

Automated title and abstract screening for scoping reviews using the GPT-4 Large Language Model

David Wilkins¹

¹Discipline of General Practice, The University of Adelaide

October 31, 2023

Abstract

Scoping reviews, a type of systematic literature review, require intensive human effort to screen large numbers of scholarly sources for their relevance to the review objectives. This manuscript introduces GPTscreenR, a package for the R statistical programming language that uses the GPT-4 Large Language Model (LLM) to automatically screen sources. The package makes use of the chain-of-thought technique with the goal of maximising performance on complex screening tasks. In validation against consensus human reviewer decisions, GPTscreenR performed similarly to an alternative zero-shot technique, with a sensitivity of 71%, specificity of 89%, and overall accuracy of 84%. Neither method achieved perfect accuracy nor human levels of intraobserver agreement. GPTscreenR demonstrates the potential for LLMs to support scholarly work and provides a user-friendly software framework that can be integrated into existing review pipelines.

1 Introduction

A scoping review is a relatively novel type of literature review that aims to map the key concepts and existing activity within an area of research [1]. Like systematic reviews, scoping reviews typically use rigorous, transparent [13], and sometimes pre-registered methods for gathering and synthesising evidence, and increasingly use formal frameworks for both performing and reporting reviews [11]. Scoping reviews can inform future system-

atic reviews or primary research in the same area [18]. However, they differ from systematic reviews in aiming to describe the breadth of coverage of the available literature rather than research findings in depth [1].

Frameworks for performing a scoping review typically involve defining a research area or question, searching bibliographic databases for potentially relevant published material ('sources'), screening these sources to identify those relevant to the area or question, and systematically extracting and reporting data from the relevant sources [20, 1]. The screening stage will usually involve initial screening of source titles and abstracts against predetermined inclusion and exclusion criteria, followed by full text screening, with both steps performed in replicate by at least two human reviewers [10, 12, 13]. Because database searches can return many hundreds or thousands of potentially relevant sources, these screening steps can require intensive human time and effort. Many software methods have been proposed or used to support or partially automate source screening for scoping reviews, including text mining to prioritise relevant sources for human screening [16, 5, 2], automated clustering and labelling of sources to support human decision-making [17], and 'crowdsourcing' screening to untrained workers via online platforms [8]. A similar but more extensive set of methods have been developed and employed for systematic reviews [6, 3] for which the process of source screening is broadly comparable.

Since the release of the first Generative Pre-trained Transformer (GPT) Large Language Model (LLM) by OpenAI (San Francisco, Califor-

nia, United States of America) in 2018 [15], transformer-based LLMs and the GPT lineage in particular have seen rapid and widespread adoption for a range of automation tasks. Broadly, these models generate a probabilistically weighted list of ‘tokens’ (parts of text such as letter combinations and punctuation) to continue or complete some input text (a ‘prompt’), having been trained to do so by practising such predictions on large human-written corpora. When this generative process is iterated, it allows for a range of applications involving analysis and production of text, such as summarising documents, generating fiction in a particular genre or style, or conversing with humans [9].

While LLMs are not yet widely used to screen sources for literature reviews, early work suggests they may perform well in this role. Guo et al. [4] reported the use of a GPT-lineage model (they do not specify which, though their published code suggests OpenAI’s ‘gpt-3.5-turbo’ model) to screen 24,307 titles and abstracts from five systematic reviews and one scoping review, achieving weighted average sensitivity of 76% and specificity of 91% when compared to human reviewers. Their approach involves giving the model a brief prompt instructing it to take on the persona of a researcher screening titles and abstracts, followed by a source’s title and abstract as well as the inclusion and exclusion criteria. The model is instructed to respond with a decision to include or exclude the source, and the process is iterated across the full set of sources to be screened. Syriani et al. [19] similarly reported the use of ‘gpt-3.5-turbo’ to screen titles and abstracts for a systematic review and achieved sensitivities of above 70%. They also systematically evaluated prompts given to the LLM to identify a prompt that performed best at the screening task; their chosen prompt, like that of Guo et al., placed the LLM in the role of an academic reviewer.

Both of these approaches made use of a single, fixed text prompt template, which the LLM then completes with additional text representing its response (the decision to include or exclude a source), a method sometimes called ‘zero-shot prompting’. Recent work has identified a number of methods which can be superior to zero-shot prompting when using LLMs for tasks that require complex or multi-step reasoning. These methods include ‘chain-of-thought prompting’ [21], in which a complex task is broken down into a series of intermediate steps,

and the ‘tree of thoughts’ strategy [23], in which multiple parallel chains of thought are generated, compared, and integrated. The LLM is induced to follow these complex reasoning strategies either by being given examples of multi-step reasoning on similar tasks, or by being lead through the process with a series of intermediate prompts.

In this paper, I introduce a package for the R programming language [14] called GPTscreenR that implements a chain-of-thought approach to using GPT-4 for scoping review title and abstract screening, and evaluate its performance by comparison to human reviewers. The purpose of this package is to assist and augment rather than replace human reviewers in performing scoping reviews. This package represents the first LLM-based screening tool designed specifically for scoping reviews. Further, this study provides the first report on the accuracy of LLM-based screening using the most recent iteration of the GPT model lineage, GPT-4, and using the recently developed chain-of-thought approach.

2 Methods

2.1 The GPTscreenR package

GPTscreenR is an R [14] package released under the MIT open source licence. The source code is available for download from GitHub at <https://github.com/wilcox/GPTscreenR>. At the time of writing the most recent package version was 0.0.3, and the results presented in this paper were obtained using this version.

GPTscreenR consists of two main components. The first is a set of internal functions for interfacing with the OpenAI Application Programming Interface (API), which allows for ‘conversations’ with the GPT-4 LLM, as well as functions for representing and manipulating those conversations. These internal functions are designed to be model-agnostic, so that future versions of the package or users with particular needs can use different GPT models. The OpenAI API requires an OpenAI account, and OpenAI charges fees for use of the API. In order to access the API, GPTscreenR requires a secret key to be registered prior to source screening, and instructions for doing so are provided in the package documentation and on loading of the package in R if the key has not been correctly reg-

istered.

The second component is a set of user-facing functions to perform source screening with GPT-4. The `review_description()` function assists in generating a text description of the review’s objectives and inclusion and exclusion criteria, using the Population, Concept, and Context (PCC) framework [10, 12] for defining the review’s inclusion and exclusion criteria. The use of this function is optional, and users may instead choose to provide a description of the review and criteria for source selection using any framework or format they see fit.

The `screen_source()` function performs the main task of the package. This function mediates an conversation with GPT-4 in which chain-of-thought prompting [21] is used to guide GPT-4 through screening a source title and abstract against the study inclusion criteria. The template for this conversation is given in Figure 1. The OpenAI API defines a conversation as a series of messages, each of which originates from one of three roles: *system*, representing an authoritative voice that can instruct GPT-4 on its task and behaviour; *user*, representing a human user that can interact with GPT-4; and *assistant*, representing the responses generated by GPT-4. In `screen_source()`, the *system* role gives GPT-4 general instructions, while the *user* role provides the user-written review description and the source title and abstract.

The specific phrase ‘work step by step’ is derived from the ‘let’s think step by step’ prompt phrase which significantly improves LLM performance on multi-step reasoning tasks with a zero-shot prompt [7], adapted to this chain-of-thought approach.

GPT-4 is instructed to summarise the inclusion criteria for the scoping review, compare the title and abstract against these summarised criteria, and make a final recommendation on whether to include or exclude the source (Figure 1). This approach was chosen after noting that a major source of error when attempting to screen sources with zero-shot, one-shot, or few-shot prompts (i.e. with a single prompt and no, one, or a few examples) was that GPT-4 would fail to consider important inclusion criteria. As an example, Figure 2a presents a conversation with GPT-4 using a zero-shot prompt. The screening task in this example is intentionally adversarial and designed to lead the model towards making an error. In order to correctly recommend exclusion of the source, GPT-4 must notice that

the review is looking for research on therapy alpaca interventions, but that the source reports on a therapy camel intervention. The presence of multiple other inclusion criteria which are met by the source, as well as the mention of alpacas in the source abstract, serve as distractors. In this example, GPT-4 incorrectly recommends inclusion. If the conversation is then continued to draw GPT-4’s attention to the error, it is able to identify and correct it (Figure 2b), suggesting that the error arises from a failure of GPT-4 to properly consider the relevant inclusion criterion rather than an inability to do so. Using the chain-of-thought approach overcomes this problem without the need for human intervention and correction (Figure 2c). GPT-4 identifies ‘The context of the study must involve the use of therapy alpacas as a part of a programme of care in residential aged care facilities.’ as an inclusion criterion, and correctly assesses that this is the only inclusion criterion not met. GPT-4 then correctly recommends exclusion of the source.

The `screen_source()` function returns a list comprising the complete transcript of the conversation with GPT-4 and GPT-4’s final recommendation as a logical (Boolean) value. The conversation transcript can be used to interrogate cases where GPT-4 may have returned an incorrect or unexpected result. The package also provides a function `screen_sources()`, which applies `screen_source()` iteratively across a data frame of sources. `screen_sources()` caches screening outcomes to a file as they accumulate, so that screening can be split across multiple sessions and recover from interruptions.

2.2 Validation

To validate GPTscreenR’s approach, six scoping reviews were identified from the Open Science Framework (OSF; <https://osf.io>) where the review inclusion criteria and the results of title and abstract screening were publicly available. A summary of the review characteristics is provided in Table 1. Random subsets were used where the large number of sources made it prohibitive in time and cost to screen the full set. The total number of sources available for screening and the number used for validation from each review are given in Table 1.

Some of the reviews did not include source full abstracts in the publicly available files, and when

System: You are helping academic researchers perform a scoping review. Your task is to screen a single source against the review criteria. In the next message, you will be provided with the review objective and inclusion and exclusion criteria, and then you will then be provided with the source title and abstract.

User: *The user-provided review description.*

System: You must work step by step. FIRST, generate a numbered list of criteria that must be met for a source to be included.

GPT-4: *GPT-4 replies with a numbered list of criteria.*

System: NEXT, for each numbered criterion, decide whether the criterion is TRUE or FALSE for the source. It is normal for the title and abstract to not have enough information to make a clear decision for every statement. For these situations, give your best guess. After giving your response of TRUE or FALSE, give a one sentence explanation for your response.

GPT-4: *GPT-4 replies with a decision and explanation for each criterion.*

System: FINALLY, consider your decisions on whether the source meets the conclusion criteria. Respond with a single word, either INCLUDE or EXCLUDE, representing your recommendation on whether the source meets the inclusion criteria. Do not write anything other than INCLUDE or EXCLUDE.

GPT-4: *GPT-4 replies with its final recommendation.*

Figure 1: Template for the conversation with GPT-4 mediated by the `screen_source()` function. Messages with variable content, including user-provided data as well as GPT-4's responses, are given in italics.

User: You are being used to help researchers perform a scoping review. A scoping review is a type of systematic review used to map the published scholarship on a topic. To gather relevant sources for a scoping review, the researchers search bibliographic databases for sources that match a selected Population, Concept, and Context (the inclusion criteria). The titles and abstracts of sources that are found in this search are then screened against the inclusion criteria.

This review has the following objective and inclusion criteria:

OBJECTIVE: This scoping review will examine the existing primary research on the role of therapy alpacas in enhancing the mental health and emotional well-being of elderly residential aged care facility residents

POPULATION: Elderly people living in residential aged care facilities. 'Elderly' is defined as 65 years of age or older.

CONTEXT: Residential aged care facilities that have used therapy alpacas as part of their programme of care for residents. 'Residential aged care facility' is defined as a residential setting intended as a long-term place of residence for elderly people which includes provision of support for activities of daily living (e.g. meal preparation, bathing, housekeeping) and nursing support (e.g. medication management). Such facilities will also typically offer other structured programmes and facilities to provide entertainment, diversion, and wellbeing. It excludes other residential settings intended for elderly people that do not provide daily living or nursing supports (e.g. independent living villages) or that are not long-term (e.g. hospitals or hospices).

CONCEPT: The impact of therapy alpaca programmes on stress reduction, emotional well-being, mental health, overall life satisfaction, or similar outcomes for residents.

Here is the title and abstract of the source for you to screen:

TITLE: Therapy Ungulate Visits Reduce Depression, Anxiety, and Stress in Australian Nursing Home Residents

ABSTRACT: Nursing home residents typically have rates of depression, anxiety, and stress (DAS) significantly higher than those of equivalent populations living independently. As well as reducing quality of life and increasing morbidity and mortality, DAS significantly increases the burden on both nursing homes and health systems in caring for the elderly. Recent pilot trials of the use of therapy ungulates (alpacas, llamas, and camels) have shown promising initial results in reducing DAS among nursing home residents. We conducted a randomised control trial involving 8,124 residents across 17 nursing homes in urban and rural Australia in which residents were randomised to either a weekly 2-hour therapy camel visit or equivalent time in a television viewing lounge. Residents receiving the intervention had significantly ($p < 0.01$) reduced levels of depression (mean DASS-21 score for depression reduction 8.2 points), anxiety (mean reduction 11 points) and stress (mean reduction 13 points) compared to the control arm. This study represents the first large-scale randomised trial of ungulate therapy for nursing home residents, and provides strong support for its effectiveness in reducing DAS.

Please respond with a single word, either INCLUDE or EXCLUDE, representing your recommendation

GPT-4: INCLUDE

Figure 2a: A conversation with GPT-4, using a zero-shot prompt to instruct the model to screen a single source for a hypothetical scoping review. This adversarial task has been intentionally designed to influence GPT-4 towards making an error.

Table 1: Scoping reviews used to validate the performance of GPTscreenR. ‘total available’ is the number of sources available for screening following attempted retrieval of missing abstracts, correction of malformed data, and removal of sources with missing, irreparable, or duplicate data; it may differ from the number reported by the review authors.

Short name	OSF project title	OSF URL	File(s)	Sources screened (total available)
‘COVID’	Complementary, Alternative, and Integrative Medicine-Specific COVID-19 Misinformation on Social Media: A Scoping Review	https://osf.io/ytz5e	‘TITLE ABSTRACT FULL TEXT SCREENING DATA_CAIM COVID-19 SM Misinfo_Mar2223.xlsx’	324 (458)
‘melanoma’	Scoping Review – Melanoma	https://osf.io/knje4	‘Melanoma Data.xlsx’	72 (128)
‘smartphones’	Smartphone Addiction Scoping Review	https://osf.io/f9huw	‘Scop_rev_full.csv’	256 (5,376)
‘solastalgia’	Australian Solastalgia Scoping Review	https://osf.io/qxe3n	‘Solastalgia re-screening and extraction spreadsheet Dec 2022.xlsx’	150 (150)
‘suicide’	Social Norms and Suicidality - Scoping Review	https://osf.io/btpzc	‘June 2022 Screenings Extraction TOP UP.xlsx’, ‘Nov20 Stage 2 Removal of duplicates and screenings.xlsx’	100 (2,094)
‘teachers’	Teachers’ soft skills: a Scoping Review	https://osf.io/n9rkd	‘Teachers’ Soft Skills a SR - Screening NM & OJ(1).xlsx’	245 (355)

User: This doesn't seem right. The inclusion criteria includes the concept of therapy alpaca programmes, while the abstract describes a study of a therapy camel programme. Can you explain your reasoning?

GPT-4: I apologize for the oversight. You're right. While the abstract mentions "therapy ungulates" (which includes alpacas, llamas, and camels), the specific intervention studied was therapy camel visits, not alpacas. Given the specific criteria you've set for the scoping review, this study does not directly address the role of therapy alpacas.

Based on this clarification, the correct recommendation would be:

EXCLUDE.

Figure 2b: A continuation of the conversation in Figure 2a, in which GPT-4 is lead to recognise and correct the error.

these abstracts could not be obtained from external databases these sources were excluded from validation. There were also many cases where missing, malformed, or duplicate data required either manual correction or exclusion of sources. The scoping review data, code used to prepare this data for validation, and code for calculating summary statistics are available in a reproducible form in the package repository on GitHub (<https://github.com/wilkox/GPTscreenR/tree/master/validation>).

The consensus human reviewer decision at the title and abstract screening level was used as the gold standard outcome. Accuracy, sensitivity and specificity were calculated by comparing GPT-4's recommendation against this gold standard. Three of the scoping reviews (COVID, solastalgia, and teachers) included individual human reviewer decisions in addition to the final decision in their publicly available datasets, and these were used to calculate human intraobserver agreement (Cohen's kappa) using the R function `cohen.kappa()` from the `psych` package [22]. This was compared to human/GPT-4 agreement across all screened sources, calculated with the same method.

2.3 Comparison to zero-shot method

In order to directly compare the chain-of-thoughts approach to a zero-shot approach (i.e. a conversation consisting of a single prompt with no examples, followed by GPT-4's response), the validation screening task was repeated using the prompt designed by Guo et al. [4], substituting the permit-

ted responses **INCLUDE** and **EXCLUDE** for **included** and **excluded** respectively in order to maintain compatibility with GPTscreenR's parsing of the response (Figure 3). In cases where a random subset of sources had been used for validation of GPTscreenR, the same subset was used. The code used to prepare this data and calculate summary statistics was otherwise identical to that used for the chain-of-thoughts method validation and is available in the 'zeroshot' branch of the package repository on GitHub (<https://github.com/wilkox/GPTscreenR/tree/zeroshot>).

3 Results

1,147 sources were screened from the six scoping reviews. GPTscreenR achieved an overall accuracy of 84% compared to the gold standard of the consensus human reviewer decision, with weighted average sensitivity of 71% and weighted average specificity of 89% (Figure 4). For the three reviews that provided individual reviewer decisions, the weighted average Cohen's kappa was 0.67, while the weighted average Cohen's kappa between final human and GPT-4 decisions was 0.52.

The zero-shot method achieved an overall accuracy of 83%, with weighted average sensitivity of 72% and weighted average specificity of 87%. The weighted average Cohen's kappa between human and GPT-4 decisions was 0.52.

System: You are helping academic researchers perform a scoping review. Your task is to screen a single source against the review criteria. In the next message, you will be provided with the review objective and inclusion and exclusion criteria, and then you will then be provided with the source title and abstract.

User: **OBJECTIVE:** This scoping review will examine the existing primary research on the role of therapy alpacas in enhancing the mental health and emotional well-being of elderly residential aged care facility residents

POPULATION: Elderly people living in residential aged care facilities. 'Elderly' is defined as 65 years of age or older.

CONTEXT: Residential aged care facilities that have used therapy alpacas as part of their programme of care for residents. 'Residential aged care facility' is defined as a residential setting intended as a long-term place of residence for elderly people which includes provision of support for activities of daily living (e.g. meal preparation, bathing, housekeeping) and nursing support (e.g. medication management). Such facilities will also typically offer other structured programmes and facilities to provide entertainment, diversion, and wellbeing. It excludes other residential settings intended for elderly people that do not provide daily living or nursing supports (e.g. independent living villages) or that are not long-term (e.g. hospitals or hospices).

CONCEPT: The impact of therapy alpaca programmes on stress reduction, emotional well-being, mental health, overall life satisfaction, or similar outcomes for residents.

TITLE: Therapy Ungulate Visits Reduce Depression, Anxiety, and Stress in Australian Nursing Home Residents

ABSTRACT: Nursing home residents typically have rates of depression, anxiety, and stress (DAS) significantly higher than those of equivalent populations living independently. As well as reducing quality of life and increasing morbidity and mortality, DAS significantly increases the burden on both nursing homes and health systems in caring for the elderly. Recent pilot trials of the use of therapy ungulates (alpacas, llamas, and camels) have shown promising initial results in reducing DAS among nursing home residents. We conducted a randomised control trial involving 8,124 residents across 17 nursing homes in urban and rural Australia in which residents were randomised to either a weekly 2-hour therapy camel visit or equivalent time in a television viewing lounge. Residents receiving the intervention had significantly ($p < 0.01$) reduced levels of depression (mean DASS-21 score for depression reduction 8.2 points), anxiety (mean reduction 11 points) and stress (mean reduction 13 points) compared to the control arm. This study represents the first large-scale randomised trial of ungulate therapy for nursing home residents, and provides strong support for its effectiveness in reducing DAS.

System: You must work step by step. FIRST, generate a numbered list of criteria that must be met for a source to be included.

GPT-4:

- 1.The research study must involve elderly people aged 65 years or older.
- 2.The study's population must be residents of residential aged care facilities.
- 3.The context of the study must involve the use of therapy alpacas as a part of a programme of care in residential aged care facilities.
- 4.The study must examine the impact of therapy alpaca programmes on outcomes such as stress reduction, emotional well-being, mental health, or overall life satisfaction for the residents.
- 5.The study must be primary research, not purely theoretical or hypothetical work. It must be based on new research and data collection, not a review or summary of previous studies.

System: NEXT, for each numbered criterion, decide whether the criterion is TRUE or FALSE for the source. It is normal for the title and abstract to not have enough information to make a clear decision for every statement. For these situations, give your best guess. After giving your response of TRUE or FALSE, give a one sentence explanation for your response.

GPT-4:

- 1.TRUE - The research study involves elderly individuals as it is focused on nursing home residents, who are typically over 65 years of age.
- 2.TRUE - The population of the study constitutes residents of nursing homes, which are a type of residential aged care facilities.
- 3.FALSE - The study does involve therapy ungulates, but it specifically mentions therapy camels being used, not therapy alpacas.
- 4.TRUE - The study examines the impact of therapy ungulate programmes on reducing depression, anxiety, and stress in nursing home residents, which falls within the scope of emotional well-being and mental health outcomes.
- 5.TRUE - The study is primary research, as it reports fresh findings from a randomized control trial rather than summarizing or reviewing previous studies.

System: FINALLY, consider your decisions on whether the source meets the conclusion criteria. Respond with a single word, either INCLUDE or EXCLUDE, representing your recommendation on whether the source meets the inclusion criteria. Do not write anything other than INCLUDE or EXCLUDE.

GPT-4: EXCLUDE

Figure 2c: A conversation with GPT-4, instructing it to perform the same adversarial task presented in Figure 2a, but using the chain-of-thought approach. This is not a continuation of the conversation presented in Figures 2a and 2b but a new conversation. In this case, GPT-4 correctly recommends excluding the source. The model also correctly identifies that the source meets all of the relevant inclusion criteria except for the requirement that the source report on an alpaca intervention.

System: Instructions: You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article. Only type “INCLUDE” or “EXCLUDE” to indicate your decision. Do not type anything else.

Title: *The source title.*

Abstract: *The source abstract.*

The review description.

GPT-4: *GPT-4 replies with a decision.*

Figure 3: Zero-shot prompt used as a comparator for the chain-of-thoughts approach, derived from the approach of Guo et al. [4]. Variable content, including user-provided data as well as GPT-4’s response, is given in italics.

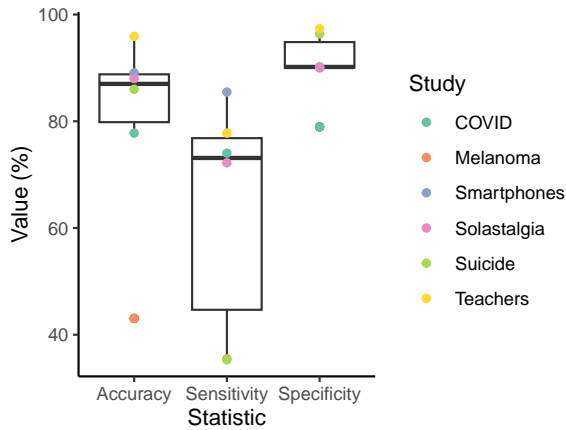


Figure 4: Accuracy, sensitivity, and specificity of GPTscreenR compared to the gold standard of the consensus human reviewer decision. The box-and-whisker plot represents for each statistic the median (central bar), 25th and 75th percentiles (upper and lower hinges of the box), and largest value $\pm 1.5 \times$ the interquartile range (whiskers). Points given the values for the individual scoping reviews used for validation.

4 Discussion

4.1 Performance of GPTscreenR

GPTscreenR and its chain-of-thought approach performed comparably to the zero-shot method reported by Guo et al. [4]. Also similarly to their report, Cohen’s kappa was higher for inter-human than human-GPT-4 agreement, suggesting that despite this generally good performance GPT-based methods still do not match the performance of a human reviewer when the consensus human decision is used as the gold standard.

4.2 Comparison to the zero-shot method

The replication of Guo et al.’s zero-shot prompt on the reviews used for validation in this study resulted in similar weighted average sensitivity (78%) compared to their report (76%), though somewhat poorer specificity (84% compared to 91%). There are a number of possible factors contributing to this difference. Firstly, this study used the GPT-4 model while Guo et al. likely used the ‘gpt-3.5-turbo’ model, although it might be expected that the more advanced GPT-4 would generally perform better on the same task. Secondly, this study included only scoping reviews, while Guo et al. examined five systematic reviews and only one scoping review, although the reported sensitivity (100%) and specificity (94%) for that scoping review were

higher than the weighted average for this study. Thirdly, validation with both the chain-of-thoughts and zero-shot methods was dependent on the availability of source and screening data from the included reviews, and it was noted that most of the reviews had issues with missing, malformed, missing, and duplicate data, including lacking the text of some or all source abstracts and sufficient information to retrieve them from public databases. While effort was made to rectify these issues, many sources from the included reviews could not be used for validation, and the differing quality of the datasets used in the two studies may have affected the performance of the zero-shot method. Finally, the different reviews used for validation in the two studies likely posed different levels of difficulty the zero-shot approach. If LLM-based approaches to source screening become more widely adopted, the performance of different methods can be more accurately determined across a growing sample of reviews.

Compared against the zero-shot method, the chain-of-thought method achieved poorer sensitivity but higher specificity. This result is consistent with rationale for selecting the chain-of-thought method, which was to reduce type I errors (false positives) of the type demonstrated in Figure 2a. However, this came at the cost of higher type II error. It is not possible to retrospectively examine the reasoning process that lead GPT-4 to make an incorrect recommendation such as a false negative, as asking it to explain its reasoning post-hoc will result in a confabulated statistically likely response; as they have no direct access to their internal processes, LLMs are unable to meaningfully introspect. However, unlike the zero-shot method, the chain-of-thought method as implemented by GPTscreenR does allow reviewers to view and assess the reasons for GPT-4’s final decision, by reviewing the transcript of the conversation with GPT-4 and in particular its assessment of the source against its summarised list of criteria. This may be useful for resolving disagreements between human reviewers and GPT-4, or disagreements among human reviewers.

4.3 Limitations of this approach

Compared to the prompts used by Guo et al. [4] and Syriani et al. [19], both the prompt used for

this approach and the response generated by GPT-4 have substantially higher token counts. This results in both a longer time to screen a single source (typically 20–30 seconds) and a higher dollar cost, as the OpenAI API currently charges on a per-token basis. This may make GPTscreenR less appealing to some reviewers, particularly when screening a large number of sources.

Validation of GPTscreenR was limited by a small number of scoping reviews with publicly available, high-quality data on human title and abstract screening decisions. Because a subset of these reviews were used for both testing and validation, there is a risk of over-fitting of the prompt to these particular reviews. The public release of GPTscreenR encourages users to contribute data from their own scoping reviews to support more accurately measuring GPTscreenR’s real-world performance, and to guide further refinement of the approach.

5 Conclusions

The use of Large Language Models to screen sources for scoping reviews is a promising technology to reduce the human time and effort required to perform these reviews. The GPTscreenR package performs comparably to a zero-shot based approach to source screening. However, as with the zero-shot approach, agreement between its decisions and the human reviewer consensus fell short of inter-human agreement, suggesting that LLM-based screening is not yet as reliable as a human reviewer. Future work may help to further quantify and refine GPTscreenR’s performance, and to expand this approach to other types of scholarly review and to other stages of the review process.

6 References

References

- [1] Hilary Arksey and Lisa O’Malley. “Scoping studies: towards a methodological framework”. In: *International Journal of Social Research Methodology* 8.1 (2005), pp. 19–32. DOI: 10.1080/1364557032000119616.

- [2] Kevin E. K. Chai et al. "Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews". In: *Systematic Reviews* 10.1 (2021), p. 93. DOI: 10.1186/s13643-021-01635-3.
- [3] Allison Gates et al. "Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools". In: *Systematic Reviews* 8.1 (2019), p. 278. DOI: 10.1186/s13643-019-1222-2.
- [4] Eddie Guo et al. "Automated Paper Screening for Clinical Reviews Using Large Language Models". In: *arXiv* (2023). DOI: 10.48550/arxiv.2305.00844.
- [5] Brian E. Howard et al. "SWIFT-Review: a text-mining workbench for systematic review". In: *Systematic Reviews* 5.1 (2016), p. 87. DOI: 10.1186/s13643-016-0263-z.
- [6] Hanan Khalil, Daniel Ameen, and Armita Zarnegar. "Tools to support the automation of systematic reviews: a scoping review". In: *Journal of Clinical Epidemiology* 144 (2022), pp. 22–42. DOI: 10.1016/j.jclinepi.2021.12.005.
- [7] Takeshi Kojima et al. "Large Language Models are Zero-Shot Reasoners". In: *arXiv* (2022). DOI: 10.48550/arxiv.2205.11916.
- [8] Michael L. Mortensen et al. "An exploration of crowdsourcing citation screening for systematic reviews". In: *Research Synthesis Methods* 8.3 (2017), pp. 366–386. DOI: 10.1002/jrsm.1252.
- [9] OpenAI. *Introducing ChatGPT*. "https://openai.com/blog/chatgpt". 2023.
- [10] Micah D. J. Peters et al. "Updated methodological guidance for the conduct of scoping reviews". In: *JBIM Evidence Synthesis* 18.10 (2020), 2119–2126. DOI: 10.1112/jbies-20-00167.
- [11] Micah D. J. Peters et al. "Scoping reviews: reinforcing and advancing the methodology and application". In: *Systematic Reviews* 10.1 (2021), p. 263. DOI: 10.1186/s13643-021-01821-3.
- [12] Micah DJ Peters et al. "Chapter 11: Scoping Reviews (2020 version)". In: *JBIM Manual for Evidence Synthesis*. Ed. by E Aromataris and Z Munn. JBI, 2020. ISBN: 9780648848806. DOI: 10.46658/jbimes-20-12. URL: https://synthesismanual.jbi.global/.
- [13] Mai T. Pham et al. "A scoping review of scoping reviews: advancing the approach and enhancing the consistency". In: *Research Synthesis Methods* 5.4 (2014), pp. 371–385. DOI: 10.1002/jrsm.1123.
- [14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.
- [15] Alec Radford et al. *Improving language understanding by generative pre-training*. "https://openai.com/research/language-unsupervised". 2018.
- [16] Ian Shemilt et al. "Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews". In: *Research Synthesis Methods* 5.1 (2014), pp. 31–49. DOI: 10.1002/jrsm.1093.
- [17] Claire Stansfield, James Thomas, and Josephine Kavanagh. "'Clustering' documents automatically to support scoping reviews of research: a case study". In: *Research Synthesis Methods* 4.3 (2013), pp. 230–241. DOI: 10.1002/jrsm.1082.
- [18] Anthea Sutton et al. "Meeting the review family: exploring review types and associated information retrieval requirements". In: *Health Information & Libraries Journal* 36.3 (2019), pp. 202–222. DOI: 10.1111/hir.12276.
- [19] Eugene Syriani, Istvan David, and Gauransh Kumar. "Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews". In: *arXiv* (2023).
- [20] The Joanna Briggs Institute. *The Joanna Briggs Institute Reviewers' Manual 2015: Methodology for JBI Scoping Reviews*. The Joanna Briggs Institute, 2015.

- [21] Jason Wei et al. “Chain of Thought Prompting Elicits Reasoning in Large Language Models”. In: *arXiv* (2022). DOI: 10.48550/arxiv.2201.11903.
- [22] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.3.6. Northwestern University. Evanston, Illinois, 2023. URL: <https://CRAN.R-project.org/package=psych>.
- [23] Shunyu Yao et al. “Tree of Thoughts: Deliberate Problem Solving with Large Language Models”. In: *arXiv* (2023). URL: <https://arxiv.org/abs/2305.10601>.