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

David Wilkins
Discipline of General Practice
10 October 2023

Outline

- What is a Large Language Model?
- Introducing the GPTscreenR package
- The chain-of-thought approach
- GPTscreenR's performance

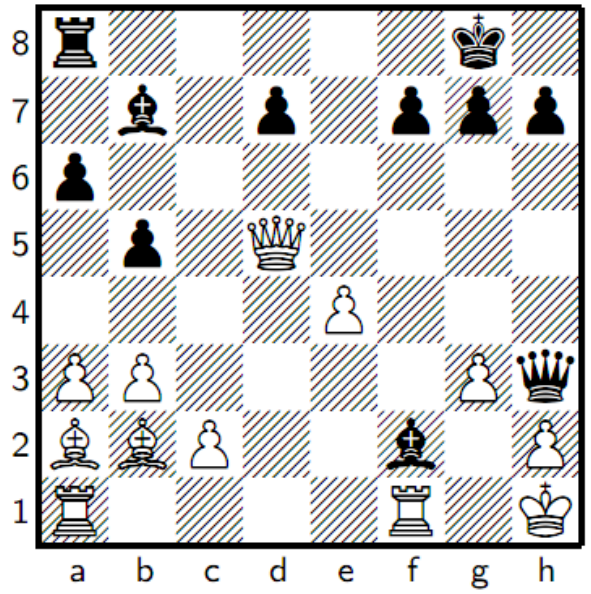
What is a large language model?

- Neural networks can be trained to produce certain outputs in response to certain inputs

Wind speed = 12 km/hour
Wind direction = NNE
Humidity = 65%
Temperature = 23 °C

“2% chance of rain in the next hour”

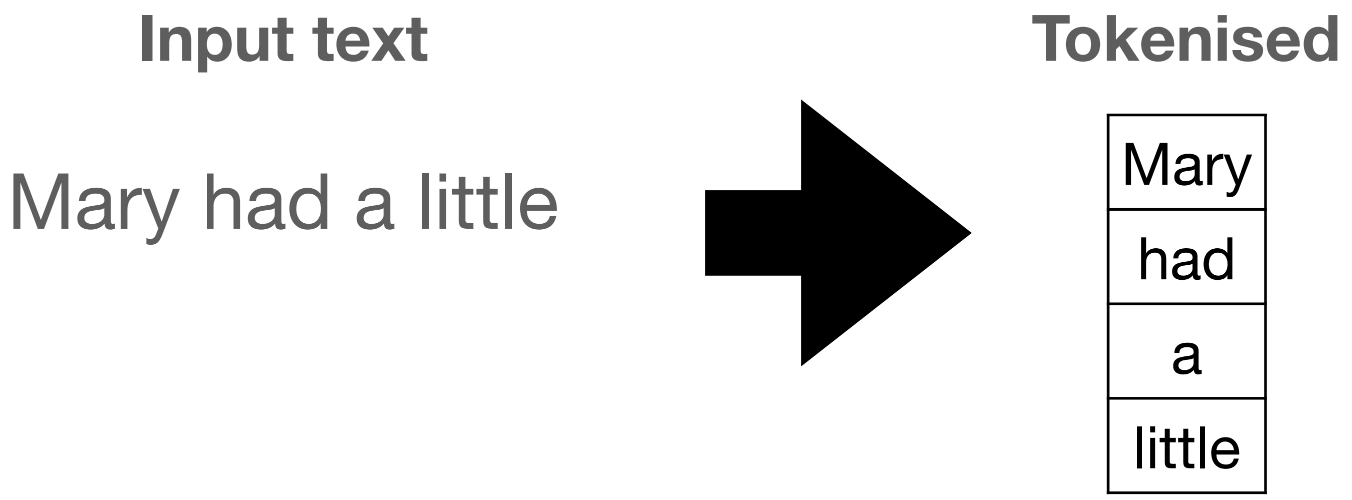
“a cat”



“the optimal move is…”

What is a large language model?

- Large Language Models (LLMs) take text as an input
- The text is broken down into small pieces called tokens
- A token is often an entire word, but it can also be a part of a word, or another part of text like punctuation



Input text

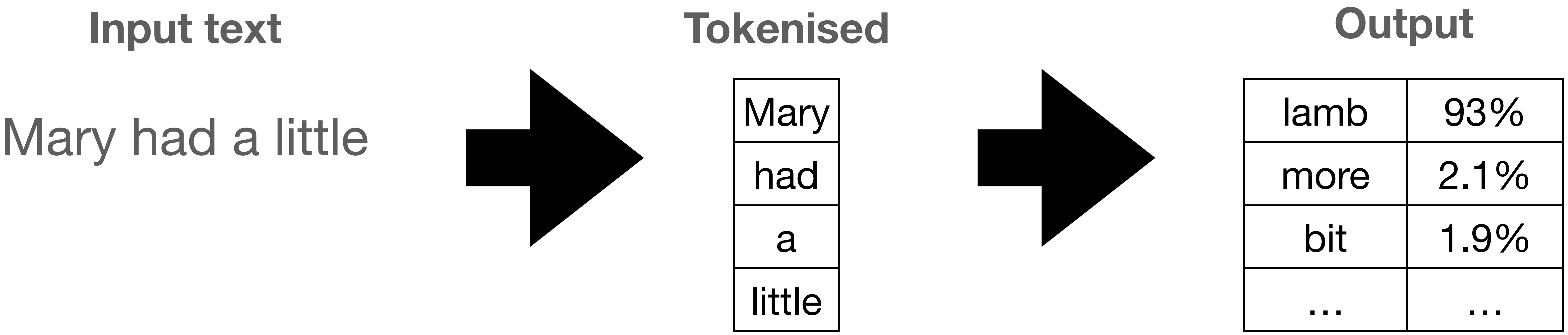
Mary had a little

Tokenised

Mary
had
a
little

What is a large language model?

- LLMs are trained to output the predicted next token after a series of input tokens
- They learn how to do this by looking at huge amounts of human-written text
- They output a list of possible next tokens, paired with the probability that each token will come next



Input text

Mary had a little

Tokenised

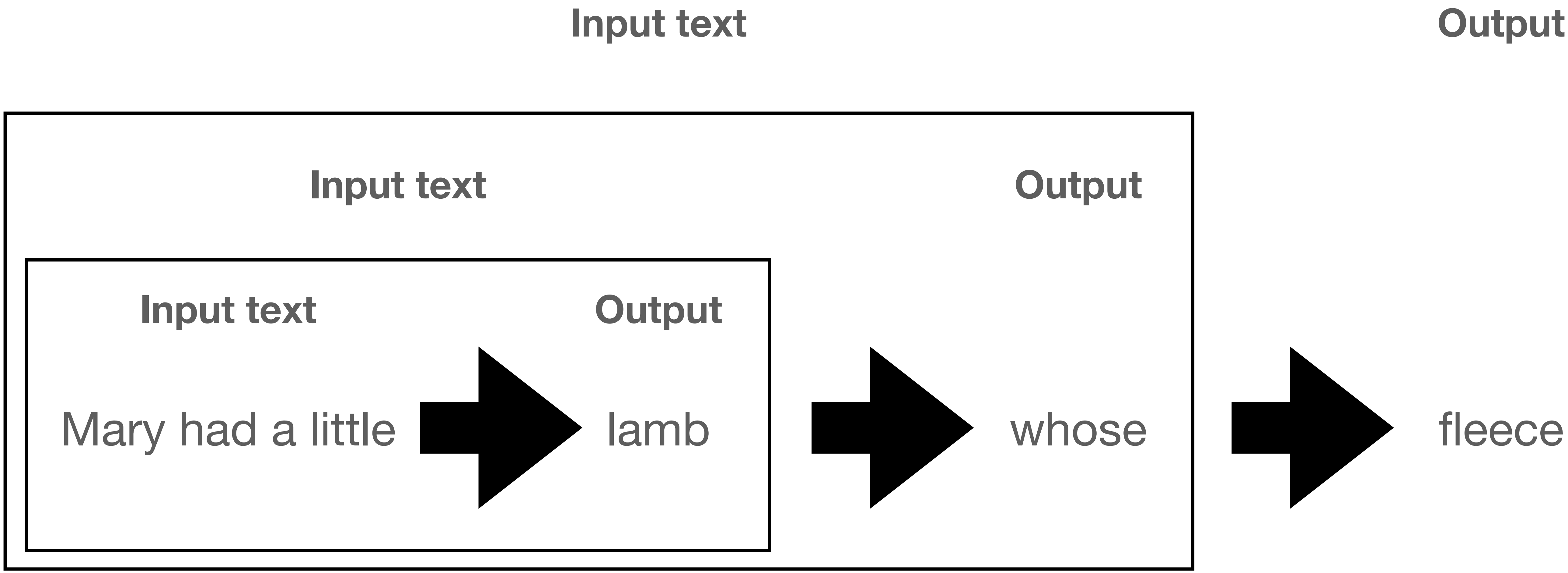
Mary
had
a
little

Output

lamb	93%
more	2.1%
bit	1.9%
…	…

What is a large language model?

- To generate a long string of output text, this process is iterated and repeated, extending the text one token at a time



Input text **Output**

Input text **Output**

Input text **Output**

Mary had a little → lamb → whose → fleece

What is a large language model?

- To have a ‘conversation’ with an LLM (e.g. ChatGPT), the input text is a transcript of the conversation so far, which the LLM extends one token at a time

Input text

ChatGPT: Hello, how can I help you today?

User: What is the capital of France?

ChatGPT: The capital of France is

Output

Paris	97%
’nt	1.2%
Rome	0.09%
…	…

What is a large language model?

- Neural networks have been around for decades, but LLMs have become much more capable recently for several reasons:
  - More computing power
  - Bigger training datasets
  - The transformer architecture
- ChatGPT is based on the Generative Pre-trained Transformer (GPT) series of LLMs, first released in 2018
- The latest (and most capable) GPT model is GPT-4 (March 2023)

Introducing GPTscreenR

- An R package to help reviewers screen sources for a scoping review
- Uses GPT-4 (via the OpenAI API) to look at titles and abstracts and make a recommendation to include or exclude in the review
- Previous work has used LLMs (including earlier versions of GPT) to screen sources for systematic reviews
    - Guo, E. et al. Automated Paper Screening for Clinical Reviews Using Large Language Models. arXiv (2023) doi:10.48550/arxiv.2305.00844.
    - Syriani, E., David, I. & Kumar, G. Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. arXiv (2023) doi:10.48550/arxiv.2307.06464.

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

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.

The chain-of-thought method

- Wei, J. et al. Chain of Thought Prompting Elicits Reasoning in Large Language Models. arXiv (2022) doi:10.48550/arxiv.2201.11903.
- Decompose a complex problem into smaller steps, and prompt the LLM to work through each step separately
- Significantly increases accuracy on complex problems

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 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.

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.

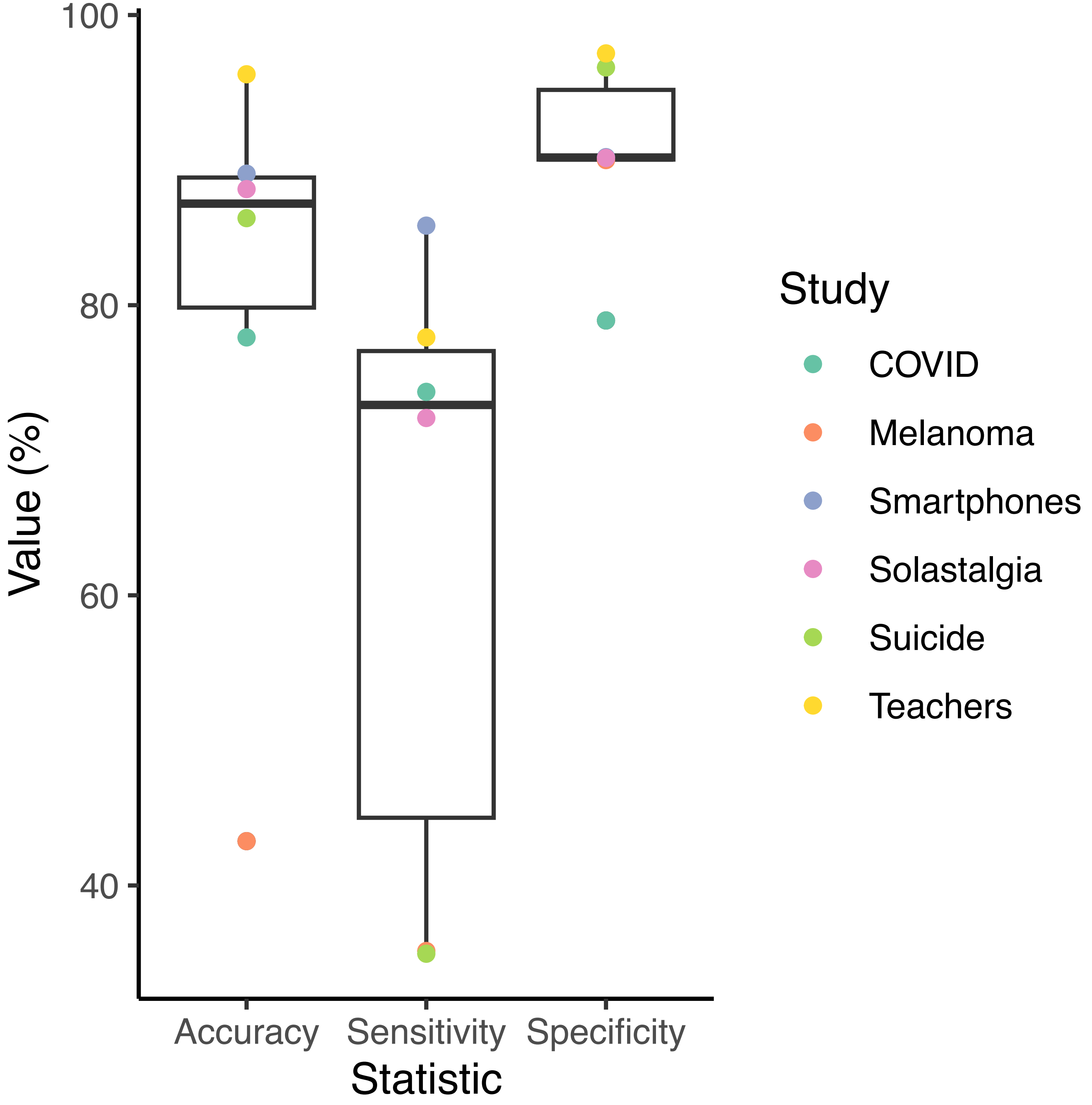
System: FINALLY, consider your decisions on whether the source meets the conclusion criteria. Respond with a single word, either INCLUDE or EXCLUDE, representing your recommendation on whether the source meets the inclusion criteria. Do not write anything other than INCLUDE or EXCLUDE.

GPT-4: EXCLUDE

GPTscreenR’s performance

- Validated against five scoping reviews with the consensus human decision as the gold standard
- 1,147 titles and abstracts screened with GPTscreenR (chain-of-thought) and with zero-shot (Guo *et. al.*)
    - Accuracy: GPTscreenR = 84%, zero-shot = 83%
    - Weighted average sensitivity: GPTscreenR = 71%, zero-shot = 72%
    - Weighted average specificity: GPTscreenR 89%, zero-shot = 87%
    - Weighted Cohen’s kappa (agreement between LLM and humans): GPTscreenR = 0.52, zero-shot = 0.52
        - For comparison, weighted Cohen’s kappa between humans = 0.67

GPTscreenR’s performance



Conclusions

- GPTscreenR works about as well as zero-shot method
- LLM based screening not as reliable as human screening, but still very promising as method to reduce workload for humans
- Needs more real-world validation