Could large language models and/or AI-based automation tools assist the screening process?

Dr. Siw Waffenschmidt

Head of the Information Management Department

Institute for Quality and Efficiency in Health Care (IQWiG)

Agenda

- Background
- What does the literature say about ML?
- Where are we going? Where are we now!
- Conclusion

BACKGROUND

IQWiG: scientifically independent HTA institution in Germany

- examines the benefits and harms of medical interventions for patients and other affected persons
- provides information on the advantages and disadvantages of different treatments and diagnostic procedures

IQWiG‘s work is

- evidence-based: specified in IQWiG’s General Methods
- independent: no influence on content of reports by payers, service providers, industry organizations or politicians
- patient-orientated: assessment of patient-relevant outcomes, involvement of patients and other affected persons
- transparent: publication of all documents relevant for reports and of the methods paper; disclosure of conflicts of interest by all persons involved in reports (employees, external experts etc.)

Study selection process – „screening“

- Methodological standard:

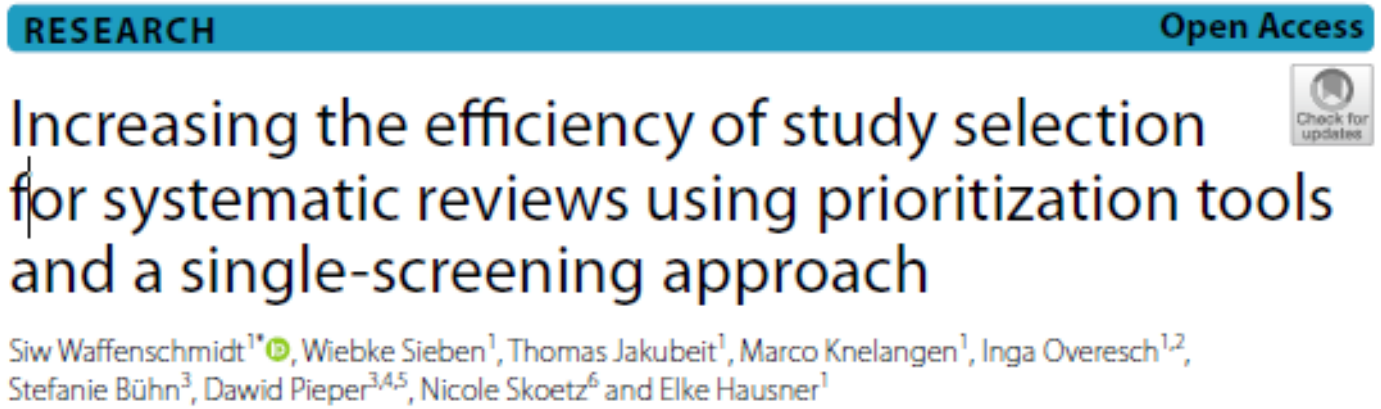
All selection steps are performed by 2 persons independently of each other. Discrepancies are resolved by discussion.

- huge work load / time savings possible
- (Rule of thumb: 1000 citations = 100 full texts = 10 included)

IQWiG:

- ca. 200-300 searches (between 100-5000 hits mostly)
- Screening tool

IQWiG activities: prospective validation study on ML



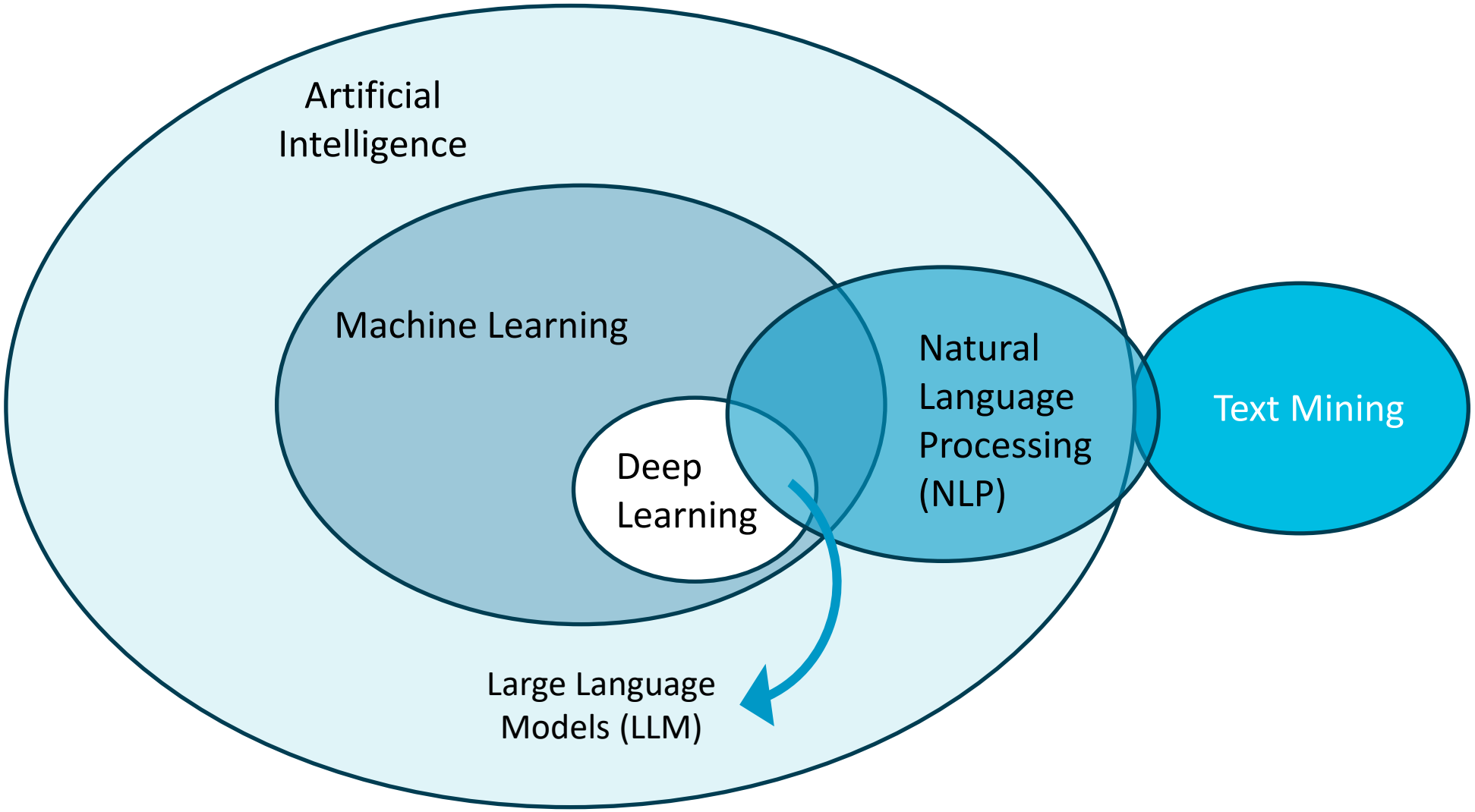
RESEARCH Open Access

Increasing the efficiency of study selection for systematic reviews using prioritization tools and a single-screening approach

Siw Waffenschmidt[1*], Wiebke Sieben[1], Thomas Jakubeit[1], Marco Knelangen[1], Inga Overesch[1,2], Stefanie Bühn[3], Dawid Pieper[3,4,5], Nicole Skoetz[6] and Elke Hausner[1]

	Number of screenings	Proportion of relevant citations after 50%
EPPI	N=10	88% [43-100]
Rayyan	N= 7	66% [0-100]

What is AI?


Artificial Intelligence
Machine Learning
Deep Learning
Natural Language Processing (NLP)
Text Mining
Large Language Models (LLM)

Reproducible ML-Approaches

Ranking

(Most Screening Tools)

- Machine Learning Style: **Active** Learning
- Needs Human Screening Decisions (IN and OUT)
- Machine Learning Algorithm determines ranking order

Pre-trained Classifier

(e.g. RobotSearch, EPPI)

- Machine Learning Style: **Supervised** Learning
- Needs a Labelled Development Set (e.g. RCTs versus not RCT) for Training
- Classification according to Machine Learning training result

Clustering (Instant Classifier)

- Machine Learning Style: **Unsupervised** Learning
- No labelling or pre-training necessary
- Classification according to Machine Learning algorithm

LLM versus ML-Screening

ML Screening

- Reproducible Results
- Custom Screening Algorithms
- Moderate Algorithm Size
- Validated Algorithms Available
- Training Data Is Transparent

Large Language Models

- No Perfectly Reproducible Results
- Language Models Predict the Similarity and Co-Occurrence of Words
- Very Large Algorithms
- No Validated Screening Process Available
- Training Data Unknown/ Intransparent

Foundation Models in Comparison

ChatGPT

- created by OpenAI
- fee-based API

Claude

- created by Anthropic
- fee-based API

Gemini

- created by Google (Alphabet)
- fee-based API

Meta Llama

- created by Meta (Facebook)
- open-source

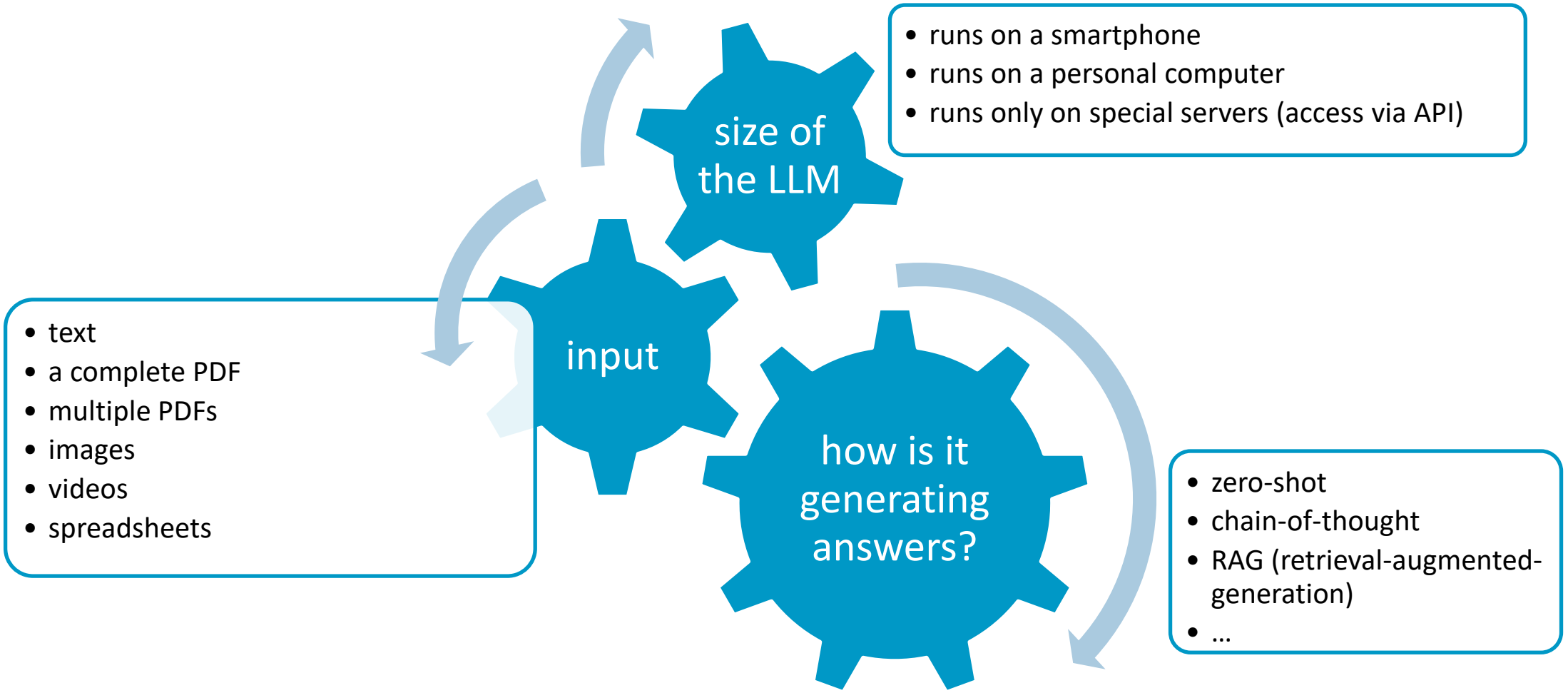
deepseek

- created by Deepseek (chinese)
- open-source

MISTRAL AI_

- created by Mistral AI (french)
- open-source

Differences in LLMs and LLM-based tools



WHAT DOES THE LITERATURE SAY ABOUT ML?

Background

<u>One year ago…..</u>

- ML algorithms mainly assist screening
  - Jiminez 2022 identified 63 tools; for screening 35 (55%)
  - Khalil 2022 identified 26 tools

The most common and used tools with Machine Learning applications

validated tools [according to Khalil 2022]

- **Rayyan**
- **AbstrackR**
- SWIFT-Active Screener
- **DistillerAI**
- EPPI-Reviewer
- Covidence (new ML feature)
- Cochrane RCT classifier (incorporated in various tools)

Practical applications of ML in screening

Blaizot 2022

AI approaches in published systematic reviews

12 systematic reviews, using 15 different AI methods, **11 methods for screening**

EPPI Reviewer, Abstrackr, Rayyan, K-means clustering algorithm, SWIFT-active screener, Wordstat/ QDA Miner

Tercero-Hidalgo 2022

application of AI tools in COVID-19 L.OVE database

28 of 3,000 COVID-19 reviews

EPPI Reviewer, SWIFT-Active Screener, Abstrackr, Evidence Prime

Summary

No significant uptake in systematic reviews

Feng 2022: Systematic review on accuracy of ML screening

Results

71 studies were included in the meta-analysis

The combined recall was **0.928** when achieving the maximized recall by optimizing the AI model.

Subgroup analysis (SVM/ other, number of hits, fraction of included studies) = still no recall above 95%

Conclusion

recall over 0.95 should be prioritized

At the current stage manual literature screening is still indispensable

WHERE ARE WE GOING? WHERE ARE WE NOW!

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

David Wilkins[1]

[1]Discipline of General Practice, The University of Adelaide

IQWiG

Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews

EUGENE SYRIANI, DIRO, Université de Montréal, Canada
ISTVAN DAVID, DIRO, Université de Montréal, Canada
GAURANSH KUMAR, DIRO, Université de Montréal, Canada

Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study

Eddie Guo[1]; Mehul Gupta[1], MD

[1]Cumming School of Medicine, Uni
[2]Temerty Faculty of Medicine, Univ

Annals of Internal Me

Sensitivity and Sp
and Abstract Scre

Viet-Thi Tran, MD, PhD; Gerald G
Lukas Schwingshackl, PhD, MSc;
Joerg Meerpohl, MD, PhD; and F

Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages

Qusai Khraisha[1,2] | Sophie Put[3] | Johanna Kappenberg[2] | Azza Warraitch[1,2] | Kristin Hadfield[1,2]

[1]Trinity Centre for Global Health, Trinity College Dublin, Dublin, Ireland
[2]School of Psychology, Trinity College Dublin, Dublin, Ireland
[3]Department of Education, York University, York, UK

Example LLM screening study: Tran 2024

Comparison

- retrospective diagnostic study
- “Indextest”: ChatGPT (GPT-3.5 Turbo)
- Reference standard: Conventional (human) consensus title/abstract double screening decision
- 5 systematic reviews: 2 COVID interventions, 1 methodological, 1 nutritional, 1 pharmacologic
- 22.665 citations (672, 4077, 6334, 6478, 5104)

Prompting

- (zero-shot) prompt chaining with instructions to provide reasoning for each PICOS element/ Outcome
- Balanced interpretation: <= 1 EXCLUDED PICS elements
- Sensitive interpretation: <= 2 EXCLUDED PICS elements

Tran 2024 zero-shot prompt example for PICO element population

Population

For the review on outpatient treatment for confirmed COVID-19 (Sommer I, Ann Intern Med, 2023) (3)

" I'm performing a systematic review. I am reading abstracts of clinical studies to assess whether or not they should be included in my review. Assess the population and answer using the following algorithm:
If the study includes hospitalized patients, your answer should contain the word \"EXCLUDE\" (in capital letters).
If the study includes patients in ICU, your answer should contain the word \"EXCLUDE\" (in capital letters).
If the study includes severe COVID-19 patients, your answer should contain the word \"EXCLUDE\" (in capital letters).
If the study includes outpatients (that is patients outside of hospital or non-hospitalized patients), your answer should contain the word \"INCLUDE\" (in capital letters)
If it the population is unclear, your answer should contain the word \"UNKNOWN\" (in capital letters)"

Tran 2024: Results

Balanced interpretation

Sensitivity: 81 – 97%

Specifitiy: 25 – 80%

Sensitive interpretation

Sensitivity: 94 – 99%

Specifitiy: 2 – 47%

Workload Savings

WSS@95%: 54 – 98% could be excluded without human screening

Re-test reliabilitity

Does ChatGPT always give the same answer?

ChatGPT makes different errors over time, but the overall error rate stays the same

AI/LLM-based screening approaches

Zero-shot prompting

- single prompt per screening decision
- without examples

Few-shot prompting

- single prompt per screening decision
- one or multiple examples for the correct answer

Prompt chaining

- multiple prompts per screening decision
- goal: breaking down a complex task
- each prompt is solving a simpler task (e.g. appraising one PICO element)

Chain-of-thought prompting (CoT):

- A technique where the LLM is guided to reason through a problem step-by-step in its response, by breaking down complex tasks into simpler parts to improve accuracy (Fleurence et al. 2024)
- reasoning can either take place in the background or be spelled out in the answer of the LLM

Majority voting

- considering multiple answers from multiple runs
- can be repeated answers of one LLM
- can be multiple LLMs each returning a single answer

Zero-shot

You

You are conducting a systematic review and meta-analysis, focusing on a specific area of medical research. Your task is to evaluate research studies and determine whether they should be included in your review. To do this, each study must meet the following criteria:

Target Patients: Adult patients (18 years old or older) diagnosed with or suspected of having infection, bacteremia, or sepsis.
Intervention: The study investigates the effects of balanced crystalloid administration.
Comparison: The study compares the above intervention with 0.9% sodium chloride administration.
Study Design: The study must be a randomized controlled trial.
Additionally, any study protocol that meets these criteria should also be included.

However, you should exclude studies in the following cases:

The study does not meet all of the above eligibility criteria.
The study's design is not a randomized controlled trial. Examples of unacceptable designs include case reports, observational studies, systematic reviews, review articles, animal experiments, letters to editors, and textbooks.
After reading the title and abstract of a study, you will decide whether to include or exclude it based on these criteria. Please answer with include or exclude only.

Title: ----------

Abstract

ChatGPT

Include

Chain-of-Though (CoT)

Zero-shot Prompt

Answer

You

You are conducting a systematic review and meta-analysis, focusing on a specific area of medical research. Your task is to evaluate research studies and determine whether they should be included in your review. To do this, each study must meet the following criteria:

Target Patients: Adult patients (18 years old or older) diagnosed with or suspected of having infection, bacteremia, or sepsis.
Intervention: The study investigates the effects of balanced crystalloid administration.
Comparison: The study compares the above intervention with 0.9% sodium chloride administration.
Study Design: The study must be a randomized controlled trial.
Additionally, any study protocol that meets these criteria should also be included.

However, you should exclude studies in the following cases:

The study does not meet all of the above eligibility criteria.
The study's design is not a randomized controlled trial. Examples of unacceptable designs include case reports, observational studies, systematic reviews, review articles, animal experiments, letters to editors, and textbooks.
After reading the title and abstract of a study, you will decide whether to include or exclude it based on these criteria. Let's think step by step. Please answer with include or exclude only.

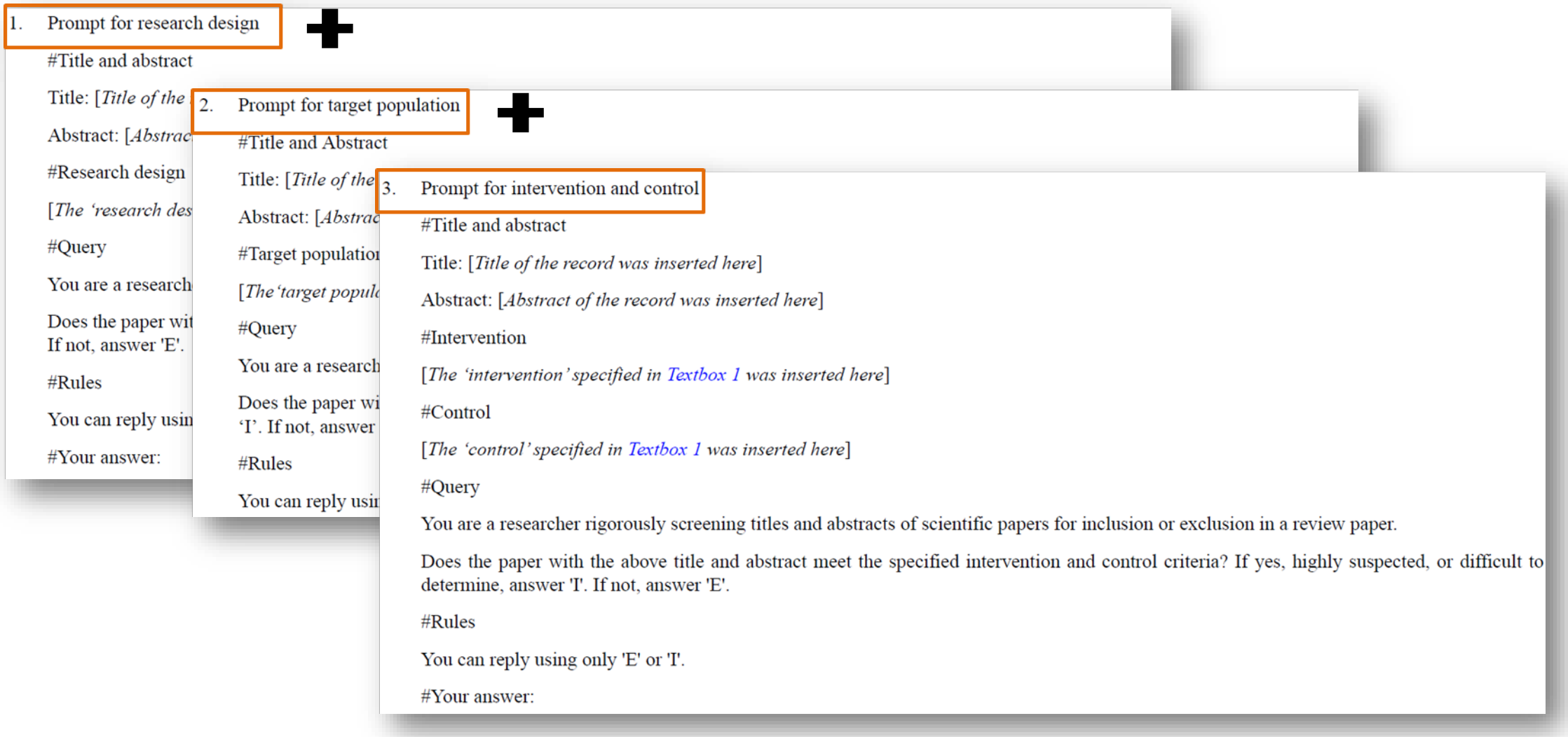
Title: ----------

Abstract

ChatGPT

Include

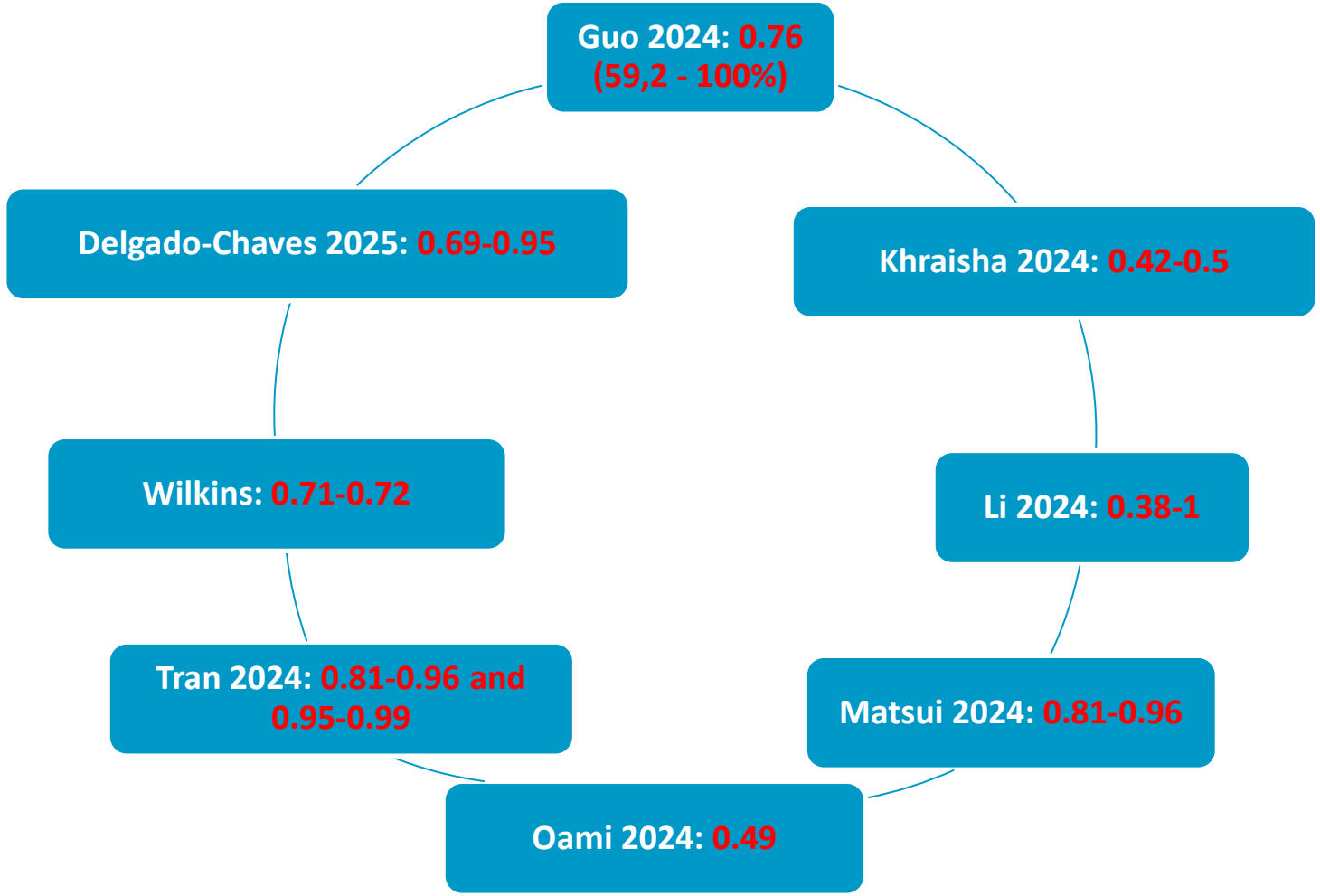
Prompt chaining



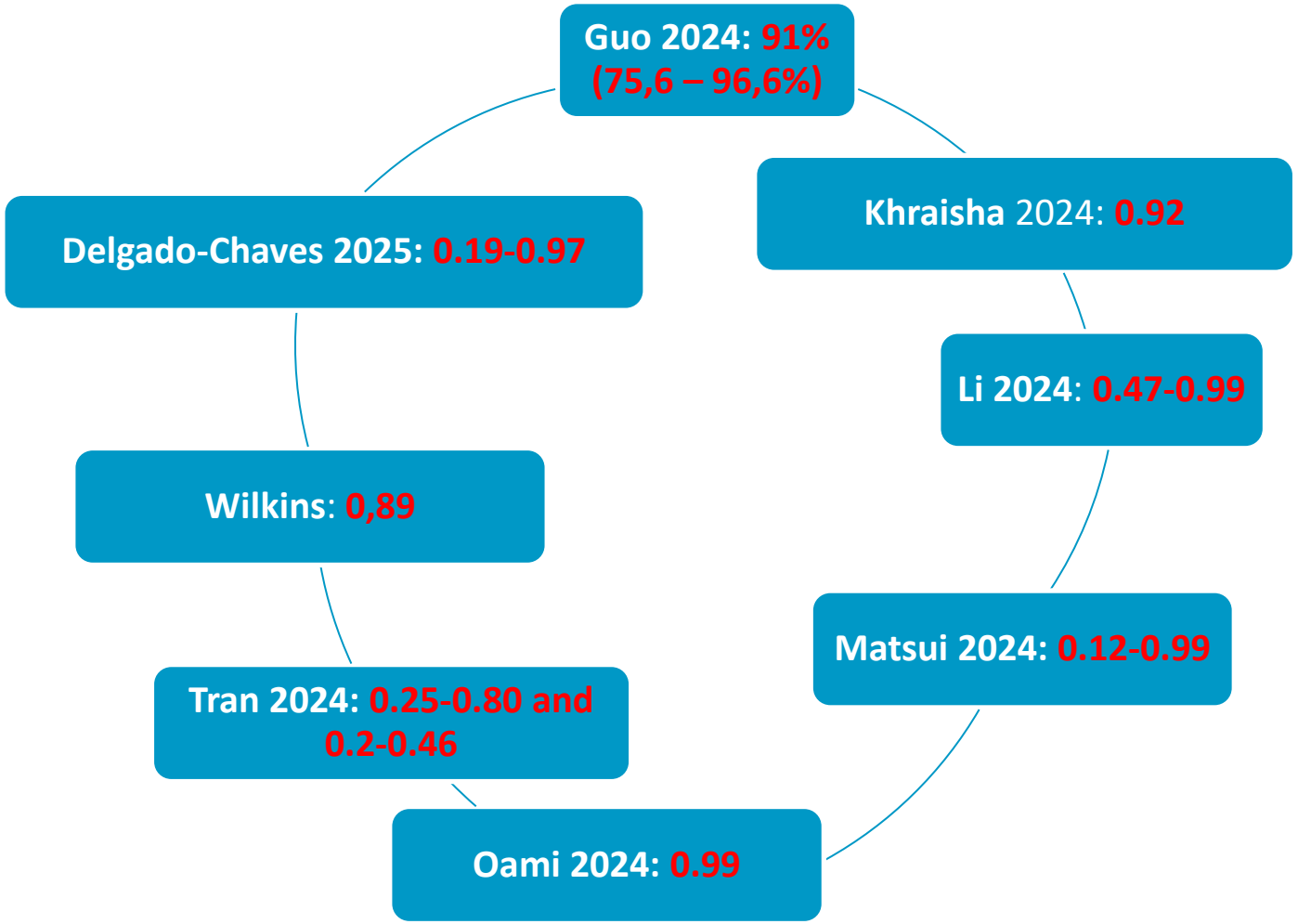
Technical approaches so far…

- Direct access
  - with chat interface
    - as a web application
    - via smartphone app
  - with API access (requires coding skills)
    - via programming language (e.g. Python, R)
    - via programming tools (Google Apps, Open refine,…)
- Indirect access
  - with intermediary service provider
    - Search engines
    - Screening tools
    - Literature software
    - Office software
    - …

Sensitivity performance so far …



Specificity performance so far …



Performance of different approaches

	Sensitivity	Specificity
GOLDSTANDARD Wong 2006: Medline – high sensitivity	99.1% (98.6 to 99.7)	71.0% (70.4 to 71.5)
Cochrane RCT classifier	99% (98%-99%)	63% (48-76)
Tran 2024 balanced	81-97%	20-80%

CONCLUSION AND DISCUSSION

Could ML tools assist the screening process?

- Uptake and implementation of automated tools slow [Khalil 2022]
- Skepticism remains [O'Connor 2019]
- Still no validated stopping rules available
- New (promising?) approaches: e.g. combined approaches
- Outdated technology?

Could large language models assist the screening process?

- ML vs. LLM:
  - easier to realize
  - sensitivity comparable results, but specificity much better
- explorative and retrospective studies – post hoc changes
- no validation study available

- LLMs might already outperforms SRs done by:
  - moderate English speakers screening English articles
  - non-Expert screeners (PhD Students, novice researchers, general practitioners)

Implementation or future application of LLMs

- Waiting for software solutions
- Learning how to use Python/ incorporate APIs seems technically not feasible for us

- Future combination of searching/ screening?
- LLMs as second screener or “RCT filter”/NOTing-Out?
- Are we (information specialists) future prompt engineers (e.g. translating PICOS for LLM)?

Implementation or future application of LLMs

- Waiting for software solutions
- Learning how to use Python/ incorporate APIs seems technically not feasible for us

- Future combination of searching/ screening?
- LLMs as second screener or “RCT filter”/NOTing-Out?
- Are we (information specialists) future prompt engineers (e.g. translating PICOS for LLM)?

References

Blaizot A, Veettil SK, Saidoung P et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods 2022; 13(3): 353-362. https://doi.org/10.1002/jrsm.1553.

Cierco Jimenez R, Lee T, Rosillo N et al. Machine learning computational tools to assist the performance of systematic reviews: A mapping review. BMC Med Res Methodol 2022; 22(1): 322. https://doi.org/10.1186/s12874-022-01805-4.

Delgado-Chaves FM, Jennings MJ, Atalaia A et al. Transforming literature screening: The emerging role of large language models in systematic reviews. Proc Natl Acad Sci U S A 2025; 122(2): e2411962122. https://doi.org/10.1073/pnas.2411962122.

Feng Y, Liang S, Zhang Y et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. Journal of the American Medical Informatics Association 2022; 29(8): 1425-1432. https://doi.org/10.1093/jamia/ocac066.

Guo E, Gupta M, Deng J et al. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res 2024; 26: e48996. https://doi.org/10.2196/48996.

Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022; 144: 22-42. https://doi.org/10.1016/j.jclinepi.2021.12.005.

Khraisha Q, Put S, Kappenberg J et al. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods 2024; 15(4): 616-626. https://doi.org/10.1002/jrsm.1715.

Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev 2024; 13(1): 219. https://doi.org/10.1186/s13643-024-02609-x.

Matsui K, Utsumi T, Aoki Y et al. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. J Med Internet Res 2024; 26: e52758. e52758. https://doi.org/10.2196/52758.

O'Connor AM, Tsafnat G, Thomas J et al. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst Rev 2019; 8: 143. https://doi.org/10.1186/s13643-019-1062-0.

Oami T, Okada Y, Nakada TA. Performance of a Large Language Model in Screening Citations. JAMA Netw Open 2024; 7(7): e2420496. https://doi.org/10.1001/jamanetworkopen.2024.20496.

Page MJ, Moher D, Bossuyt PM et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021; 372: n160. https://doi.org/10.1136/bmj.n160.

Tercero-Hidalgo JR, Khan KS, Bueno-Cavanillas A et al. Artificial intelligence in COVID-19 evidence syntheses was underutilized, but impactful: a methodological study. J Clin Epidemiol 2022; 148: 124-134. https://doi.org/10.1016/j.jclinepi.2022.04.027.

Tran VT, Gartlehner G, Yaacoub S et al. Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses. Ann Intern Med 2024. https://doi.org/10.7326/M23-3389.

Wilkins D. Automated title and abstract screening for scoping reviews using the GPT-4 Large Language Model. arxiv 2023. https://doi.org/10.48550/arXiv.2311.07918.

Institut für (IQWiG) Qualität und Wirtschaftlichkeit im Gesundheitswesen

Im Mediapark 8
50670 Köln

Telefon +49 221 35685-0
Telefax +49 221 35685-1

info@iqwig.de

www.iqwig.de
www.gesundheitsinformation.de
www.themencheck-medizin.de

@iqwig@wisskomm.social
@iqwig_gi@wisskomm.social

PRISMA Essential elements for systematic reviews using automation tools in the selection process [Page 2021]

Report how automation tools were integrated within the overall study selection process.

• If an externally derived **machine learning** classifier was applied (e.g. Cochrane RCT Classifier), […], **include a reference or URL to the version used.**

If the classifier was used to eliminate records before screening, **report the number eliminated in the PRISMA flow diagram as 'Records marked as ineligible by automation tools'.**

If an internally derived machine learning classifier was used to assist with the screening process**, identify the software/classifier and version**, describe how it was used (e.g. to remove records or replace a single screener) and trained (if relevant), and what **internal or external validation was done to understand the risk of missed studies or incorrect classifications.**

• If machine learning algorithms were used to prioritise screening (whereby unscreened records are continually re-ordered based on screening decisions), **state the software used and provide details** of any screening rules applied.