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

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 (Arksey and O'Malley 2005). Like systematic reviews, scoping reviews typically use rigorous, transparent (Pham et al. 2014), and sometimes pre-registered methods for gathering and synthesising evidence, and increasingly use formal frameworks for both performing and reporting reviews (Micah D. J. Peters et al. 2021). Scoping reviews can inform future systematic reviews or primary research in the same area (Sutton et al. 2019). However, they differ from systematic reviews in aiming to describe the breadth of coverage of the available literature rather than research findings in depth (Arksey and O'Malley 2005).

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 (Institute 2015; Arksey and O'Malley 2005). The screening stage will usually involve initial screening of source titles and abstracts against pre-determined inclusion and exclusion criteria, followed by screening of source full texts, with both steps performed in replicate by at least two human reviewers (Micah D. J. Peters et al. 2020; Pham et al. 2014). Because database searches can return many hundreds or thousands of potentially relevant sources, these screening steps can require intensive human effort and time. Many software methods have been proposed or used to support or partially automate source screening for scoping reviews, including text mining to prioritise potentially more relevant sources for human screening (Shemilt et al. 2014; Howard et al. 2016; Chai et al. 2021), automated clustering and labelling of sources to support human decision-making (Stansfield, Thomas, and Kavanagh 2013), and 'crowdsourcing' screening to untrained workers via online platforms (Mortensen et al. 2017). A similar but more extensive set of methods have been developed and employed for systematic reviews (Khalil, Ameen, and Zarnegar 2022; Gates et al. 2019) 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, California, United States of America) in 2018 (Radford et al. 2018), 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) which may continue or complete some input text (a 'prompt'), having been trained to do so by practising such predictions against 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 (OpenAI 2023).

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. (Guo et al. 2023) 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. (Syriani, David, and Kumar 2023) 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' (Wei et al. 2022), in which a complex task is broken down into a series of intermediate steps, and the 'tree of thoughts' strategy (Yao et al. 2023), 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 (R Core Team 2023) called GPTscreenR that implements a chain-of-thought based 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 a novel development in the use of LLMs for source screening in literature reviews, as it provides an open-source and immediately available screening tool which can be downloaded, used, and modified by academic reviewers, and is 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 to maximise screening accuracy.

Methods

The GPTscreenR package

GPTscreenR is an R (R Core Team 2023) package released under the MIT open source licence. The source code is available for download from GitHub at https://github.com/wilkox/GPTscreenR. At the time of writing the most recent package version was 0.0.1; the results presented in this paper were obtained with an earlier version 0.0.0.9005, since which small changes have been made to further refine performance.

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 registered.

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) frame (Micah D. J. Peters et al. 2020) 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 (Wei et al. 2022) 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 Fig. 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 phrase 'work step by step', used each time *system* gives GPT-4 a prompt with specific instructions, 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 (Kojima et al. 2022), adapted to this chain-of-thought approach. This approach thus attempts to maximise engagement of GPT-4's multi-step reasoning capabilities both across and within each step in the screening task.

GPT-4 is serially instructed to summarise the inclusion criteria for the scoping review (prior to being presented

System: You are being used to help researchers perform a scoping review. You are not interacting directly with a user.

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 search are then screened against the inclusion criteria.

Your task is to screen a single source against the inclusion criteria. You will be provided with the review objective and inclusion criteria, and then you will then be provided with the source title and abstract. You will then be instructed to work step by step through the process of comparing the source against the inclusion criteria. Finally, you will instructed to make a recommendation on whether the source should be included.

The next message will be from the user, and will contain the scoping review objective and inclusion criteria.

User: *The user-generated review description.*

System: Let's work step by step. First, generate a numbered list of statements that summarise the inclusion criteria for the scoping review, including the Population, Concept, and Context. The statements should be clear, comprehensive and complete. Any source for which all the statements are true is a source that meets the inclusion criteria. As a template, here are some example statements (these are a generic set of examples that are not related to the current scoping review):

1. The source reports the results of a randomised control trial
2. The source reports the results of a study in which:
   2a. The participants were all male; AND
   2b. The participants were all aged between 18 and 74 inclusive
3. The source reports the results of a study conducted in the European Union.

Aspects of the inclusion criteria with multiple elements should be broken down into separate points where possible. For example, instead of:

1. The source reports on a study of men who live in the European Union.

You should instead say:

1. The source reports on a study of people who are:
   1a. Male; and
   1b. Living in the European Union.

GPT-4: *GPT-4 replies with a summary of the inclusion criteria.*

System: The next message will be from the user, and will contain the title and abstract of a source to be compared against the inclusion criteria.

User: *The source title and abstract.*

System: Let's continue to work step by step. Refer back to the set of statements you developed summarising the inclusion criteria. For each statement, decide whether or not the statement is true for the source described by the title and abstract. You must select from the following permitted responses: TRUE, FALSE, LIKELY TRUE, LIKELY FALSE, or NOT APPLICABLE. No other response is permitted. It is normal for the title and abstract to not have enough information to make a clear decision for every statement. There is a natural and normal amount of ambiguity in this process. For these situations, give your best guess, making use of your general knowledge, and deciding LIKELY TRUE or LIKELY FALSE. Responses like UNCLEAR or NOT ENOUGH INFORMATION are not permitted. After giving your response, give a one sentence explanation for your response.

GPT-4: *GPT-4 replies with a decision and explanation for each of its summarised inclusion criteria.*

System: Let's continue to work step by step. Consider your decisions on whether the title and abstract meet the conclusion criteria. Overall, is it likely true that the source meets the inclusion criteria? Reply with a single word, either INCLUDE or EXCLUDE, representing your recommendation on whether the source is likely to meet the inclusion criteria. You must reply with a single word only and it must be one of these two words; any other reply will cause the automatic parsing of your response to fail, which will be troublesome for the user.

GPT-4: *GPT-4 replies with either INCLUDE or EXCLUDE.*

Figure 1: Template for the conversation with GPT-4, mediated by the GPTscreenR package. Messages with variable content, including user-provided data as well as GPT-4's responses, are given in italics. The typographic error in the first message ('search search') appeared in the version of the package used for the results reported in this study, but has been since corrected.

with the title and abstract to be screened), compare the title and abstract against these summarised criteria, and make a final recommendation on whether to include or exclude the source (Fig. 1). Following the chain-of-thought approach, an example of summarised inclusion criteria is given for GPT-4 to use as a template or exemplar. 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, Fig 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 (Fig 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 (Fig 2c). GPT-4 identifies 'The source examines the impact of therapy alpaca programmes' 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, either the word 'INCLUDE' or 'EXCLUDE'. 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.

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. Small random subsets of screened sources from four of the reviews (`COVID`, `smartphones`, `solastalgia`, and `teachers`) were used during initial testing and refinement of the `screen_source()` function, while the full set or random subsets from all six reviews were used for final validation. 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 the full abstract text of sources in the publicly available files, and where 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 final human reviewer decision at the title and abstract screening level was used as the gold standard outcome. 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 (William Revelle 2023). This was compared to human/GPT-4 agreement across all screened sources, calculated with the same method.

Comparison to zero-shot method

In order to directly compare the chain-of-thoughts approach to a zero-shot approach (i.e. an 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. (Guo et al. 2023), substituting the permitted responses `INCLUDE` and `EXCLUDE` for `included` and `excluded` respectively in order to maintain compatibility with GPTscreenR's parsing of the response (Fig. 3). The number of sources used for zero-shot validation were the same as those used for the chain-of-thoughts method validation given in Table 2, although a different subset was selected by the randomisation process in cases where the full set was not used. The code used

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 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 no 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.

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 Fig. 2a, in which GPT-4 is lead to recognise and correct the error.

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).

Results

1,138 sources were screened from the six scoping reviews. GPTscreenR achieved a weighted average sensitivity of 0.75 and weighted average specificity of 0.87. 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 0.53.

The zero-shot method achieved a weighted average sensitivity of 0.58 and weighted average specificity of 0.91. The weighted average Cohen's kappa between human and GPT-4 decisions was 0.52.

Discussion

GPTscreenR and its chain-of-thought approach were able to achieve a sensitivity and specificity closely comparable to those reported by Guo et al. (Guo et al. 2023). 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. For reasons discussed further below, this is likely because of poorer performance relative to humans on more difficult cases rather than overall poorer performance.

Comparison to the zero-shot method

Interestingly, the replication of Guo et al.'s zero-shot prompt on the reviews used for validation in this study resulted in poorer weighted average sensitivity (0.58) compared to their report (0.76), although specificity (0.91) was the same. 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 perform better rather than worse 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 (1.00) and specificity (0.94) 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 two studies both validated the zero-shot method on a sample size comprising a different set of reviews, which may pose different levels of difficulty to GPT-based approaches. 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.

System: You are being used to help researchers perform a scoping review. You are not interacting directly with a user.

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.

Your task is to screen a single source against the inclusion criteria. You will be provided with the review objective and inclusion criteria, and then you will then be provided with the source title and abstract. You will be instructed to work step by step through the process of comparing the source against the inclusion criteria. Finally, you will instructed to make a recommendation on whether the source should be included.

The next message will be from the user, and will contain the scoping review objective and inclusion criteria.

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 no 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.

System: Let's work step by step. First, generate a numbered list of statements that summarise the inclusion criteria for the scoping review, including the Population, Concept, and Context. The statements should be clear, comprehensive and complete. Any source for which all the statements are true is a source that meets the inclusion criteria. As a template, here are some example statements (these are a generic set of examples that are not related to the current scoping review):

1. The source reports the results of a randomised control trial

2. The source reports the results of a study in which:

   2a. The participants were all male; AND

   2b. The participants were all aged between 18 and 74 inclusive

3. The source reports the results of a study conducted in the European Union.

Aspects of the inclusion criteria with multiple elements should be broken down into separate points where possible. For example, instead of:

1. The source reports on a study of men who live in the European Union.

You should instead say:

1. The source reports on a study of people who are:

   1a. Male; and

   1b. Living in the European Union.

GPT-4: 1. The source includes primary research conducted on elderly people 1a. 'Elderly' is defined as 65 years or older.

2. The elderly people in the source reside in a residential aged care facilities

   2a. 'Residential aged care facility' is a long-term residence for elderly people providing support for activities of daily living and nursing support.

   2b. The facility offers structured programmes and facilities for entertainment, diversion, and well-being.

   2c. The facility does not include residential settings that do not provide daily living support or nursing support, or that are not for long-term residence.

3. The residential aged care facilities mentioned in the source have implemented therapy alpaca programs as part of their care for residents.

4. The source examines the impact of therapy alpaca programmes

   4a. The impact examined includes effects on stress reduction, emotional well-being, mental health, overall life satisfaction, or similar outcomes for residents.

System: The next message will be from the user, and will contain the title and abstract of a source to be compared against the inclusion criteria.

User: 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 82 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: Let's continue to work step by step. Refer back to the set of statements you developed summarising the inclusion criteria. For each statement, decide whether or not the statement is true for the source described by the title and abstract. You must select from the following permitted responses: TRUE, FALSE, LIKELY TRUE, LIKELY FALSE, or NOT APPLICABLE. No other response is permitted. It is normal for the title and abstract to not have enough information to make a clear decision for every statement. There is a natural and normal amount of ambiguity in this process. For these situations, give your best guess, making use of your general knowledge, and deciding LIKELY TRUE or LIKELY FALSE. Responses like UNCLEAR or NOT ENOUGH INFORMATION are not permitted. After giving your response, give a one sentence explanation for your response.

GPT-4: 1. The source includes primary research conducted on elderly people. - LIKELY TRUE: While the age of the nursing home residents is not explicitly stated, it can be reasonably inferred that they are likely to be elderly. - Explanation: 'Nursing home residents' are generally senior individuals.

1a. 'Elderly' is defined as 65 years or older. - LIKELY TRUE: Although the abstract does not explicitly define an age for "elderly", it is common understanding that elderly often refers to individuals who are 65 years or older. - Explanation: Although not explicit, the elderly population in most research is usually referred to those aged 65 or more.

2. The elderly people in the source reside in a residential aged care facilities. - TRUE: The residents are shown to live in nursing homes, which qualifies as a type of residential aged care facility. - Explanation: Nursing homes are a type of residential aged care facility.

2a. 'Residential aged care facility' is a long-term residence for elderly people providing support for activities of daily living and nursing support. - LIKELY TRUE: Although not specifically mentioned in the abstract, it is understood that nursing homes facilitate activities of daily living and nursing support for the elderly. - Explanation: It is widely known that nursing homes provide support for activities of daily living and nursing care.

2b. The facility offers structured programmes and facilities for entertainment, diversion, and well-being. - TRUE: The abstract mentions pilot trials of therapy ungulates, suggesting the presence of structured programmes at the facility. - Explanation: The implementation of therapy ungulate trials indicates the existence of structured programs for the residents' well-being.

2c. The facility does not include residential settings that do not provide daily living support or nursing support, or that are not for long-term residence. - LIKELY TRUE: Although not explicitly stated, nursing homes generally provide both daily living support and nursing care, and are meant for long-term residence of the elderly. - Explanation: As part of their operations, nursing homes generally provide these forms of supports and cater to long-term residents.

3. The residential aged care facilities mentioned in the source have implemented therapy alpaca programmes as part of their care for residents. - TRUE: The abstract mentions the use of therapy ungulates, which includes alpacas, as part of their program. - Explanation: The abstract identifies therapy ungulates, including alpacas, being utilized in the nursing home.

4. The source examines the impact of therapy alpaca programmes. - FALSE: Although the abstract mentions the use of therapy ungulates, it specifically mentions the use of therapy camels in the trial, not alpacas. - Explanation: The abstract does not mention the specific effects of therapy alpaca programs, but rather, the therapy camel program.

4a. The impact examined includes effects on stress reduction, emotional well-being, mental health, overall life satisfaction, or similar outcomes for residents. - TRUE: The study indeed examines the impact on stress, depression, and anxiety in the nursing home residents. - Explanation: The abstract mentions that the study measures impact on stress reduction, depression, and anxiety, which can be seen as elements of mental health and emotional well-being.

System: Let's continue to work step by step. Consider your decisions on whether the title and abstract meet the conclusion criteria. Overall, is it likely true that the source meets the inclusion criteria? Reply with a single word, either INCLUDE or EXCLUDE, representing your recommendation on whether the source is likely to meet the inclusion criteria. You must reply with a single word only and it must be one of these two words; any other reply will cause the automatic parsing of your response to fail, which will be troublesome for the user.

GPT-4: EXCLUDE

Figure 2c: A conversation with GPT-4, instructing it to perform the same adversarial task presented in Fig. 2a, but using the chain-of-thought approach. This is not a continuation of the conversation presented in Fig. 2a and Fig. 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*

Inclusion criteria: *The inclusion criteria*

Exclusion criteria: *The exclusion criteria*

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. (Guo et al. 2023). Variable content, including user-provided data as well as GPT-4's response, is given in italics.

When compared directly against the chain-of-thought method for the same set of reviews, the zero-shot method achieved poorer sensitivity (0.58 vs 0.75) but essentially identical specificity (0.91). This suggests that the chain-of-thought approach is able to reduce type II error (false negatives) relative to the zero-shot approach, but is essentially equivalent with regards to type I error (false positives).

References

10 Arksey, Hilary, and Lisa O'Malley. 2005. "Scoping studies: towards a methodological framework." *International Journal of Social Research Methodology* 8 (1): 19–32. https://doi.org/10.1080/1364557032000119616.

Chai, Kevin E. K., Robin L. J. Lines, Daniel F. Gucciardi, and Leo Ng. 2021. "Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews." *Systematic Reviews* 10 (1): 93. https://doi.org/10.1186/s13643-021-01635-3.

Gates, Allison, Samantha Guitard, Jennifer Pillay, Sarah A. Elliott, Michele P. Dyson, Amanda S. Newton, and Lisa Hartling. 2019. "Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools." *Systematic Reviews* 8 (1): 278. https://doi.org/10.1186/s13643-019-1222-2.

Guo, Eddie, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Mike Paget, and Christopher Naugler. 2023. "Automated Paper Screening for Clinical Reviews Using Large Language Models." *arXiv.* https://doi.org/10.48550/arxiv.2305.00844.

Howard, Brian E., Jason Phillips, Kyle Miller, Arpit Tandon, Deepak Mav, Mihir R. Shah, Stephanie Holmgren, et al. 2016. "SWIFT-Review: a text-mining workbench for systematic review." *Systematic Reviews* 5 (1): 87. https://doi.org/10.1186/s13643-016-0263-z.

Institute, The Joanna Briggs. 2015. *The Joanna Briggs Institute Reviewers' Manual 2015: Methodology for JBI Scoping Reviews.* The Joanna Briggs Institute.

Khalil, Hanan, Daniel Ameen, and Armita Zarnegar. 2022. "Tools to support the automation of systematic reviews: a scoping review." *Journal of Clinical Epidemiology* 144: 22–42. https://doi.org/10.1016/j.jclinepi.2021.12.005.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. "Large Language Models Are Zero-Shot Reasoners." *arXiv.* https://doi.org/10.48550/arxiv.2205.11916.

Mortensen, Michael L., Gaelen P. Adam, Thomas A. Trikalinos, Tim Kraska, and Byron C. Wallace. 2017. "An exploration of crowdsourcing citation screening for systematic reviews." *Research Synthesis Methods* 8 (3): 366–86. https://doi.org/10.1002/jrsm.1252.

OpenAI. 2023. "Introducing ChatGPT." "https://openai.com/blog/chatgpt".

Peters, Micah D J, Casey Marnie, Andrea C Tricco, Danielle Pollock, Zachary Munn, Lyndsay Alexander, Patricia McInerney, Christina M Godfrey, and Hanan Khalil. 2020. "Updated Methodological Guidance for the Conduct

of Scoping Reviews." *JBI Evidence Synthesis* 18 (10): 2119–26. https://doi.org/10.11124/jbies-20-00167.

Peters, Micah D. J., Casey Marnie, Heather Colquhoun, Chantelle M. Garritty, Susanne Hempel, Tanya Horsley, Etienne V. Langlois, et al. 2021. "Scoping reviews: reinforcing and advancing the methodology and application." *Systematic Reviews* 10 (1): 263. https://doi.org/10.1186/s13643-021-01821-3.

Pham, Mai T., Andrijana Rajić, Judy D. Greig, Jan M. Sargeant, Andrew Papadopoulos, and Scott A. McEwen. 2014. "A scoping review of scoping reviews: advancing the approach and enhancing the consistency." *Research Synthesis Methods* 5 (4): 371–85. https://doi.org/10.1002/jrsm.1123.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. "Improving Language Understanding by Generative Pre-Training." "https://openai.com/research/language-unsupervised".

Shemilt, Ian, Antonia Simon, Gareth J. Hollands, Theresa M. Marteau, David Ogilvie, Alison O'Mara-Eves, Michael P. Kelly, and James Thomas. 2014. "Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews." *Research Synthesis Methods* 5 (1): 31–49. https://doi.org/10.1002/jrsm.1093.

Stansfield, Claire, James Thomas, and Josephine Kavanagh. 2013. "'Clustering' documents automatically to support scoping reviews of research: a case study." *Research Synthesis Methods* 4 (3): 230–41. https://doi.org/10.1002/jrsm.1082.

Sutton, Anthea, Mark Clowes, Louise Preston, and Andrew Booth. 2019. "Meeting the review family: exploring review types and associated information retrieval requirements." *Health Information & Libraries Journal* 36 (3): 202–22. https://doi.org/10.1111/hir.12276.

Syriani, Eugene, Istvan David, and Gauransh Kumar. 2023. "Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews." *arXiv.*

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. "Chain of Thought Prompting Elicits Reasoning in Large Language Models." *arXiv.* https://doi.org/10.48550/arxiv.2201.11903.

William Revelle. 2023. *Psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." *arXiv.* https://arxiv.org/abs/2305.10601.