheffield Hallam University, College of Health, Wellbeing and Life Sciences, Sheffield, UK | [3]Department of Clinical Medicine, Aalborg University, Aalborg, Denmark | [4]Clinical Nursing Research Unit, Aalborg University Hospital, Aalborg, Denmark | [5]Department of Nursing, University College of Northern Denmark, Aalborg, Denmark

Correspondence: Alexandra Mudd (alexandra.mudd@flinders.edu.au; a.mudd@shu.ac.uk)

Received: 4 March 2025 | **Revised:** 15 April 2025 | **Accepted:** 23 April 2025

Funding: The authors received no specific funding for this work.

ABSTRACT

Aim: To examine the feasibility of using a large language model (LLM) as a screening tool during structured literature reviews to facilitate evidence-based practice.

Design: A proof-of-concept study.

Methods: This paper outlines an innovative method of abstract screening using ChatGPT and computer coding for large scale, effective and efficient abstract screening. The authors, new to ChatGPT and computer coding, used online education and ChatGPT to upskill. The method was empirically tested using 400 abstracts relating to public involvement in nursing education from four different databases (CINAHL, Scopus, ERIC and MEDLINE), using four versions of ChatGPT. Results were compared with a human nursing researcher and reported using the CONSORT 2010 extension for pilot and feasibility trials checklist.

Results: ChatGPT-3.5 Turbo was most effective for rapid screening and had a broad inclusionary approach with a false-negative rate lower than the human researcher. More recent versions of ChatGPT-4, 4 Turbo, and 4 omni were less effective and had a higher number of false negatives compared to ChatGPT-3.5 Turbo and the human researcher. These more recent versions of ChatGPT did not appear to appreciate the nuance and complexities of concepts that underpin nursing practice.

Conclusion: LLMs can be useful in reducing the time nurses spend screening research abstracts without compromising on literature review quality, indicating the potential for expedited synthesis of research evidence to bridge the research–practice gap. However, the benefits of using LLMs can only be realised if nurses actively engage with LLMs, explore LLMs' capabilities to address complex nursing issues, and report on their findings.

Implications for the Professional and/or Patient Care: Nurses need to engage with LLMs to explore their capabilities and suitability for nursing purposes.

Patient or Public Contribution: No patient or public contribution.

1 | Background

Evidenced-based practice requires nurses to integrate research evidence alongside their clinical expertise and understanding of patient needs (Craig and Dowding 2020). Consequently, searching and synthesising research evidence is key for nurses, and to support high-quality patient care, nurses undertake structured literature reviews. However, abstract screening within structured

https://doi.org/10.1111/jocn.17818

Summary

- What does this paper contribute to the wider global clinical community?
    - Evidenced-based practice is central to all aspects of nursing and requires nurses to have appropriate knowledge and tools to collate high-quality evidence.
    - Artificial intelligence is creating new opportunities for collating evidence. Previous studies have illustrated LLMs' effectiveness in abstract screening for discrete exact science projects, though nursing's complex, multidimensional nature requires further exploration. Nursing researchers must assess LLMs' suitability for nursing research and disseminate results, ensuring that the nursing perspective is represented.
    - This paper demonstrates that using ChatGPT can be effective as a first-line screening tool. However, nurses require knowledge of LLMs, need to iteratively develop their tools, and be cautious of the variations present within different versions of LLMs.

literature reviews creates four interconnected issues. First, abstract screening is non-creative, repetitive, and fatigue inducing whilst remaining a cognitively demanding task (Kerr et al. 2023). Second, there has been an exponential growth in academic publishing (Bornmann and Mutz 2015; Bornmann et al. 2021), exacerbated by an imbalance between relevant and irrelevant article abstracts (Hamel et al. 2020). Third, as screening is a manual human task, it is acknowledged that human error is part of the process (Hamel et al. 2020; Alshami et al. 2023). Studies report an error rate of between 2.5% (Gartlehner et al. 2019) and 5% (Bannach-Brown et al. 2019). Fourth, the most complex and intractable issue is the amount of researcher time that abstract screening requires (Gartlehner et al. 2020; Van De Schoot et al. 2021). Such is the extent of the human researcher time involved in abstract screening that there is a risk of reviews being outdated at the time of publication (Grbin et al. 2022) impacting on nurses' ability to provide high-quality, evidenced-based care to patients. Furthermore, this time intensive activity likely contributes to the deteriorating effort to new ideas ratio in the sciences (Bloom et al. 2020). Overall, the four issues arising with abstract screening (non-creative activity, growing academic literature, error, and time intensity) individually and synergistically negatively impact nursing research. According to methodological best practice, the 'solution' to the problem of human error is two-person abstract review (Gartlehner et al. 2019; Peters et al. 2020; Lefebvre et al. 2019). However, this 'solution' (two-person review) addresses one of the four factors (error) whilst simultaneously augmenting two other factors (time intensity and non-creative activity). Clearly, current practice is suboptimal and creates delays to research implementation in practice with implications for patients and the wider community.

2 | Large Language Models as Research Tools

Industry is developing tools to assist researchers with academic literature reviews (Blaizot et al. 2022). The advent of free to use Large Language Models (LLM) such as ChatGPT, is growing evidence that researchers can leverage the capabilities of LLMs in abstract screening (Qureshi et al. 2023; Li et al. 2024) and potentially address all four of issues related to abstract screening. To date, research on the use of LLMs to screen abstracts has focused on 'realist' natural sciences and often involving specialist information technologists or software engineers (Alshami et al. 2023; Kebede et al. 2023; Issaiy et al. 2024). This is logical as exact science and AI computer scientists may appear natural companions for LLMs. However, the power and expertise of LLMs is developing at pace (Boiko et al. 2023), to the extent that OpenAI in their technical report purport that GPT-4 has human level performance on multiple professional and academic benchmarks (OpenAI 2023). Still, it is not clear the extent to which LLMs can be used to screen abstracts pertaining to nursing specific research questions and there was no previous literature exploring this issue.

3 | Nursing Knowledge and Nursing Specific Research Questions

There are four reasons why a nursing specific focus on the use of LLMs is required. First, it is documented that nursing does not use a standard terminology and there are variations across organisations and countries (Bertocchi et al. 2023; Johnson et al. 2024). The result is that nursing data is not easy to label and classify. Since LLMs such as ChatGPT have been 'trained' using publicly available internet sources (OpenAI 2023), ChatGPT's 'training' is based on data that does not have consistent terminology which may have implications on ChatGPT's efficacy for nursing related tasks. Second, training ChatGPT on publicly available internet sources includes media portrayals of nurses and nursing which may include inaccuracies (Kalisch et al. 2007; Buresh and Gordon 2013; Allen 2014; Garcia and Qureshi 2022; Allen 2024). Third, nursing knowledge is not always discrete and clearly documented. In their framework of nursing work, Jackson et al. (2021) outline the multi-facetted complexities of nursing requiring physical, emotional, cognitive, and organisational labour. A specific challenge Jackson et al., identified was that 'nursing work is complex with numerous unrecognised aspects that are difficult to specify' (pp. 9). Abdulai and Hung (2023) expand on this theme highlighting the experiential nature of nursing and caring with a focus separate and distinct from the biomedical model, containing activity that is not always consciously acknowledged or documented. This focus on tacit knowledge and interaction with caring, aligns with previous work on understanding nursing as a profession (Benner and Wrubel 1989; Smith 2012; Garcia and Qureshi 2022; Allen 2024). In this context, nurses have queried whether the organic and complex nature of nursing is conducive to a computer-based reductionist algorithmic approach (Abdulai and Hung 2023). However, current LLMs such as ChatGPT, use deep learning with multiple layers of neural networks and large data sets to perform complex, 'human like' tasks (Kumar 2024; Collins 2025). Though LLMs have shown promise in abstract screening in 'hard' sciences (Qureshi et al. 2023; Li et al. 2024), to date there has not been exploration of LLMs use in screening abstracts for nursing specific research questions. Fourth, nurses are the largest healthcare professional group, spend most time with patients, have unique insights when developing and conducting research to inform clinical practice (Siedlecki and

Albert, 2017; May 2021) and hold a pivotal position when developing technological innovations (Philips 2024). As LLMs' capabilities grow, it is imperative that nurses harness the potential for efficiencies in nursing research to advance healthcare outcomes (Hoelscher et al. 2024).

4 | The Study

4.1 | Aim

Our aim was to critically study the capabilities of using ChatGPT to screen abstracts for nursing research where concepts are less well defined, practices are multifaceted, roles are diverse, and contexts are variable. We sought to address the following questions:

- Can we leverage the abilities of LLMs to reduce the resources (time) required for nurses to undertake abstract screening when conducting structured literature reviews in nursing-specific topics?
- Is it feasible and beneficial for nurses to use LLMs to screen abstracts to reduce non-creative and yet cognitively demanding activity?
- Does the use of LLMs to screen abstracts reduce the incidence of error?

4.2 | The Development of the Method

This proof-of-concept study compares the abilities of ChatGPT and a human when screening research abstracts for a scoping review pertaining to public involvement in nursing education. Since it is known that researchers across disciplines face challenges in trying to adapt, learn about, and engage with AI for research purposes (Van Noorden and Perkel 2023), the following sections use an explanatory approach to describe the incremental and adaptive development that took place to reach our final method (using a computer programming script to engage with ChatGPT). Four steps informed the development of the final method: initial exploration of ChatGPT, improving knowledge of ChatGPT and prompting techniques, customising a GPT, and using ChatGPT as a knowledgeable companion. Thereafter, decisions pertaining to the final method will be explored and justified.

4.3 | Initial Exploration of ChatGPT

Our process started in an exploratory manner using ChatGPT 3.5T, a publicly available free-to-access website. ChatGPT was selected because it was freely available, required no special software, and was the source of lots of public conversation about LLMs. We started by having conversations with ChatGPT to explore its ability to screen research abstracts. We posted an abstract into ChatGPT and instructed ('prompted') it to read the abstract and decide on its inclusion in a scoping review based on specific criteria. We customised our ChatGPT website, providing background information so it had context for our conversations and could adapt its responses to meet our needs. See Figure 1. We noted ChatGPT's ability to make decisions on individual abstracts and provide a justification for its decision that appeared logical. Specifically, ChatGPT could summarise abstracts and recognise those that overtly involved our scoping review topic of public involvement in nursing education. A central facet of our approach was requesting ChatGPT to provide a rationale for its decision. Commentators refer to LLM operating a 'black box' (Hutson 2024; Li et al. 2024) consequently, asking ChatGPT to include a rationale is essential to understanding ChatGPT's interpretation of the prompt and what refinements are required (Khan 2024). scoping review topic of public involvement in nursing education. A central facet of our approach was requesting ChatGPT to provide a rationale for its decision. Commentators refer to LLM operating a 'black box' (Hutson 2024; Li et al. 2024) consequently, asking ChatGPT to include a rationale is essential to understanding ChatGPT's interpretation of the prompt and what refinements are required (Khan 2024).

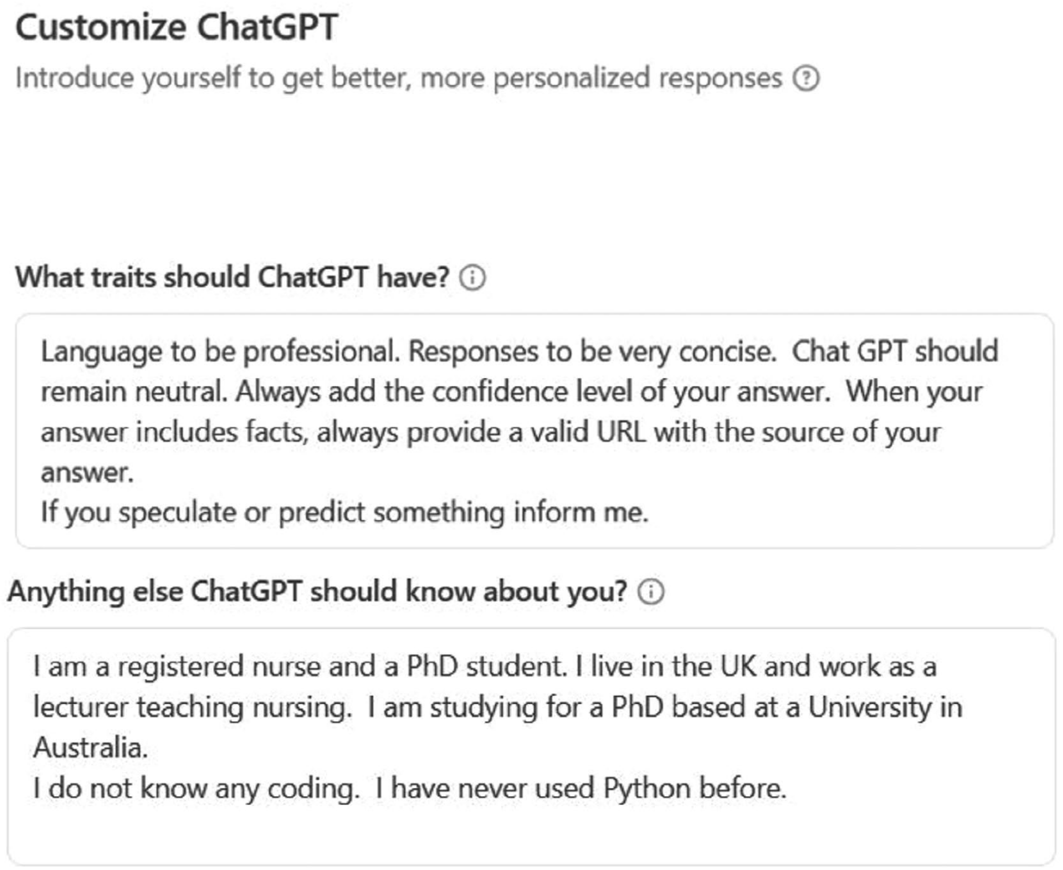
Although we recognised the potential of ChatGPT to screen abstracts, decide on inclusion and provide a rationale for its decision, we were aware of challenges with this mechanism. First, conversing with ChatGPT in a conversation via the website was time intensive; each abstract was submitted individually, and this manual process increased the risk of human error. Second, we needed a mechanism that could deal with a high volume of abstracts without repetitive human intervention. Third, the mechanism must be compatible with receiving data generated from research databases such as MEDLINE and Scopus. Fourth, a mechanism was required to provide output (decisions and a rationale) in a format that is amenable to upload for reviewing and referencing software, for example, Covidence and EndNote. Our exploration of the ChatGPT website indicated using an LLM could be useful; however, we were using it in a sub-optimal manner.

4.4 | Improving Knowledge of ChatGPT and Prompting Techniques

To optimise our knowledge around prompting, we sought support from university library services. However, as the field was developing at pace, it was challenging to find contemporary expert advice from information scientists. Similarly, we searched for short educational courses about the optimal use of ChatGPT and AI prompting from higher education providers; however, we struggled to find suitable courses available at a convenient time. Thus, the internet was the main source of education, using online platforms such as YouTube and LinkedIn Learning alongside search engines (Bing and Google) and conversing with ChatGPT itself. Consequently, our education relating to using ChatGPT and prompting took place outside traditional educational institutions and was personalised to our requirements. We learnt that ChatGPT operates a GPT Marketplace, a platform within ChatGPT that allows users to customise their own GPTs (agents) to fulfil a specific purpose. Searching the GPT Marketplace, we struggled to find a GPT that met our needs and decided to create and customise our own.

4.5 | Customising a GPT

Creating a customised GPT is an iterative process. The user provides instructions to the GPT under a 'configure' tab and then



Customize ChatGPT

Introduce yourself to get better, more personalized responses

What traits should ChatGPT have?

Language to be professional. Responses to be very concise. Chat GPT should remain neutral. Always add the confidence level of your answer. When your answer includes facts, always provide a valid URL with the source of your answer.
If you speculate or predict something inform me.

Anything else ChatGPT should know about you?

I am a registered nurse and a PhD student. I live in the UK and work as a lecturer teaching nursing. I am studying for a PhD based at a University in Australia.
I do not know any coding. I have never used Python before.

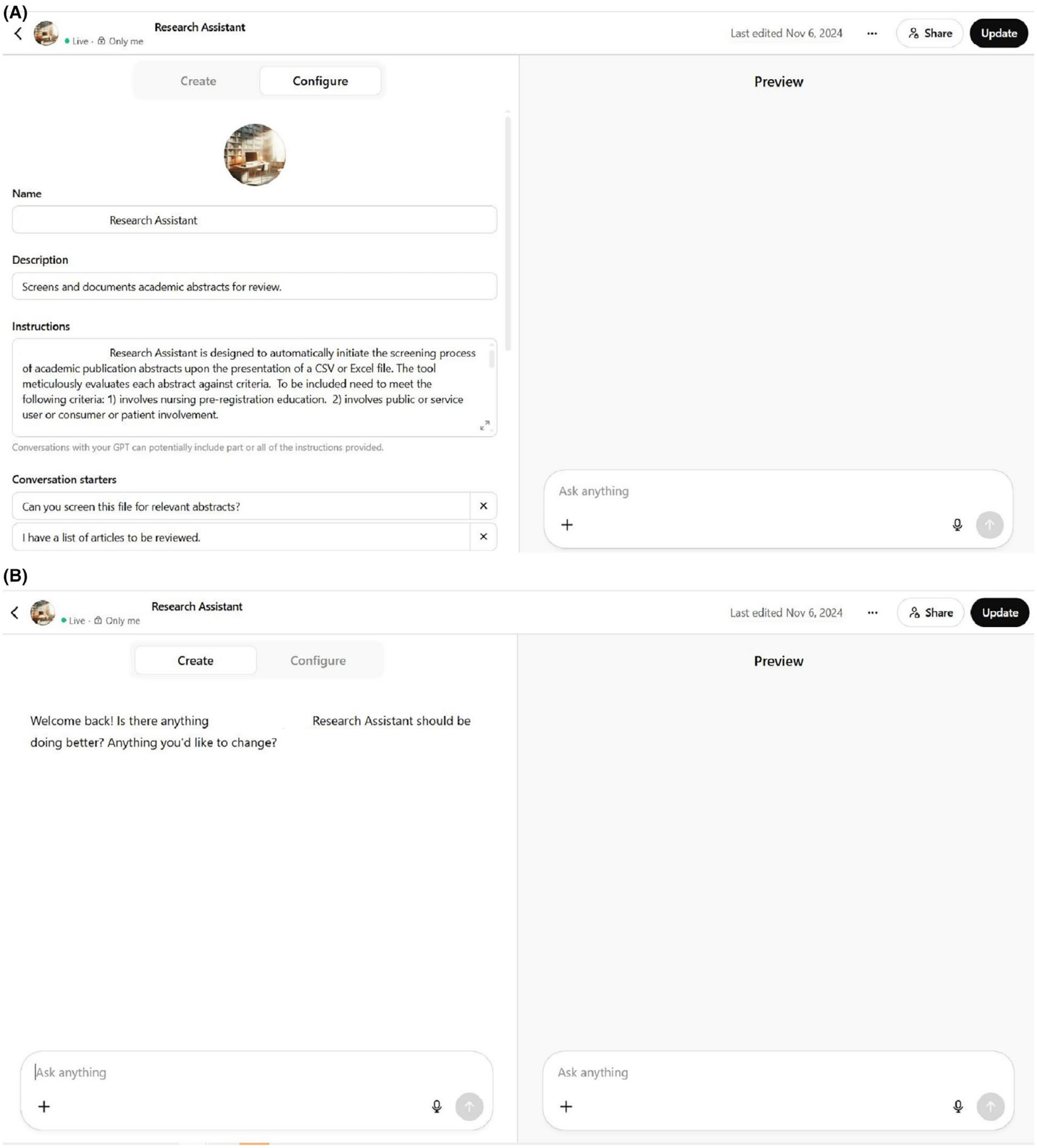
FIGURE 1 | Customisation of ChatGPT via the ChatGPT website.

tests and converses with the GPT 'create' chatbot to optimise results. The GPT 'create' chatbot responds to the user's responses and amends the GPT instructions in the 'configure' tab. Therefore, customised GPT instructions are co-created with ChatGPT. Figure 2 illustrates the two tabs (configure and create) within the customised GPT process. We provided several examples to the configure tabs so that the GPT would be aware of the type of data it would receive and our expectation for output. Our example GPT is available as a Supporting Information file (Method S1). Similar to using ChatGPT website, we encountered problems using the customised GPT. We knew it was possible for GPT to read and screen abstracts according to our criteria, however, it was difficult to get the GPT to read the inputted data correctly and respond with data in the required format. Using an iterative process, we amended the instructions and conversed with the GPT builder to try to correct input and output issues. However, amendments to the GPT instructions via the configure tab, and conversations with the create chatbot had large 'butterfly effect' on the overall prompt to ChatGPT. The result was that the process was not standardised and suitable for repetition at scale.

4.6 | Using ChatGPT Website as a Knowledgeable Companion

Being naïve to technology and GPT, we sought advice from ChatGPT on the best way to achieve our aim of using a large language model to screen academic abstracts using a replicable process, able to operate at scale in a standardised manner. ChatGPT advised us to access OpenAI using an application programming interface (API) via a Python script (see Method S2). In lay terms, this means creating a computing code (a Python script) to contact OpenAI (the parent company of ChatGPT) asking it to read our input file of abstracts one by one, apply the screening criteria via the prompt and respond with a decision, a rationale, and an output file.

Previously we would have considered that using a Python Script to access OpenAI was not within our skillset; however, we used ChatGPT as a learning tool, akin to a knowledgeable companion, to guide us. Often, in our interactions with ChatGPT, we had to ask for simplified instructions and clarification of its responses in lay language. Consequently, we amended our customisation of ChatGPT (Figure 1) to inform it that we were Python naïve and had no coding background. Specifically, whilst installing and practising with Python software, we were able to feed error messages to ChatGPT and gain assistance in how to resolve issues. We spoke with a companion who works in website development and asked for assistance on writing a Python code. Our companion advised us to use Visual Studio Code software, ChatGPT, and Google together to start to create our Python coding script. Our Python script was confirmed for accuracy prior to use by our companion, and amendments were made.



Name

Research Assistant

Description

Screens and documents academic abstracts for review.

Instructions

Research Assistant is designed to automatically initiate the screening process of academic publication abstracts upon the presentation of a CSV or Excel file. The tool meticulously evaluates each abstract against criteria. To be included need to meet the following criteria: 1) involves nursing pre-registration education. 2) involves public or service user or consumer or patient involvement.

Conversations with your GPT can potentially include part or all of the instructions provided.

Conversation starters

Can you screen this file for relevant abstracts?

I have a list of articles to be reviewed.

Welcome back! Is there anything Research Assistant should be doing better? Anything you'd like to change?

FIGURE 2 | Examples of screenshots of the process of creating and customising one's own GPT. (A) Configure display—where completion of text boxes allows a user to set instructions and parameters for their GPT. (B) Create display—where users can converse with ChatGPT to amend the configure tab.

4.7 | The Method—Using a Python Script to Engage With ChatGPT

Our Python script is publicly available (via GitHub and Method S4). The script contains three sections: information about the software and the inputted data, the prompt to OpenAI, and instructions regarding the format of output data. The first section outlines the different programmes and software required to operate the script. In addition, it explains and classifies the data accessed by the Python script. Our planned scoping review will incorporate searches across several databases including MEDLINE, Scopus, CINAHL, and ERIC, and since these databases all facilitate data export via a csv file (.csv) we designed our script to accommodate this format. Furthermore, a csv file is easily converted to Excel and can be transferred to a RIS format to allow integration with software such as Covidence. When designing our script, we were aware of societal concerns that LLMs have the potential to assume human biases (Zack et al. 2024). To avoid bias related to author names (Van De Schoot et al. 2021), citation bias, or journal bias (Lund et al. 2023) we designed the

Python script so that only the publicly available title and the abstract of the article would be accessed by OpenAI. This avoided issues where OpenAI is accessing Doi webpages of abstracts and therefore exposed to all the other irrelevant information on that page (Hill et al. 2024). Only publicly available information, the article abstracts, were accessible to OpenAI, avoiding issues related to accessing information from behind paywalls (Lund et al. 2023).

The second section of the code contains instructions and a prompt to OpenAI, including specifying which version of ChatGPT to use. Most of our preliminary testing was conducted with ChatGPT-3.5 Turbo, the default version at that time. Another aspect within the instructions is the 'temperature'. In lay terms this is set as a number between 0 and 1 and pertains to the degree of randomness that we would like ChatGPT to apply. We used the default value 0, reducing the model's creativity and seeking ChatGPT to provide the most predictable response according to its data set (Khan 2024). This recognises that responses generated by ChatGPT, in our case its written justifications for its decisions, may appear repetitive and not using embellished language (Khan 2024). We were not seeking to use the generative abilities of ChatGPT to write interesting summaries of why the abstracts were to be included or not, instead the focus is on the accuracy of the prompt.

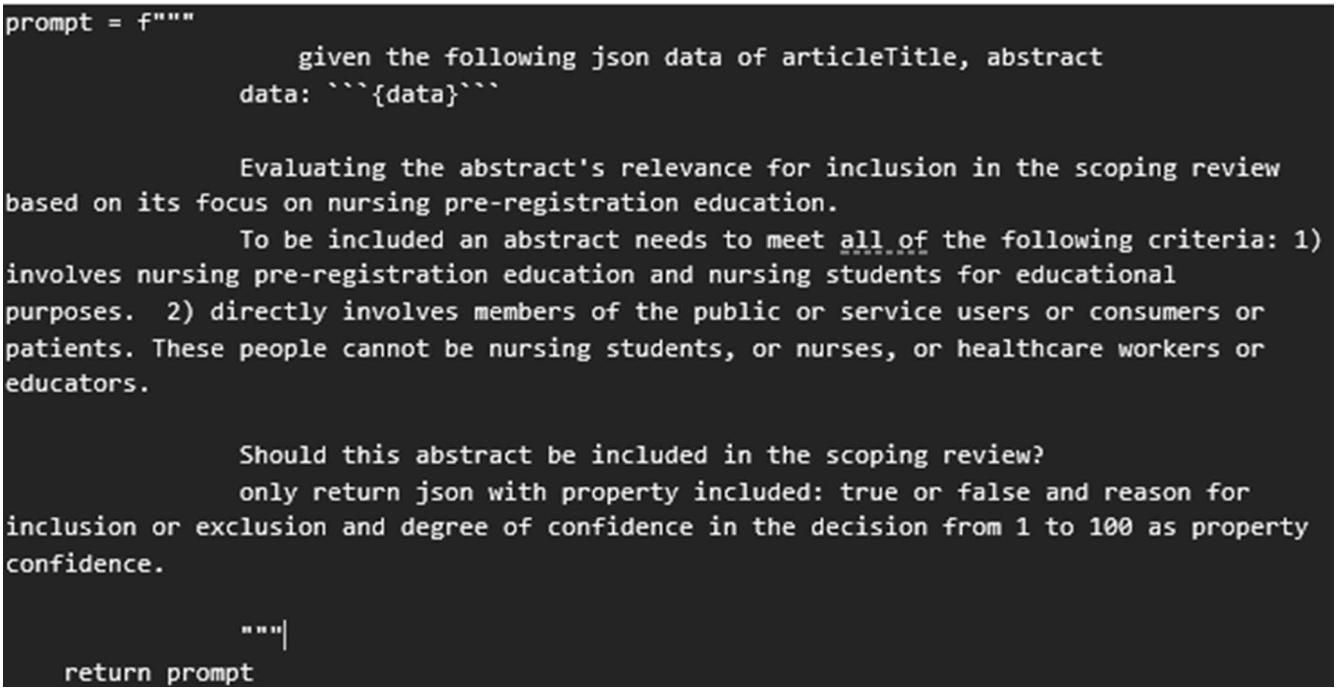
4.8 | Crafting and Refining the Prompt

To formulate the specific prompt, we started by using prompts we had tested on the ChatGPT website and during our attempts to customise our GPT, akin to a Reinforcement Learning from Human Feedback (RLHF) approach. We had learnt that the best prompts were concise, clear, and refined through iteration and review of results (Alshami et al. 2023; Khan 2024). Interestingly, prompts that appeared clear to us as users were sub-optimal. For example, a lengthy explanation of our purpose and an exhaustive list of criteria appeared to create confusion regarding what was most important. Through a process of iteration, we refined the prompt so that it produced results that screened abstracts on specific criteria, even if this resulted in the use of language or grammar we would not naturally choose.

4.9 | Using Prompts Within the Complexities of Nursing

Our interactions with ChatGPT were essential to gaining insight into its interpretation of our prompt. This insight was crucial since we were asking ChatGPT to explore nursing education and public involvement, which relate to diverse interrelated human activities and where authors use a wide variety of language. Besides the development of a tool for high-volume abstract screening as opposed to a single-shot GPT, the complex nature of our academic field is where our study builds on previous work. Nursing education is a broad concept involving scheduled learning sessions at university, alongside practical healthcare placements facilitating experiential learning. Consequently, when screening for nursing education, one needs to appreciate the spectrum of learning contexts, from formal to informal. Similarly, public involvement in nursing education creates further subtleties within activity. For example, public involvement in nursing education may include people with lived experience being a guest speaker at a designated education session, ward patients providing ad hoc instant feedback to students on their clinical skills, or public engagement events when redesigning the nursing curriculum. We wanted ChatGPT to include abstracts relating to active public involvement, rather than members of the public receiving healthcare which could involve a student, without any definitive educative activity taking place. This distinction between active involvement in education and participation in healthcare is nuanced. At first, we prompted ChatGPT to be specific on active involvement, however, this narrowly restricted results to cases where members of the public were leading educational sessions. After multiple amendments to the prompt, the best fit for our purposes was to change the design of the prompt so that it covered nursing education for educational purposes (rather than healthcare provision). ChatGPT appeared to understand the potential for human actors to have multiple roles, for example it noted that healthcare workers or university lecturers are also members of the public. Thus, we adapted our prompt to focus ChatGPT on the specific roles of interest, specifying that for our purposes, we were not considering healthcare workers or educators as members of the public. We were also aware that the language in this field is varied, there being many ways authors refer to members of the public. We wanted to acknowledge the different labels in existence to demonstrate the concept without being too restrictive. Consequently, we amended the prompts to include abstracts that 'directly involve members of the public or service users or consumers or patients. These people cannot be nursing students, or nurses, or healthcare workers or educators'. This amendment was to give context to the multiple potential roles that we are seeking to include whilst recognising the limits. This refining period was significant as the final prompt was different from how we would have instructed a human reviewer. Instead of focusing on three concepts, nursing students, members of the public and engagement for educational activity, we had two criteria which incorporated the three concepts. It was crucial that the prompt was concise because too much detail made it difficult for ChatGPT to determine the principal factors. Moreover, we wanted ChatGPT to take a broad approach allowing for diversity in the results, thus we were only restrictive when distinctly necessary. This approach took time, professional insight, and energy to obtain optimal results.

Overall, our prompt (see Figure 3) has three elements: (1) it asks ChatGPT to evaluate data (read and understand each individual abstract) (2) decide upon its inclusion to the scoping review based on the specific criteria and provide a degree of confidence in the decision (include? True or False and a value from 0 to 100 on the degree of confidence in the decision) (3) generate a rationale to explain its decision. In our previous interactions with ChatGPT via the website and via our custom GPT, we provided ChatGPT with some example results. However, given the 'black box' nature of LLM (Hutson 2024; Collins 2025) we were uncertain the extent to which ChatGPT was weighting these examples when it already draws on an extensive knowledge base. A concern was that if we asked ChatGPT to focus heavily or exclusively on data examples provided we may constrict its ability to function. For example, if it came across an abstract in the literature which



```
prompt = f"""
                    given the following json data of articleTitle, abstract
                data: ```{data}```

                Evaluating the abstract's relevance for inclusion in the scoping review
based on its focus on nursing pre-registration education.
                To be included an abstract needs to meet all of the following criteria: 1)
involves nursing pre-registration education and nursing students for educational
purposes.  2) directly involves members of the public or service users or consumers or
patients. These people cannot be nursing students, or nurses, or healthcare workers or
educators.

                Should this abstract be included in the scoping review?
                only return json with property included: true or false and reason for
inclusion or exclusion and degree of confidence in the decision from 1 to 100 as property
confidence.

                """
    return prompt
```

FIGURE 3 | The prompt used across all versions of ChatGPT.

included elements that were not in the training data it may not work optimally. Moreover, a key aspect of screening academic abstracts is the imbalance in abstracts (usually larger numbers of exclusions compared to inclusions) which makes obtaining a large and authentic test dataset challenging (Van De Schoot et al. 2021). In their exploration of accessing OpenAI via an API, Issaiy et al. (2024) used a 'zero shot approach' which provides a prompt without specific test data (recognising the large language model's capabilities to process data independently) and we followed that approach.

The ultimate aspect of the code is formatting the data so that the output is usable. The Python script creates an output file on the user's computer which contains a list of articles for inclusion and exclusion (in formats compatible for manipulation and upload to other software, for example, Covidence and clear reporting to meet PRISMA guidance) and files containing the decision, confidence level and ChatGPT's rationale for thedecision.

4.10 | Evaluation of the Method

To evaluate the accuracy of the Python script and ChatGPT's decisions, we ran a proof-of-concept study. Our scoping review pertained to public involvement in nursing education and contained two questions: (1) What methods of public involvement are currently used in nursing pre-registration education? (2) What pedagogical or theoretical principles underpin public involvement methods in nursing education? First, we ran our proposed search string across four databases: SCOPUS, Medline, CINAHL (Cumulated Index to Nursing and Allied Health Literature) and ERIC (Education Resources Information Center) and selected the first 100 abstracts. A total of 100 abstracts were selected from each database (total 400) to explore the feasibility of processing large numbers of abstracts and to explore LLM's ability to make inclusion decisions. This number was chosen partly for feasibility reasons relating to human reviewer time and is comparable to previous research (Alshami et al. 2023; Khraisha et al. 2024; Issaiy et al. 2024). We chose a several different databases as this reflects realistic practice and to test the Python script's ability to work with different presentations of data. Second, we manually screened these articles for inclusion in our scoping review based on the same inclusion criteria as provided within the prompt to GPT (Where the human was ambiguous about whether to include an abstract based on the paucity of information in the abstract, the abstract was included. This action was to replicate the stages of screening for a scoping review where an abstract would proceed to full text review.). Third, we ran the Python script to screen the same 100 articles using GPT-3.5 Turbo, GPT-4, GPT-4 Turbo and GPT-4 omni (Note, GPT-3.5 Turbo is the API version of ChatGPT 3.5 which is currently available on ChatGPT website. GPT-4 Turbo was the predecessor of GPT-4 omni, GPT-4 Turbo remains accessible via API though no longer via the ChatGPT website.). Manual screening took place in April 2024 and Python/ChatGPT screening between 12 and 13 May 2024. Where there was a conflict in the decision between ChatGPT and the human reviewer, abstracts were re-reviewed by the human and a final determination made to determine whether ChatGPT or the human had made an error.

5 | Results

5.1 | Overview

The combined results of our Python and ChatGPT screening across 400 abstracts from four databases (MEDLINE, Scopus, CINAHL and ERIC) are displayed in Tables 1–5. Each version of ChatGPT provided unique results. ChatGPT-3.5 Turbo was general and inclusive and the most effective for our purposes. Regarding false negatives, ChatGPT-3.5 Turbo outperformed our

TABLE 1 | Results of abstract screening involving human, ChatGPT 3.5 Turbo, 4, 4 Turbo, and 4 omni on articles from the CINAHL database.

	Includes	Excludes	Tech glitch	True positives	False positives	True negatives	False negatives
GPT 3.5T	38	62	0	9	29	62	0
GPT 4	20	79	1	7	13	77	2
GPT 4T	5	95	0	4	1	90	5
GPT 4O	5	95	0	4	1	90	5
Human	8	92	0	8	0	91	1

TABLE 2 | Results of abstract screening involving human, ChatGPT 3.5 Turbo, 4, 4 Turbo and 4 omni on articles from ERIC database.

	Includes	Excludes	Tech glitch	True positives	False positives	True negatives	False negatives
GPT 3.5T	41	59	0	1	40	59	0
GPT 4	19	81	0	1	18	81	0
GPT 4T	1	99	0	1	0	99	0
GPT 4O	2	98	0	1	1	98	0
Human	1	99	0	1	0	99	0

TABLE 3 | Results of abstract screening involving human, ChatGPT 3.5 Turbo, 4, 4 Turbo, and 4 omni on articles from SCOPUS database.

	Includes	Excludes	Tech glitch	True positives	False positives	True negatives	False negatives
GPT 3.5T	47	53	0	33	14	52	1*
GPT 4	40	59	1	27	12	55	5
GPT 4T	26	74	0	22	4	63	11
GPT 4O	25	75	0	21	4	63	12
Human	34	66	0	33	1	66	0

*This relates to an occasion when an abstract was not available and so unable to run prompt as planned.

TABLE 4 | Results of abstract screening involving human, ChatGPT 3.5 Turbo, 4, 4 Turbo and 4 omni on articles from the MEDLINE database.

	Includes	Excludes	Tech glitch	True positives	False positives	True negatives	False negatives
GPT 3.5T	33	67	0	11	22	65	2*
GPT 4	20	74	6	11	9	73	1
GPT 4T	10	90	0	9	1	86	4
GPT 4O	11	89	0	10	1	86	3
Human	12	88	0	12	0	87	1

*1 of the 2 relates to an occasion when the abstract not available and so unable to run prompt as planned.

human reviewer, making only one error to the human's two when reviewing full abstracts (In both cases, on rereading the abstract the human admitted that they had made an error and that on reflection the human recognised that the abstract should have been included.). There were occasions when no abstracts were available; this created difficulties for our prompt which assessed the text of abstracts. There were two abstracts that the human reviewer included based on title alone. In future, minor changes to the Python code could separate the papers with no abstract available so that these could be reviewed manually. We experienced some technical challenges processing a small number of abstracts usually related to formatting. With regard to timings, all versions of ChatGPT completed the screening of the 400 articles within 2 h, including human administrative time, whereas human screening time was approximately 6 h. The cost of accessing ChatGPT for all the development and evaluation was less than $20 USD.

5.2 | Consideration of ChatGPT-3.5 Turbo

ChatGPT-3.5 Turbo was not as specific as the human reviewer and had a broader range of inclusion (see Table 5). Reading

TABLE 5 | Summarising results from across all databases and all methods of screening (human, ChatGPT 3.5 Turbo, 4, 4 Turbo and 4 omni).

	Includes	Excludes	Tech glitch	True positives	False positives	True negatives	False negatives
GPT 3.5T	159	240	0	54	105	239	3*
GPT 4	99	293	8	46	52	286	8
GPT 4T	42	358	0	36	6	338	20
GPT 4O	43	357	0	36	7	337	20
Human	55	345	0	54	1	343	2

*2 of the 3 relate to occasions when the abstracts were not available and so were unable to run the prompt as planned.

the rationales for ChatGPT-3.5 Turbo's rationales there were some nursing concepts that it appeared to not to be as familiar with. For example, articles which described nursing education using simulated patients or simulated patient experiences were included by ChatGPT as it appeared to classify these as involving patients. In contrast, human reviewers would appreciate that unless specified as being derived from patient, public or consumer involvement, an abstract solely describing simulated patients was unlikely to involve public involvement. At times ChatGPT-3.5 Turbo's lack of specificity creates results that may be interesting for the authors of a review, though not strictly relevant. For example, when ChatGPT-3.5 Turbo encountered secondary research describing reviews of public involvement in nursing education or an abstract validating a tool for measuring service user involvement in nursing education, it marked these abstracts as for inclusion, though strictly speaking, they were not required. This is interesting as one of the points raised about manual screening of reviews is that it is part of the research process and potentially the learning of the individual researcher (Vaishya et al. 2023). In this regard, our results from ChatGPT-3.5 Turbo suggest that it may assist the researcher to view results within a broader context, helping the researcher to better understand relevant factors and generate ideas. Overall, the result is an increase in false positives, impacting on human time; however, without false negatives which impact on the review quality and results.

5.3 | Consideration of ChatGPT-4, 4 Turbo and 4 Omni

In contrast ChatGPT-4, 4 Turbo and 4 omni appeared to produce increasingly 'cautious' results and wanted inclusion criteria to be explicitly mentioned within the abstract. Examples of where ChatGPT-4, 4 Turbo and 4 omni had difficulties with concepts such as health professionals' education and healthcare programmes, co-design and co-creation are available in the Method S3. ChatGPT-4, 4 Turbo and 4 omni appeared to require specific direction and specific details prior to prompting. This creates difficulties for researchers as it requires them to have a solid understanding of the literature under review without blind spots. In empirical or exact sciences, such specificity and explicit requirements may exist and may be represented clearly and accurately in the data. However, in areas with more complexity regarding language and cultural practices such as nursing, more fluidity and flexibility is required.

6 | Discussion

Our study demonstrated the nuances involved in using LLMs as abstract screening tools for nursing research where boundaries are not clearly defined, practices are multifaceted, roles are diverse, and contexts are variable. Our research indicates that LLMs can reduce the time required to screen abstracts. Time is a central issue for any researcher, however, in nursing there are additional challenges. For example, unlike medicine, nursing does not have a longstanding tradition and culture of clinical nursing academics where nursing research is an embedded part of the role (Westwood et al. 2018). Without a research tradition and supportive culture, nurses' research time can get consumed by clinical pressures (May 2021). However, considerable nurse researcher time is required to understand the specifics of the model and attune to its workings. Specifically, this study demonstrated the uniqueness of the different versions of ChatGPT and that one cannot adopt a one 'prompt' fits all approach. The development work behind this method was conducted in ChatGPT-3.5 Turbo, hence the prompt was developed iteratively with this model. Overall, with appropriate prompting, ChatGPT 3.5 Turbo can screen abstracts rapidly, enabling the researcher to access an ever-growing body of academic literature.

Our study illustrated the potential for abstract screening using LLMs to be a creative task with learning opportunities. Specifically, the knowledge, understanding and skills of a human nurse researcher are essential to devising a prompt, responding to ChatGPT appropriately and amending the prompt as part of an iterative process. Considerable professional knowledge, experience and academic curiosity are required. In this regard, this study exemplifies the principle of 'co-intelligence' (Mollick 2024) and the significance of the role of the human-in-the-loop (Fui-Hoon Nah et al. 2023). This research evolved due to education from non-traditional sources (YouTube, ChatGPT, LinkedIn Learning, alongside professional discussions, articles, and publicly available resources). ChatGPT was essential in this process in acting as virtual tutor (Lin 2024) helping to identify steps to meet educational goals and as an inexhaustible 'knowledgeable assistant' when issues arose whatever time of day or night (Nazir and Wang 2023). This study exemplifies the impact LLMs are having on education noted elsewhere in academic literature (Lim et al. 2023; Kohler 2024) and how nurses can take advantage of such opportunities.

With regards accuracy, our study ChatGPT-3.5 Turbo was very effective at recognising irrelevant abstracts and correctly excluding them. However, ChatGPT-3.5 Turbo was not as sensitive as the human in discerning true positives and thus included articles that were potentially irrelevant, aligning with previous research (Kebede et al. 2023). Reading the rationales for inclusion, ChatGPT-3.5 Turbo adopted a broad definition of the specified concepts such as 'nursing education' and 'public involvement.' Due to reduced specificity compared to the human, ChatGPT-3.5 Turbo was not a 'magic bullet' solution to abstract screening. Nonetheless, it created efficiencies whilst not compromising review quality, since false positives could be identified easily at full text review stage. The number of false negatives was comparable to our nurse researcher and could be improved with minor modifications to the Python code to separate articles without an available abstract so that these could be reviewed manually. In contrast, GPT-4, 4 Turbo and 4 omni appeared increasingly specific and sought explicit confirmation within the abstract on the specified concepts. This appears to be a 'hyper-conservative' approach noted in other ChatGPT-4 research (Strachan et al. 2024). Previous studies have not explored the different versions of ChatGPT and how developments to the algorithm are impacting on abilities to screen in complex topics. Our results suggest a hybrid approach where ChatGPT-3.5 Turbo is used as 'first line' screening tool prior to human nurse researcher abstract screening.

Challenges It is recognised that the generative nature of ChatGPT is built on probabilistic responses which means that responses are not standardised. For example, asking the same question will not necessarily receive the same response each time. The authors were aware of this lack of standardisation prior to commencing this work and used this knowledge to craft a prompt that facilitated a broad inclusionary approach facilitating a high number of false positives (high sensitivity) and low number of false negatives (high specificity). The purpose of this study was to explore ChatGPT's potential to reduce the human nurse researcher resources required for abstract screening, whether it is feasible and beneficial, faster, can help reduce errors, and most importantly how ChatGPT uses and 'understands' nursing specific information. We wanted to create an approach which would assist in removing irrelevant results so that nurses research time and energy can be maximised. If we had sought absolute accuracy in ChatGPT's response this would have required extensive prompting time which would not be in keeping with the realities of time nurses usually have available.

Transparency in methods, decisions and practice is recommended in healthcare research (Aromataris et al. 2024; Thomas et al. 2004). A challenge when working with AI and specifically LLMs is that we can interact with and observe the outputs from LLMs, however we do not understand how LLMs reach their decisions or the learning processes involved—the purported 'black box' (Hutson 2024; Collins 2025). Previous researchers have called for transparency on the inner workings of the 'black box' (Johnson et al. 2024; Li et al. 2024). However, it is unlikely that a commercial company such as OpenAI will willingly share its confidential business information. Open Science Framework's recent draft guidance on Responsible AI in Evidence Synthesis (Thomas et al. 2004) reiterate calls for transparency in the processes around AI use and the importance of researchers being part of the research AI ecosystem for disseminating information and feeding back on experiences of AI. This paper is an example of this collaboration and knowledge sharing and specifically is an example of nursing adding their insights to the conversation.

7 | Recommendations for Further Research

LLMs are constantly evolving, though interestingly, we did not find that the most recent iteration of ChatGPT was the most effective for our purposes. We recommend further nursing research to investigate the efficacy of ChatGPT and other LLMs in abstract screening, particularly in relation to complex and diverse nursing practice issues and when using larger datasets. This evaluation focused exclusively on ChatGPT, the most popular LLM (Zhu 2025) and the LLM reportedly most used by academics (Lenharo 2024). However other LLMs are available, for example, Copilot, DeepSeek, Claude, and Gemini, and their unique understanding of nursing related issues is unknown. Without engagement with LLMs and dissemination of these findings via academic publication, nurses' risk being excluded in societal conversations about the useability and appropriateness of this technology (Abdulai and Hung 2023) and lose access to information to optimise evidenced-based practice. Already this study has demonstrated that relying upon research and insights from 'hard sciences' (e.g., providing ChatGPT the inclusion and exclusion criteria as a prompt without further refinement) is not directly appropriate for the complexity required for nursing specific issues. In our study, dedicated time was required to craft the prompt and gain insight into ChatGPT's 'understanding' of the concept prior to screening. Since LLMs are trained using documented internet sources (Collins 2025), there may be elements of nursing which are not adequately represented within these sources given the tacit knowledge and skills surrounding the nursing role (Castonguay et al. 2023; Allen 2024) and the ethical values involved (Abdulai and Hung 2023). In the technical report for ChatGPT-4 (OpenAI 2023), Open AI acknowledge that they aim to make their systems reflect users' values and obtain 'public input on what those bounds should be' (pp. 11). Therefore, nurses need to engage with, explore, and report on their interactions with ChatGPT to influence future iterations of the LLM to support evidence for practice. The risk is that nurses' pivotal position is underutilised in technological developments (Philips 2024) and without engaging with new technology nurses risk their professional development (Taskiran 2023) and ability to keep pace with an advancing evidence base. What may assist here is using AI as a collaborative tool (Nashwan and Abujaber 2023) and supportive partner (Shin et al. 2023) (akin to our knowledgeable companion example above), to extend and complement and nurses learning regarding AI (Jeyaraman et al. 2023).

8 | Implications for Policy and Practice

Evidenced-based practice is the cornerstone of safe and effective nursing care. Tools which help nurses to navigate the growing body of evidence have the potential to optimise nursing research and, in turn, enhance nursing practice more broadly. This study demonstrates that when prompted appropriately, with an understanding of ChatGPT's capabilities, nurses can use ChatGPT as a tool to assist in first line research abstract screening and reduce time spent screening. However, nurses must gain experience engaging with LLMs to use them optimally and this will require time. As LLMs develop, it is important that nurses, a group that hold a unique and important role within society, are involved in evaluating LLMs' abilities to solve nursing-specific issues and be part of societal discussions on the use of LLMs for healthcare purposes.

9 | Conclusion

To conclude, abstract screening is an essential aspect of academic literature reviews that are required to ensure nursing practice is based on the best available evidence. Manual abstract screening is a non-creative, yet cognitively demanding task, is subject to human error, is complicated by the exponential growth of academic literature and requires significant researcher time. For nursing researchers, these challenges are even more intense. This study shows how to leverage the abilities of LLMs to use ChatGPT and Python coding to enable rapid, accurate and scalable first-line abstract screening prior to human review. This proof-of-concept study demonstrates, for the first time, the ability of LLMs to screen abstracts in nursing research that involve complexity, diverse terminology, loosely defined boundaries, multifaceted practices, diverse roles, and variable contexts. This study demonstrates the constructive collaboration required between nursing researchers and LLMs to optimise abstract screening practices to answer the important questions that nurses have and to continue to optimise clinical practice and healthcare outcomes. Nurses are encouraged to engage with LLMs as research tools, ensuring that nurses have a professional voice in the future development of AI and to bring about greater efficiencies in bringing nursing research to practice.

Author Contributions

L.S. conceived the idea for the study. A.M. conducted the interactions with ChatGPT, screened the abstracts and worked with the interdisciplinary colleague on the Python script; these actions were taken in collaboration with all authors. A.M. wrote the paper with input from all authors (T.C., L.S., S.L.V., R.F. and A.G).

Acknowledgements

This work forms part of a PhD supported by the Erik Elgaard Sørensen Scholarship from the International Learning Collaborative. The authors thank Philip Mudd for assistance with Python coding. Open access publishing facilitated by Flinders University, as part of the Wiley - Flinders University agreement via the Council of Australian University Librarians.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

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

References

Abdulai, A. F., and L. Hung. 2023. “Will ChatGPT Undermine Ethical Values in Nursing Education, Research, and Practice.” *Nursing Inquiry* 30, no. 3: e12556.

Allen, D. 2014. “Re-Conceptualising Holism in the Contemporary Nursing Mandate: From Individual to Organisational Relationships.” *Social Science & Medicine* 119: 131–138.

Allen, D. 2024. *Care Trajectory Management for Nurses-E-Book*. Elsevier Health Sciences.

Alshami, A., M. Elsayed, E. Ali, A. E. E. Eltoukhy, and T. Zayed. 2023. “Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions.” *System* 11: 351.

Aromataris, E., C. Lockwood, K. Porritt, B. Pilla, and Z. Jordan. 2024. *JBI Manual for Evidence Synthesis*. JBI. https://synthesismanual.jbi.global.

Bannach-Brown, A., P. Przybyła, J. Thomas, et al. 2019. “Machine Learning Algorithms for Systematic Review: Reducing Workload in a Preclinical Review of Animal Studies and Reducing Human Screening Error.” *Systematic Reviews* 8: 1–12.

Benner, P. E., and J. Wrubel. 1989. *The Primacy of Caring: Stress and Coping in Health and Illness*. Addison-Wesley.

Bertocchi, L., A. Dante, C. La Cerra, et al. 2023. “Impact of Standardized Nursing Terminologies on Patient and Organizational Outcomes: A Systematic Review and Meta-Analysis.” *Journal of Nursing Scholarship* 55, no. 6: 1126–1153. https://doi.org/10.1111/jnu.12894.

Blaizot, A., S. K. Veettil, P. Saidoung, et al. 2022. “Using Artificial Intelligence Methods for Systematic Review in Health Sciences: A Systematic Review.” *Research Synthesis Methods* 13, no. 3: 353–362. https://doi.org/10.1002/jrsm.1553.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb. 2020. “Are Ideas Getting Harder to Find?” *American Economic Review* 110, no. 4: 1104–1144.

Boiko, D. A., R. MacKnight, B. Kline, and G. Gomes. 2023. “Autonomous Chemical Research With Large Language Models.” *Nature* 624, no. 7992: 570–578.

Bornmann, L., R. Haunschild, and R. Mutz. 2021. “Growth Rates of Modern Science: A Latent Piecewise Growth Curve Approach to Model Publication Numbers From Established and New Literature Databases.” *Humanities and Social Sciences Communications* 8, no. 1: 1–15.

Bornmann, L., and R. Mutz. 2015. “Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References.” *Journal of the Association for Information Science and Technology* 66, no. 11: 2215–2222.

Buresh, B., and S. Gordon. 2013. *From Silence to Voice: What Nurses Know and Must Communicate to the Public*. 3rd ed. ILR Press. https://doi.org/10.7591/9780801467370.

Castonguay, A., P. Farthing, S. Davies, et al. 2023. “Revolutionizing Nursing Education Through AI Integration: A Reflection on the Disruptive Impact of ChatGPT.” *Nurse Education Today* 129: 105916.

Collins, H. 2025. “Why Artificial Intelligence Needs Sociology of Knowledge: Parts I and II.” *Ai & Society* 40, no. 3: 1249–1263. https://doi.org/10.1007/s00146-024-01954-8.

Craig, J. V., and D. Dowding, eds. 2020. *Evidence-Based Practice in Nursing*. 4th ed. Elsevier.

Fui-Hoon Nah, F., R. Zheng, J. Cai, K. Siau, and L. Chen. 2023. “Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration.” *Journal of Information Technology Case and Application Research* 25, no. 3: 277–304.

Garcia, R., and I. Qureshi. 2022. “Nurse Identity: Reality and Media Portrayal.” *Evidence Based Nursing* 25, no. 1: 1–5.

Gartlehner, G., L. Affengruber, V. Titscher, et al. 2020. “Single-Reviewer Abstract Screening Missed 13 Percent of Relevant Studies: A Crowd-Based, Randomized Controlled Trial.” *Journal of Clinical Epidemiology* 121: 20–28.

Gartlehner, G., G. Wagner, L. Lux, et al. 2019. "Assessing the Accuracy of Machine-Assisted Abstract Screening With DistillerAI: A User Study." *Systematic Reviews* 8: 1–10.

Grbin, L., P. Nichols, F. Russell, M. Fuller-Tyszkiewicz, and C. A. Olsson. 2022. "The Development of a Living Knowledge System and Implications for Future Systematic Searching." *Journal of the Australian Library and Information Association* 71, no. 3: 275–292.

Hamel, C., S. E. Kelly, K. Thavorn, D. B. Rice, G. A. Wells, and B. Hutton. 2020. "An Evaluation of Distillersr's Machine Learning-Based Prioritization Tool for Title/Abstract Screening–Impact on Reviewer-Relevant Outcomes." *BMC Medical Research Methodology* 20: 1–14.

Hill, J. E., C. Harris, and A. Clegg. 2024. "Methods for Using Bing's AI-Powered Search Engine for Data Extraction for a Systematic Review." *Research Synthesis Methods* 15, no. 2: 347–353.

Hoelscher, S. H., K. Taylor-Pearson, and H. Wei. 2024. "Charting the Path: Nursing Leadership in Artificial Intelligence Integration Into Healthcare." *Nurse Leader* 22, no. 6: 763–772.

Hutson, M. 2024. "How Does ChatGPT 'Think'? Psychology and Neuroscience Crack Open AI Large Language Models." *Nature* 629, no. 8014: 986–988.

Hutson, M. 2024. "How Does ChatGPT 'Think'? Psychology and Neuroscience Crack Open AI Large Language Models." *Nature* 629, no. 8014: 986–988.

Issaiy, M., H. Ghanaati, S. Kolahi, et al. 2024. "Methodological Insights Into Chatgpt's Screening Performance in Systematic Reviews." *BMC Medical Research Methodology* 24, no. 1: 78. https://doi.org/10.1186/s12874-024-02203-8.

Jackson, J., J. E. Anderson, and J. Maben. 2021. "What Is Nursing Work? A Meta-Narrative Review and Integrated Framework." *International Journal of Nursing Studies* 122: 103944.

Jeyaraman, M., S. Ramasubramanian, S. Balaji, N. Jeyaraman, A. Nallakumarasamy, and S. Sharma. 2023. "ChatGPT in Action: Harnessing Artificial Intelligence Potential and Addressing Ethical Challenges in Medicine, Education, and Scientific Research." *World Journal of Methodology* 13, no. 4: 170–178. https://doi.org/10.5662/wjm.v13.i4.170.

Johnson, E. A., K. M. Dudding, and J. M. Carrington. 2024. "When to Err Is Inhuman: An Examination of the Influence of Artificial Intelligence Driven Nursing Care on Patient Safety." *Nursing Inquiry* 31: e12583.

Kalisch, B. J., S. Begeny, and S. Neumann. 2007. "The Image of the Nurse on the Internet." *Nursing Outlook* 55, no. 4: 182–188.

Kebede, M. M., C. Le Cornet, and R. T. Fortner. 2023. "In-Depth Evaluation of Machine Learning Methods for Semi-Automating Article Screening in a Systematic Review of Mechanistic Literature." *Research Synthesis Methods* 14, no. 2: 156–172.

Kerr, J. A., A. N. Gillespie, M. O'Connor, et al. 2023. "Intervention Targets for Reducing Mortality Between Mid-Adolescence and Mid-Adulthood: A Protocol for a Machine-Learning Facilitated Systematic Umbrella Review." *BMJ Open* 13, no. 10: e068733. https://doi.org/10.1136/bmjopen-2022-068733.

Khan, I. 2024. *The Quick Guide to Prompt Engineering: Generative AI Tips and Tricks for ChatGPT, Bard, Dall-E, and Midjourney*. 1st ed. John Wiley & Sons, Incorporated.

Khraisha, Q., S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield. 2024. "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." *Research Synthesis Methods* 15, no. 4: 616–626.

Kohler, K. 2024. "You Only Need to Change Your Direction: A Look at the Potential Impact of ChatGPT on Education." *Technology in Language Teaching & Learning* 6, no. 1: 1–18.

Kumar, P. 2024. "Large Language Models (LLMs): Survey, Technical Frameworks, and Future Challenges." *Artificial Intelligence Review* 57, no. 9: 260. https://doi.org/10.1007/s10462-024-10888-y.

Lefebvre, C., J. Glanville, S. Briscoe, et al. 2019. "Searching for and Selecting Studies." In *Cochrane Handbook for Systematic Reviews of Interventions*, 67–107. Wiley.

Lenharo, M. 2024. "ChatGPT Turns Two: How the AI Chatbot Has Changed Scientists' Lives." *Nature* 636, no. 8042: 281–282. https://www.nature.com/articles/d41586-024-03940-y.

Li, M., J. Sun, and X. Tan. 2024. "Evaluating the Effectiveness of Large Language Models in Abstract Screening: A Comparative Analysis." *Systematic Reviews* 13, no. 1: 219.

Lim, W. M., A. Gunasekara, J. L. Pallant, J. I. Pallant, and E. Pechenkina. 2023. "Generative AI and the Future of Education: Ragnarök or Reformation? A Paradoxical Perspective From Management Educators." *International Journal of Management Education* 21, no. 2: 100790.

Lin, X. 2024. "Exploring the Role of ChatGPT as a Facilitator for Motivating Self-Directed Learning Among Adult Learners." *Adult Learning* 35, no. 3: 156–166.

Lund, B. D., T. Wang, N. R. Mannuru, B. Nie, S. Shimray, and Z. Wang. 2023. "ChatGPT and a New Academic Reality: Artificial Intelligence-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing." *Journal of the Association for Information Science and Technology* 74, no. 5: 570–581.

May, R. 2021. *Making Research Matter, Chief Nursing Officer for England's Strategic Plan for Research*. National Health Service. https://www.england.nhs.uk/wp-content/uploads/2021/11/B0880-cno-for-englands-strategic-plan-fo-research.pdf.

Mollick, E. 2024. *Co-Intelligence: Living and Working With AI*. Penguin Publishing Group.

Nashwan, A. J., and A. A. Abujaber. 2023. "Nursing in the Artificial Intelligence (AI) Era: Optimizing Staffing for Tomorrow." *Cureus* 15, no. 10: e47275. https://doi.org/10.7759/cureus.47275.

Nazir, A., and Z. Wang. 2023. "A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges." *Meta-Radiology* 1, no. 2: 100022. https://doi.org/10.1016/j.metrad.2023.100022.

OpenAI. 2023. "GPT-4 Technical Report." arxiv.org.

Peters, M. D., C. Marnie, A. C. Tricco, et al. 2020. "Updated Methodological Guidance for the Conduct of Scoping Reviews." *JBI Evidence Synthesis* 18, no. 10: 2119–2126.

Philips, N. 2024. "Nurses Must Be Involved in the Development of Digital Tools." *Nursing Times* 120, no. 10: 15. https://www.nursingtimes.net/opinion/nurses-must-be-involved-in-the-development-of-digital-tools-02-10-2024/.

Qureshi, R., D. Shaughnessy, K. A. Gill, K. A. Robinson, T. Li, and E. Agai. 2023. "Are ChatGPT and Large Language Models 'the Answer' to Bringing Us Closer to Systematic Review Automation?" *Systematic Reviews* 12, no. 1: 72.

Shin, H., J. C. De Gagne, S. S. Kim, and M. Hong. 2023. "The Impact of Artificial Intelligence-Assisted Learning on Nursing Students' Ethical Decision-Making and Clinical Reasoning in Pediatric Care: A Quasi-Experimental Study." *Computers, Informatics, Nursing* 42: 10–1097.

Siedlecki, S. L., and N. M. Albert. 2017. "Research-Active Clinical Nurses: Against All Odds." *Journal of Clinical Nursing* 26, no. 5–6: 766–773.

Smith, P. 2012. *The Emotional Labour of Nursing Revisited. Can Nurses Still Care?* 2nd ed. Palgrave Macmillan.

Strachan, J. W., D. Albergo, G. Borghini, et al. 2024. "Testing Theory of Mind in Large Language Models and Humans." *Nature Human Behaviour* 8: 1–11.

Taskiran, N. 2023. "Effect of Artificial Intelligence Course in Nursing on Students' Medical Artificial Intelligence Readiness: A Comparative Quasi-Experimental Study." *Nurse Educator* 48, no. 5: E147–E152. https://doi.org/10.1097/NNE.0000000000001446.

Thomas, J., E. Flemyng, A. Noel-Storr, et al. 2004. "Responsible AI in Evidence Synthesis (RAISE): Guidance and Recommendations v.0.9 (draft)." Cochrane. https://osf.io/cn7x4.

Vaishya, R., A. Misra, and A. Vaish. 2023. "ChatGPT: Is This Version Good for Healthcare and Research?" *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* 17, no. 4: 102744.

Van De Schoot, R., J. De Bruin, R. Schram, et al. 2021. "An Open-Source Machine Learning Framework for Efficient and Transparent Systematic Reviews." *Nature Machine Intelligence* 3, no. 2: 125–133.

Van Noorden, R., and J. M. Perkel. 2023. "AI and Science: What 1,600 Researchers Think." *Nature* 621, no. 7980: 672–675.

Westwood, G., A. Richardson, S. Latter, J. Macleod Clark, and M. Fader. 2018. "Building Clinical Academic Leadership Capacity: Sustainability Through Partnership." *Journal of Research in Nursing* 23, no. 4: 346–357.

Zack, T., E. Lehman, M. Suzgun, et al. 2024. "Assessing the Potential of GPT-4 to Perpetuate Racial and Gender Biases in Health Care: A Model Evaluation Study." *Lancet Digital Health* 6, no. 1: e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X.

Zhu, K. 2025. "Ranked: Most Popular AI Tools by Monthly Site Visits. Visual Capitalist." https://www.visualcapitalist.com/ranked-most-popular-ai-tools-by-monthly-site-visits/.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.