McMaster University

Family Medicine

David Braley Primary Care Research Collaborative

Using Generative AI in Systematic Reviews

Watch a recording of the session

Istvan David

October 22nd, 2024

Presenter's Twitter Handle
@McMasterFamMed

Michael G. DeGroote SCHOOL OF MEDICINE

Using Generative AI in Systematic Reviews

Istvan David
McMaster University, Canada
istvandavid.com

October 22, 2024
@Department of Family Medicine, McMaster University

McMaster University

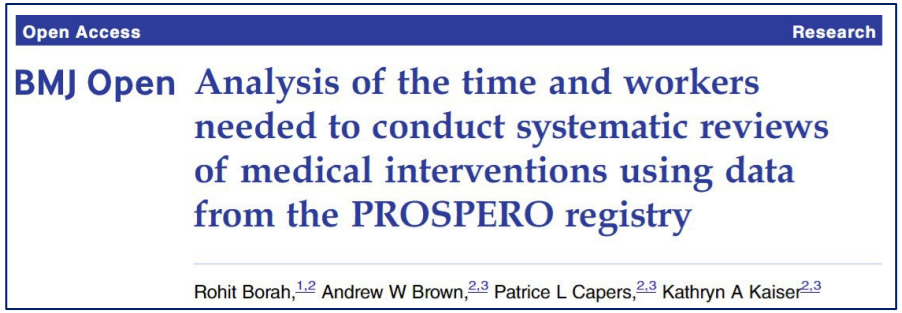
Sustainable Systems and Methods

Systematic reviews

- Goals
  - Synthesize and organize knowledge from primary studies
  - Document the SOTA and provide a foundation for scholarly research
  - SE/CS: steady increase in the number of published systematic reviews

- Most problematic part: selection and screening
  - Time-consuming
  - Error-prone
  - A significant barrier

We need automation!


Open Access Research

BMJ Open

Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry

Rohit Borah,[1,2] Andrew W Brown,[2,3] Patrice L Capers,[2,3] Kathryn A Kaiser[2,3]

Information and Software Technology 91 (2017) 72–81

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

ELSEVIER

Vision for SLR tooling infrastructure: Prioritizing value-added requirements

Ahmed Al-Zubidy[a], Jeffrey C. Carver[a,*], David P. Hale[a], Edgar E. Hassler[b]

[a] The University of Alabama, Tuscaloosa, Alabama, USA
[b] Appalachian State University, Boone, North Carolina, USA

McMaster University

Screening automation for systematic reviews

- Traditional AI/ML methods: typically supervised
  - Training input: labeled data (e.g., previously included/excluded articles)
  - Task input: unlabeled data (e.g., articles to be included/excluded)
  - Output: labeled data (e.g., inclusion/exclusion decisions)
- Active learning
  - The human is queried whenever a label is required
  - Often combined with ranking methods
- Support vector machines, decision trees, Bayesian networks…

- Why these methods fail?
  - Not enough data or data is of low quality
  - Need for re-training for each SR

Generative AI

- Class of AI that uses generative models to create text, images, etc
  - Traditional AI: pattern recognition, performing a specific task
  - GenAI: create *new* content/information
- Enabled by *foundation models*
  - Deep learning model trained on vast datasets + adaptation/fine-tuning for downstream tasks ➔ Applicable across a wide range of use cases
  - Examples: ChatGPT (LLM), BERT (LLM), DALL-E (image), MusicGen (music)
- Personalized experiences, content, and product recommendations

- No need for training, only fine-tuning by providing examples

- Input: prompt
  - Typically: context + some examples + task
  - Examples: “shots” ➔ zero-shot learning, few-shot learning
- Output: generated artifact (text, image, etc)

Generating content like a human would

Prompt:

A seascape in the expressionistic style associated with Van Gogh, complete with swirling stokes and vivid colors. The scene is dominated by the vast, dynamic ocean with waves thrashing about, catching the light from above. There's an emphasis on not just the visual depiction of the sea but also the emotions it evokes. Use colors such as ultramarine for the deep ocean, gradually transitioning to lighter hues of blues and greens where the waves crash and foam. Create the sky with Van Gogh's typical whirls in bright shades of yellows, oranges, and reds.

Generating content like a human would

Prompt:

```
I am screening papers for a systematic literature review.
The topic of the systematic review is reinforcement learning for software engineering.
The study should focus exclusively on this topic.

I give 2 examples with title and abstract that should be included.
Example 1:
-Title: A DQN-based agent for automatic software refactoring
-Abstract: Context: Nowadays, technical debt has become a very important issue in software project […]
Example 2:
-Title: A Reinforcement Learning-Based Framework for the Generation and Evolution of Adaptation Rules
-Abstract: One of the challenges in self-adaptive systems concerns how to make adaptation […]

Exclude the article if any of the following 2 criteria are true.
1: Article does not define or use a reinforcement learning method.
2: Software engineering is not the problem reinforcement learning is used for.

Decide if the article should be included or excluded from the systematic review.
I give the title and abstract of the article as input.
Only answer INCLUDE or EXCLUDE.
Be lenient. I prefer including papers by mistake rather than excluding them by mistake.

-Title: PARMOREL: a framework for customizable model repair
-Abstract: In model-driven software engineering, models are used in all phases of the development process[…]
```

Generating ~~content~~ like a human would decisions

Prompt:

```
I am screening papers for a systematic literature review.
The topic of the systematic review is reinforcement learning for software engineering.
The study should focus exclusively on this topic.

I give 2 examples with title and abstract that should be included.
Example 1:
-Title: A DQN-based agent for automatic software refactoring
-Abstract: Context: Nowadays, technical debt has become a very important issue in software project […]
Example 2:
-Title: A Reinforcement Learning-Based Framework for the Generation and Evolution of Adaptation Rules
-Abstract: One of the challenges in self-adaptive systems concerns how to make adaptation […]

Exclude the article if any of the following 2 criteria are true.
1: Article does not define or use a reinforcement learning method.
2: Software engineering is not the problem reinforcement learning is used for.

Decide if the article should be included or excluded from the systematic review.
I give the title and abstract of the article as input.
Only answer INCLUDE or EXCLUDE.
Be lenient. I prefer including papers by mistake rather than excluding them by mistake.

-Title: PARMOREL: a framework for customizable model repair
-Abstract: In model-driven software engineering, models are used in all phases of the development process[…]
```

Journal of Computer Languages 80 (2024) 101287

Contents lists available at ScienceDirect

Journal of Computer Languages

journal homepage: www.elsevier.com/locate/cola

Screening articles for systematic reviews with ChatGPT

Eugene Syriani [a,*], Istvan David [b], Gauransh Kumar [a]

[a] DIRO, Université de Montréal, Canada
[b] McMaster University, Canada

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.10257742

Keywords:
Generative AI
GPT
Empirical research
Large language model
Literature review
Mapping study
Screening

ABSTRACT

Systematic reviews (SRs) provide valuable evidence for guiding new research directions. However, the manual effort involved in selecting articles for inclusion in an SR is error-prone and time-consuming. While screening articles has traditionally been considered challenging to automate, the advent of large language models offers new possibilities. In this paper, we discuss the effect of using ChatGPT on the SR process. In particular, we investigate the effectiveness of different prompt strategies for automating the article screening process using five real SR datasets. Our results show that ChatGPT can reach up to 82% accuracy. The best performing prompts specify exclusion criteria and avoid negative shots. However, prompts should be adapted to different corpus characteristics.

What if we used GenAI as an additional reviewer in the screening phase?

80+% accuracy

https://doi.org/10.1016/j.cola.2024.101287
https://arxiv.org/abs/2307.06464

arXiv > cs > arXiv:2307.06464

Computer Science > Software Engineering

[Submitted on 12 Jul 2023]

Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews

Eugene Syriani, Istvan David, Gauransh Kumar

By organizing knowledge within a research field, Systematic Reviews (SR) provide valuable leads to steer research. Evidence suggests that SRs have become first-class artifacts in software engineering. However, the tedious manual effort associated with the screening phase of SRs renders these studies a costly and error-prone endeavor. While screening has traditionally been considered not amenable to automation, the advent of generative AI-driven chatbots, backed with large language models is set to disrupt the field. In this report, we propose an approach to leverage these novel technological developments for automating the screening of SRs. We assess the consistency, classification performance, and generalizability of ChatGPT in screening articles for SRs and compare these figures with those of traditional classifiers used in SR automation. Our results indicate that ChatGPT is a viable option to automate the SR processes, but requires careful considerations from developers when integrating ChatGPT into their SR tools.

Prompt template

How many examples to provide?

Example: a previously classified title+abstract

- No examples: *zero-shot learning*
- A few examples: *few-shot learning*
- Positive vs negative shots

Inclusion/exclusion criteria as defined in the SR protocol

The specific title+abstract that has to be classified

```
1  <Prompt> ::= <Context> <Examples>? <SelectionCriteria>? <Instructions> <Task>
2  <Context> ::= 'I am screening papers for a systematic literature review. The topic of the systematic review is {TOPIC}. The study
       should focus exclusively on this topic.'
3  <Examples> ::= <ExampleHeader> <Example>+ (<ExampleHeader> <Example>+)?
4  <ExampleHeader> ::= 'I give ({N+}|{N-}) examples with {FEATURE} that should be (included|excluded).'
5  <Example> ::= ('Example '[1-9] ':' <Task>)+
6  <SelectionCriteria> ::= <CriteriaHeader> <Criterion>+ (<CriteriaHeader> <Criterion>+)?
7  <CriteriaHeader> ::= '(Include|Exclude) the article if (all|any) of the following ({Ni}|{Nx}) criteria (are|is) true.'
8  <Criterion> ::= [1-9] ': {CRITERION}'
9  <Instructions> ::= 'Decide if the article should be included or excluded from the systematic review. I give the {FEATURE}+ of the
       article as input. Only answer {INCLUDE_WORD} or {EXCLUDE_WORD}. Be lenient. I prefer including papers by mistake rather
       than excluding them by mistake.'
10 <Task> ::= '-{FEATURE}: {INPUT}'
```

Prompt instance (example)

Context

```
I am screening papers for a systematic literature review.
The topic of the systematic review is reinforcement learning for software engineering.
The study should focus exclusively on this topic.
```

Shots (a few)

```
I give 2 examples with title and abstract that should be included.
Example 1:
-Title: A DQN-based agent for automatic software refactoring
-Abstract: Context: Nowadays, technical debt has become a very important issue in software project […]
Example 2:
-Title: A Reinforcement Learning-Based Framework for the Generation and Evolution of Adaptation Rules
-Abstract: One of the challenges in self-adaptive systems concerns how to make adaptation […]
```

Exclusion criteria

```
Exclude the article if any of the following 2 criteria are true.
1: Article does not define or use a reinforcement learning method.
2: Software engineering is not the problem reinforcement learning is used for.
```

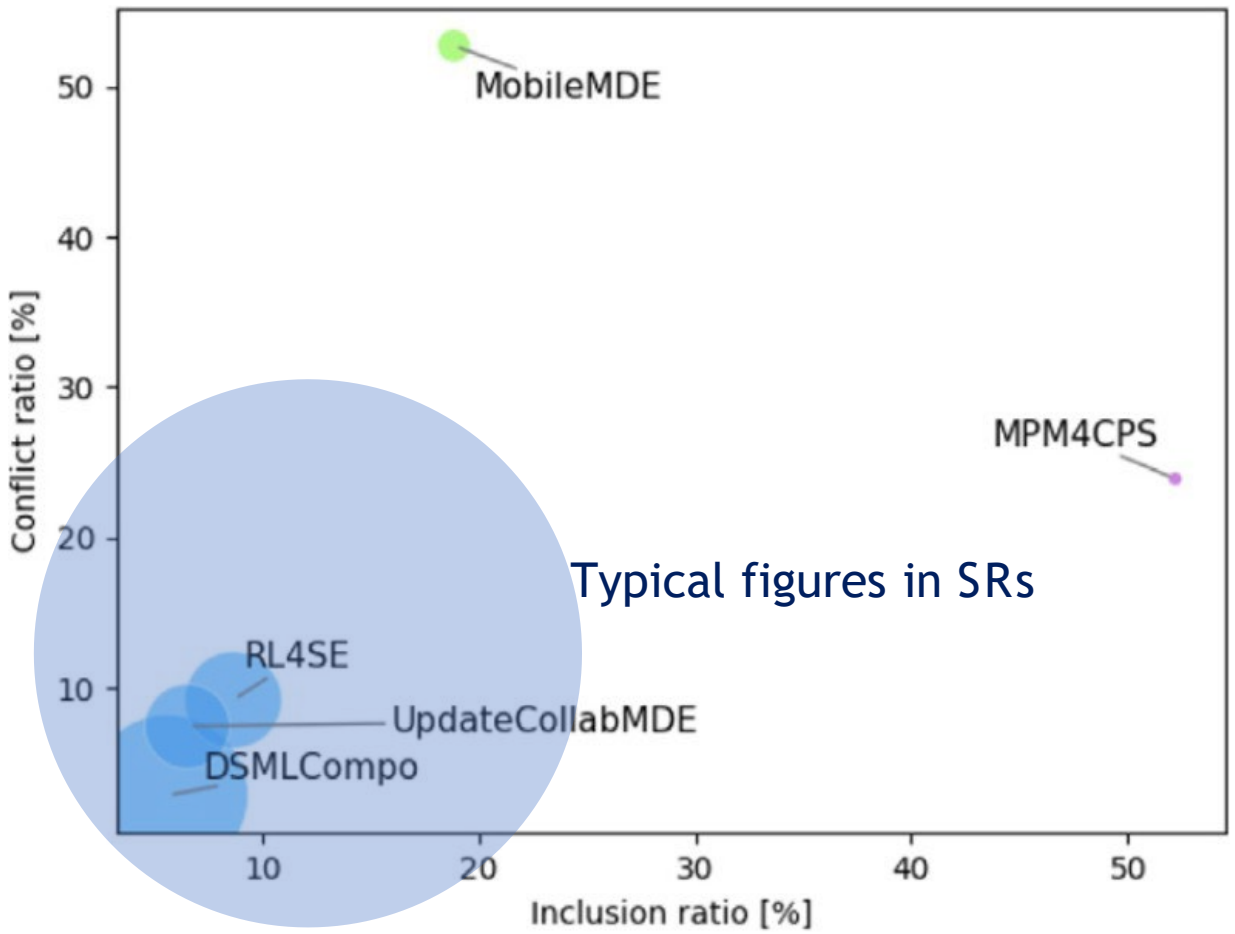
Instructions

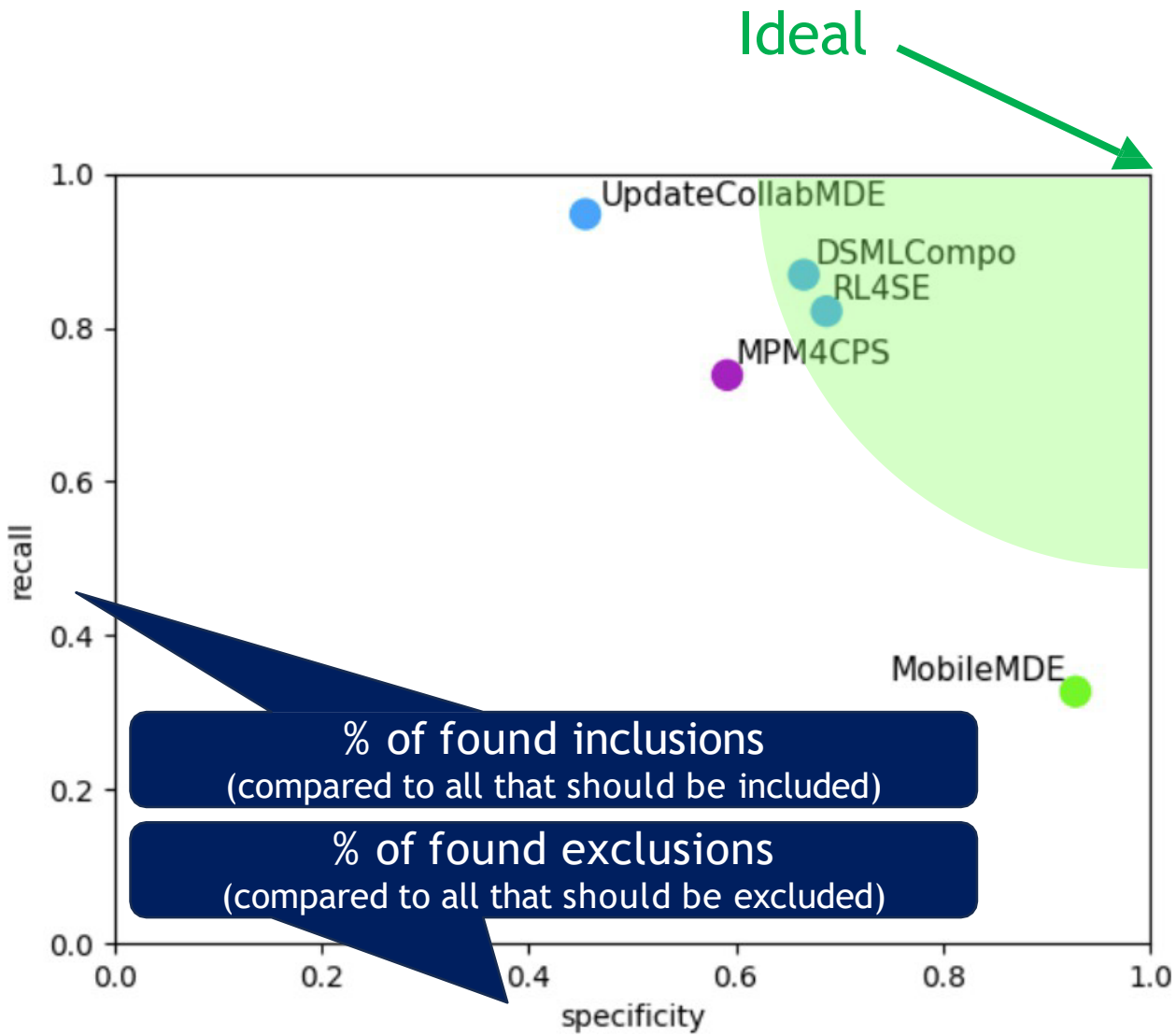
```
Decide if the article should be included or excluded from the systematic review.
I give the title and abstract of the article as input.
Only answer INCLUDE or EXCLUDE.
Be lenient. I prefer including papers by mistake rather than excluding them by mistake.
```

Task

```
-Title: PARMOREL: a framework for customizable model repair
-Abstract: In model-driven software engineering, models are used in all phases of the development process[…]
```

Experimental data and results





McMaster University

Choosing your metrics

Be lenient. I prefer including papers by mistake rather than excluding them by mistake.

FP ↑
N ↓

$$Prec = \frac{TP}{TP + FP}$$

% of **correct** inclusions
(compared to all inclusions)

$$Rec = \frac{TP}{TP + FN}$$

% of **found** inclusions
(compared to all that should have been included)

$$NPV = \frac{TN}{TN + FN}$$

% of **correct** exclusions
(compared to all exclusions)

$$Spec = \frac{TN}{TN + FP}$$

% of **found** exclusions
(compared to all that should have been excluded)

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \qquad F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

F_2: weights recall higher than precision
$F_{0.5}$: weights precision higher than recall

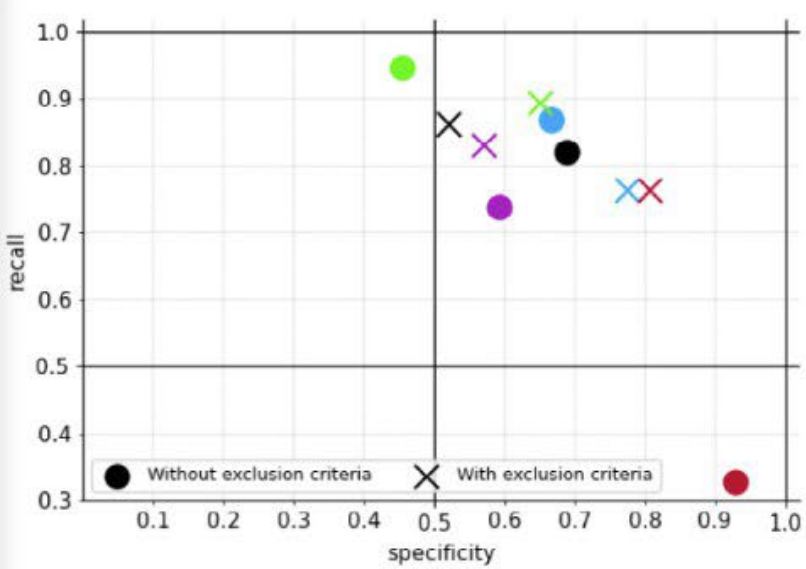
$$bAcc = \frac{Rec + Spec}{2}$$

Better for imbalanced data

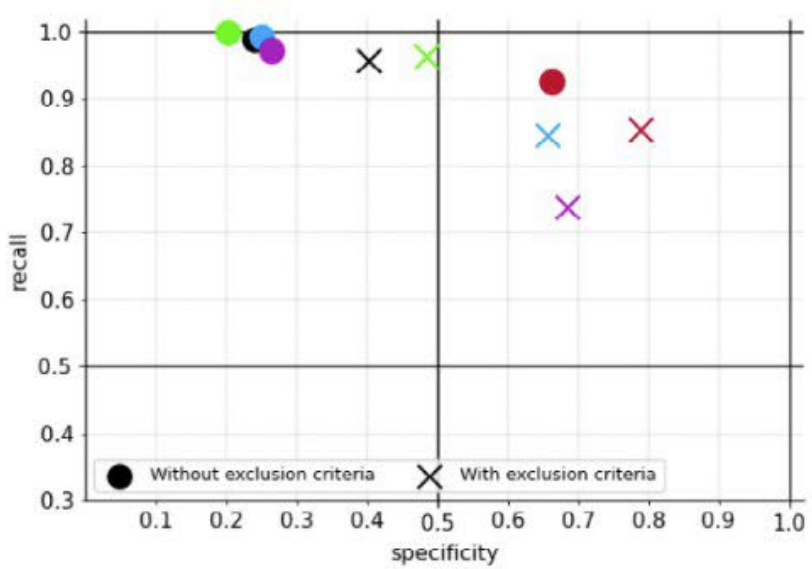
$$MCC = 0.5 + \frac{TP \times TN - FP \times FN}{2 \times \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Better for imbalanced data *and* binary classifiers

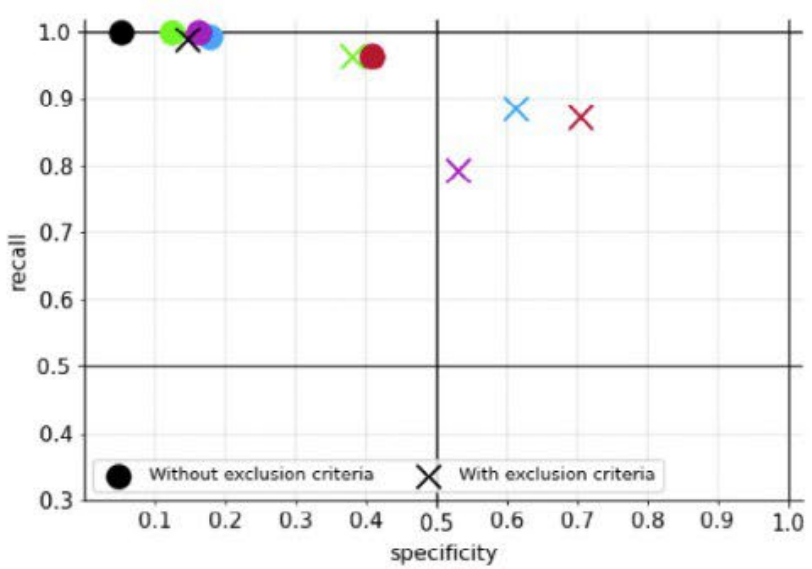
Which prompting strategy to use?



(a) Zero-shot
(*Simple* and *SimpleX*)



(b) Few-shot
(*Positive* and *PositiveX*)

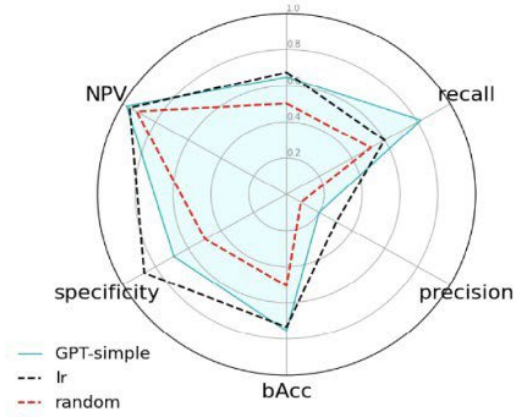


(c) Few-shot with negative
(*Balanced* and *BalancedX*)

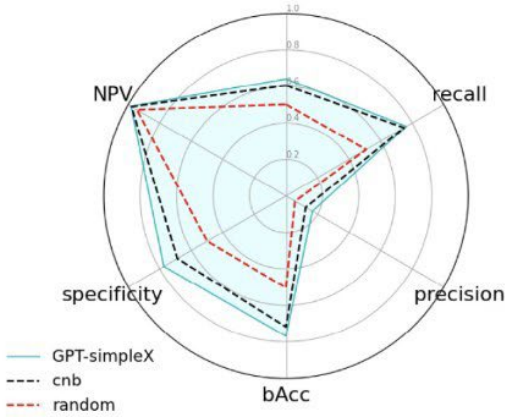
Answer to RQ1

Zero-shot prompts tend to perform better than few-shot prompts. Listing exclusion criteria tends to increase specificity. Using negative shots deteriorates the performance.

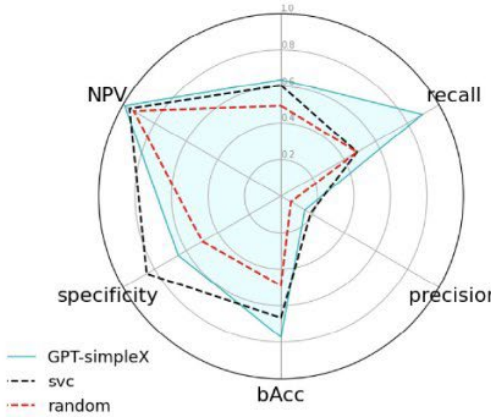
GPT vs traditional AI methods



(a) RL4SE



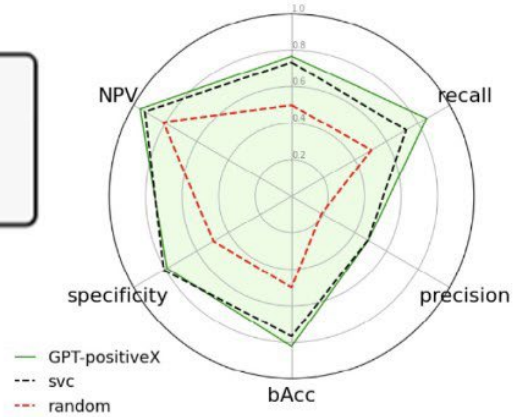
(b) DSMLCompo



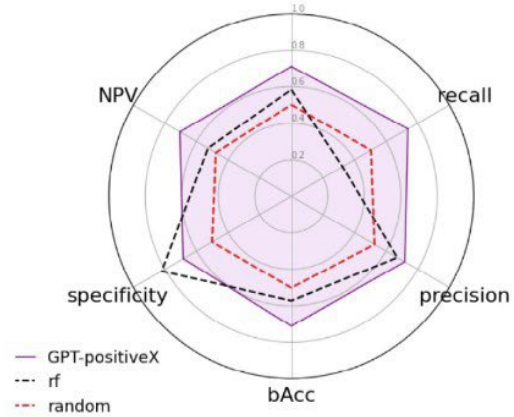
(c) UpdateCollabMDE

Answer to RQ2

ChatGPT classifies articles more accurately than baseline classifiers. Its prompts reach higher recall but lower specificity.

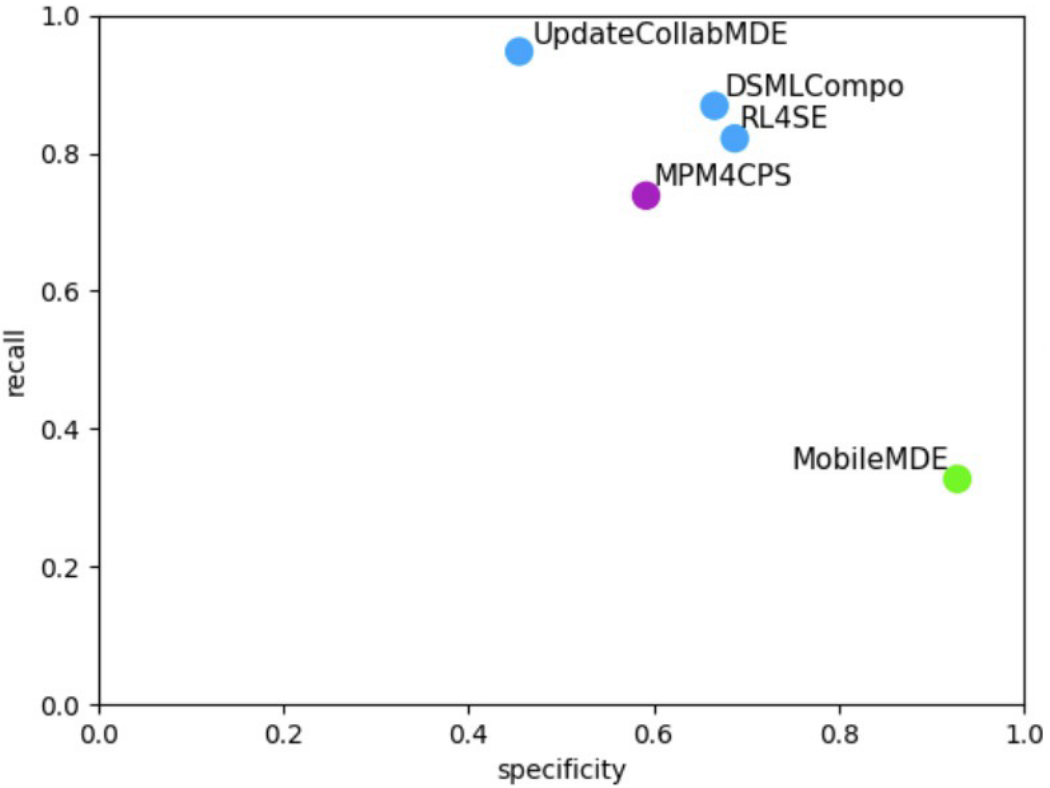


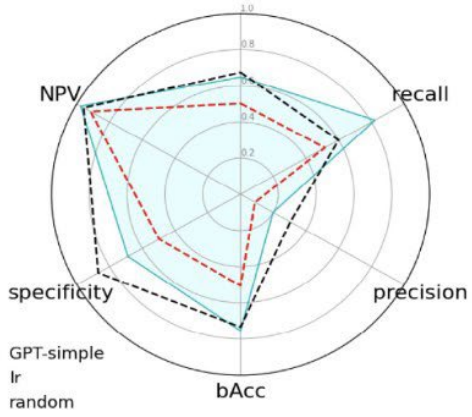
(d) MobileMDE



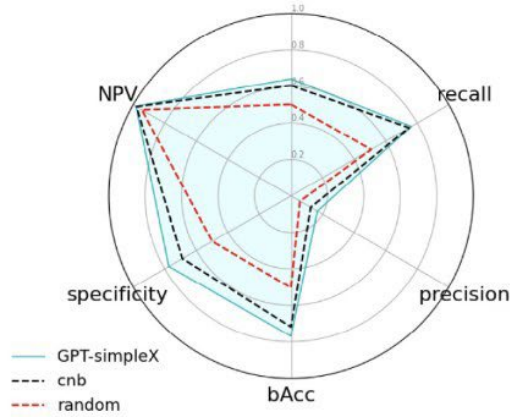
(e) MPM4CPS

Corpus characteristics

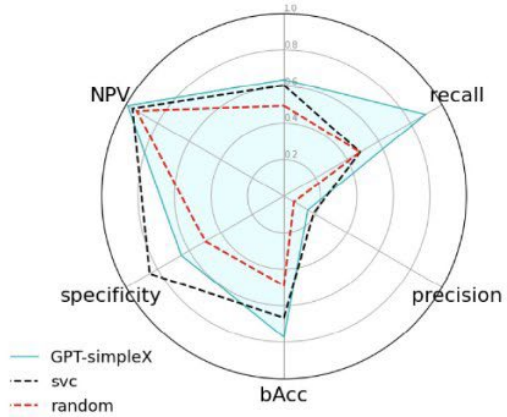




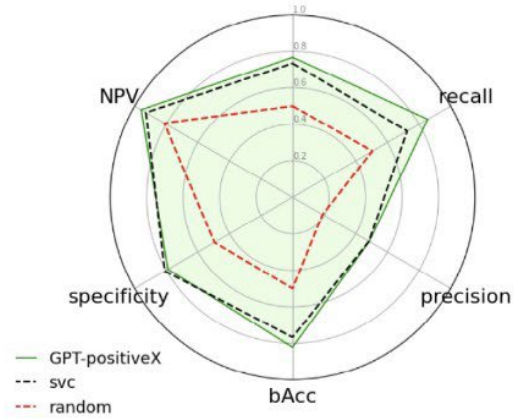
(a) RL4SE



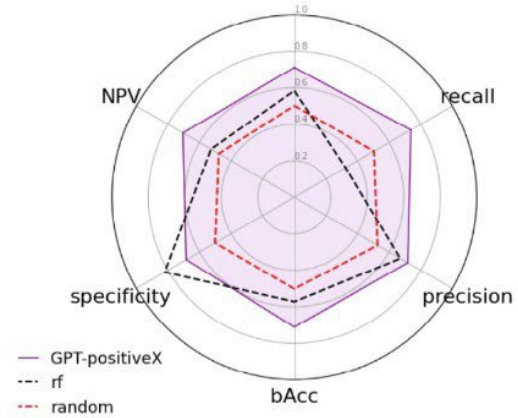
(b) DSMLCompo



(c) UpdateCollabMDE



(d) MobileMDE



(e) MPM4CPS

Answer to RQ3

The characteristics of the corpus affect the performance profile of ChatGPT. Different corpus characteristics might favor different prompt strategies.

Other uses of GenAI in SRs

- Search query generation
- Data extraction
- Risk/bias assessment

- Review partner (copilot)

Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?

Shuai Wang
The University of Queensland
Brisbane, Australia
shuai.wang5@uq.net.au

Harrisen Scells
Leipzig University
Leipzig, Germany
harry.scells@uni-leipzig.de

Bevan Koopman
CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Guido Zuccon
The University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

CAN LARGE LANGUAGE MODELS REPLACE HUMANS IN THE SYSTEMATIC REVIEW PROCESS? EVALUATING GPT-4'S EFFICACY IN SCREENING AND EXTRACTING DATA FROM PEER-REVIEWED AND GREY LITERATURE IN MULTIPLE LANGUAGES

Review

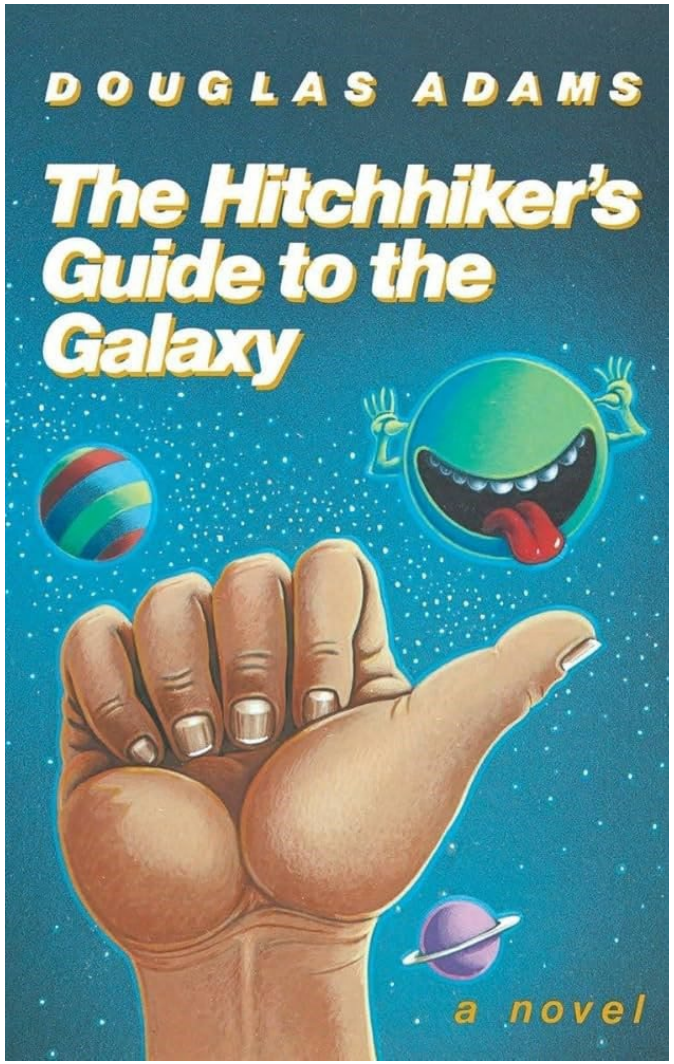
Enhancing Clinical Reasoning with Virtual Patients: A Hybrid Systematic Review Combining Human Reviewers and ChatGPT

Daniel García Torres [1], María Asunción Vicente Ripoll [2,*], César Fernández Peris [2] and José Joaquín Mira Solves [1,3,4]

Research methods and reporting

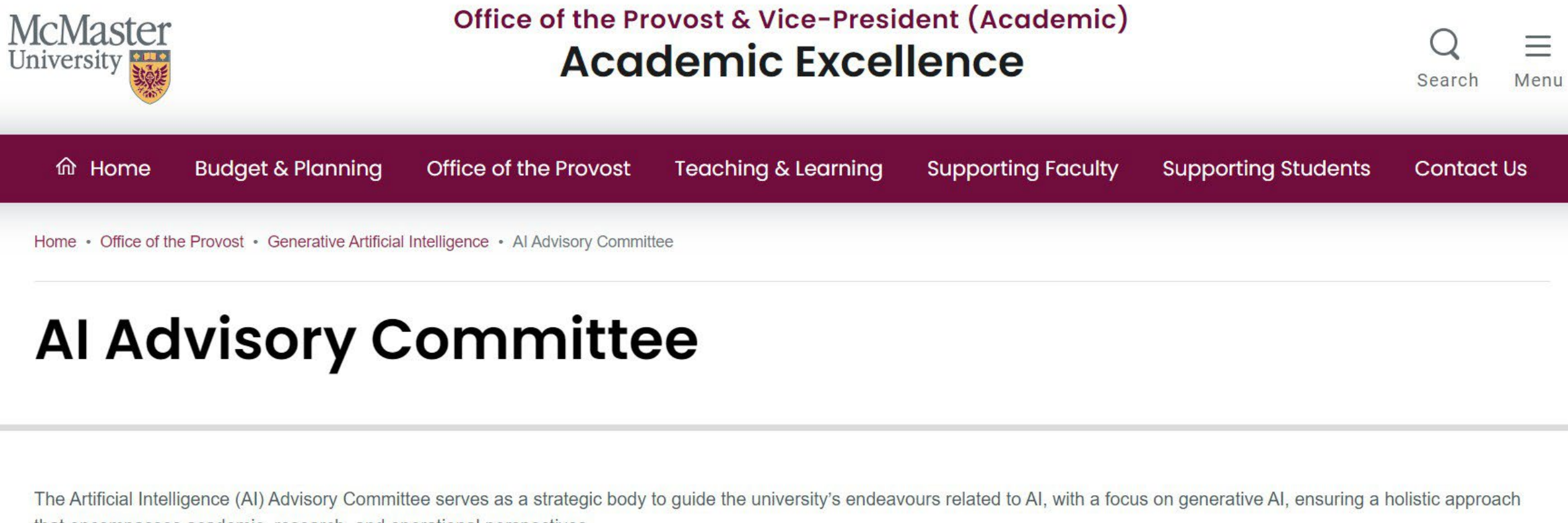
Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment

AI hallucinations



"the answer to the ultimate question of life, the universe, and everything

Perspectives on the Use of AI in Research



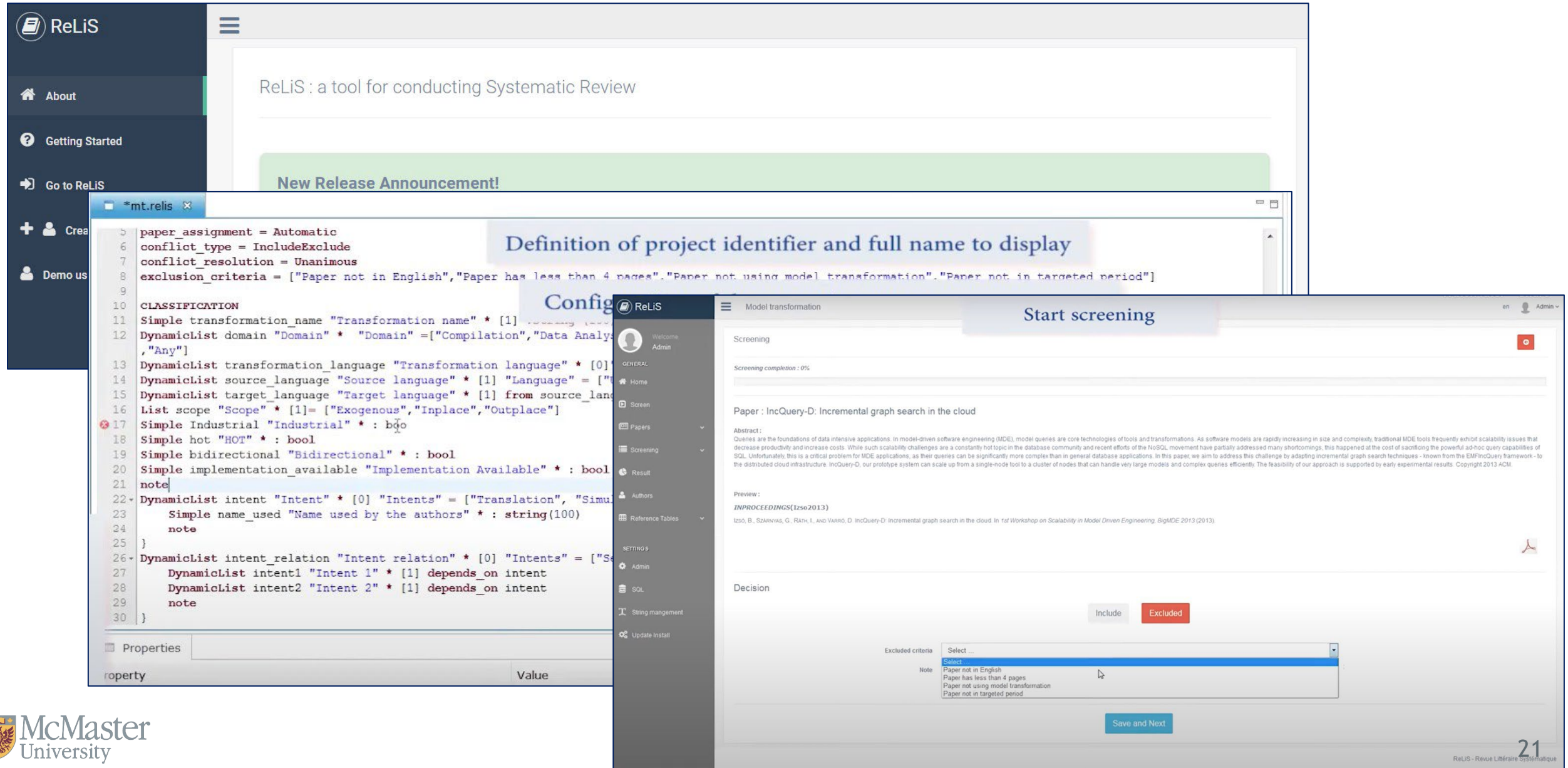
Bottom line

- Are we there yet?
  - + GenAI outperforms traditional classifiers and renders previous work obsolete
  - - GenAI cannot be trusted just yet and the human in the loop is still required

- Expected developments in 1-5 years
  - Rapid SRs by GenAI-streamlined screening [Usual team setup]
  - (Gen)AI as a copilot [Fewer humans needed]
  - Solo SRs [One human needed]
    - Mean number of authors in SE: 2.67 (over half of the articles having one or two authors)
    - ➔Gathering a team for an SR is a challenge for the majority of SE researchers
  - Conversational evidence synthesis
  - Eventually: GenAI to become a key element of SRs
    - The role of human researchers: from labor to oversight/validation/teaching

ReLiS *(Revue Littéraire Systématique)*

https://relis.iro.umontreal.ca



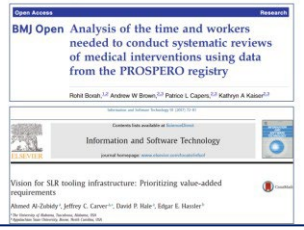
Using generative AI in Systematic Reviews

Istvan David

Systematic reviews

- Goals
  - Synthesize and organize knowledge from primary studies
  - Document the SOTA and provide a foundation for scholarly research
  - SE/CS: steady increase in the number of published systematic reviews
- Most problematic part: selection and screening
  - Time-consuming
  - Error-prone
  - A significant barrier

We need automation!



Generating content like a human would

Prompt:

```
A seascape in the expressionistic style
associated with Van Gogh, complete with
swirling stokes and vivid colors. The scene is
dominated by the vast, dynamic ocean with waves
thrashing about, catching the light from above.
There's an emphasis on not just the visual
depiction of the sea but also the emotions it
evokes. Use colors such as ultramarine for the
deep ocean, gradually transitioning to lighter
hues of blues and greens where the waves crash
and foam. Create the sky with Van Gogh's
typical whirls in bright shades of yellows,
oranges, and reds.
```

Prompt instance (example)

Context

```
I am screening papers for a systematic literature review.
The topic of the systematic review is reinforcement learning for software engineering.
The study should focus exclusively on this topic.
```

Shots (a few)

```
I give 2 examples with title and abstract that should be included.
Example 1:
-Title: A DQN-based agent for automatic software refactoring
-Abstract: Context: Nowadays, technical debt has become a very important issue in software project […]
Example 2:
-Title: A Reinforcement Learning-Based Framework for the Generation and Evolution of Adaptation Rules
-Abstract: One of the challenges in self-adaptive systems concerns how to make adaptation […]
```

Exclusion criteria

```
Exclude the article if any of the following 2 criteria are true.
1: Article does not define or use a reinforcement learning method.
2: Software engineering is not the problem reinforcement learning is used for.
```

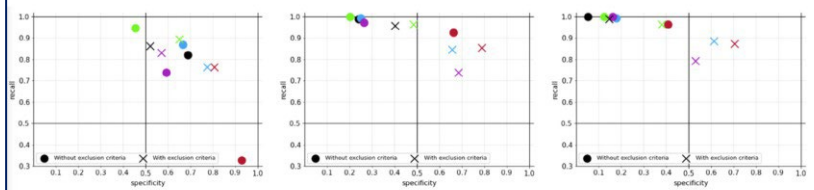
Instructions

```
Decide if the article should be included or excluded from the systematic review.
I give the title and abstract of the article as input.
Only answer INCLUDE or EXCLUDE.
Be lenient. I prefer including papers by mistake rather than excluding them by mistake.
```

Task

```
-Title: PARMOREL: a framework for customizable model repair
-Abstract: In model-driven software engineering, models are used in all phases of the development process[…]
```


Which prompting strategy to use?

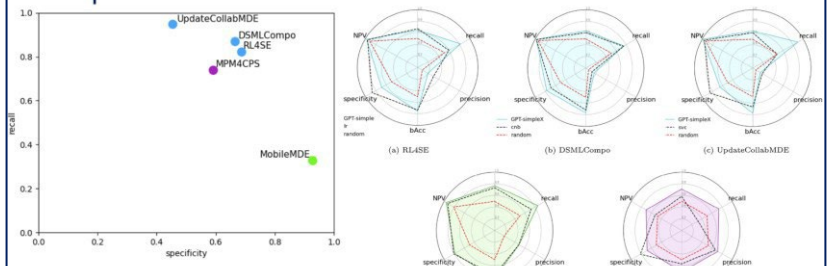


(a) Zero-shot (*Simple* and *SimpleX*) (b) Few-shot (*Positive* and *PositiveX*) (c) Few-shot with negative (*Balanced* and *BalancedX*)

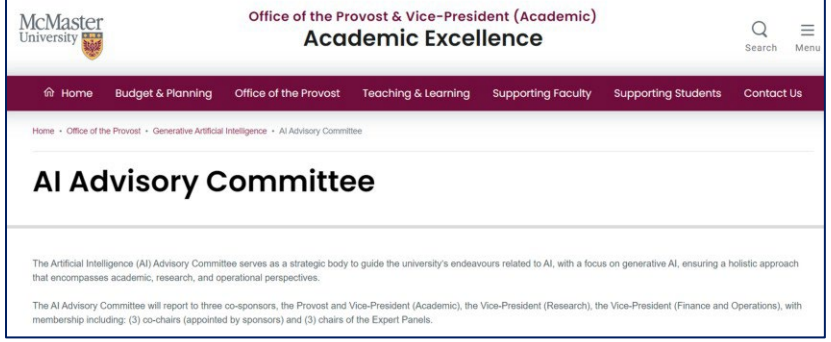
Answer to RQ1

Zero-shot prompts tend to perform better than few-shot prompts. Listing exclusion criteria tends to increase specificity. Using negative shots deteriorates the performance.

Corpus characteristics



Answer to RQ3

The characteristics of the corpus affect the performance profile of ChatGPT. Different corpus characteristics might favor different prompt strategies.

Perspectives on the Use of AI in Research