Evaluating the Reliability of a Custom GPT in Full-Text Screening of a Systematic Review

Rachel C. Davis[1, 2], Saskia S. List[1], Kendal G. Chappell[1], Espen Heen[1, 2]

[1]Department of Community Medicine and Global Health, University of Oslo, Oslo, Norway

[2] Correspondence: r.c.davis@studemed.uio.no , e.k.heen@medisin.uio.no

ABSTRACT

Objective: The purpose of this study is to evaluate the reliability and time-saving potential of a custom GPT (cGPT) in full text screening of a systematic review focusing on average 24-hour urine production and 24-hour creatinine excretion in populations.

Methods: A cGPT model, developed using ChatGPT4o (OpenAI Plus), was trained on a subset of articles previously assessed in duplicate by human reviewers. The human operator of the cGPT manually uploaded individual articles into the cGPT conversation with a standardized prompt. The outputs were coded to simulate cGPT in 3 different roles: (1) autonomous reviewer, (2) assistant to the 1st reviewer, and (3) assistant to the 2nd reviewer. Cohen's kappa was used to measure interrater agreement between cGPT and each human reviewer, as well as against human consensus decisions (the "gold standard"). The threshold for practical use was determined to be a cGPT-consensus kappa score which existed within the confidence intervals of at least one human-human pairing at inclusion/exclusion and exclusion reason.

Results: Of the three reviewer roles, cGPT as assistant to the 2nd reviewer was the only role which met the threshold for practical use, producing a cGPT-consensus kappa score of 0.733 (95% CI: 0.607, 0.859) compared to a human-human kappa range between 0.713 (95% CI: 0.606, 0.821) and 0.784 (95% CI: 0.656, 0.912) for inclusion/exclusion. In the classification of exclusion reason, cGPT-consensus kappa score was 0.632 (95% CI: 0.568, 0.696) compared to the human-human kappa range from 0.713 (95% CI: 0.606, 0.821) to 0.784 (95% CI: 0.656, 0.912). The study found that there is a clear time saving advantage to using cGPT in this way for full text screening, estimating 10.1 to 84.4 hours saved in the data set investigated here. cGPT as an autonomous reviewer or as assistant to the 1st reviewer did not meet reliability thresholds.

Conclusion: While cGPT did not have sufficiently reliable and accurate performance to replace human reviewers in full text screening, its use as an assistant holds promise in expediting the screening process, particularly with a large full-text corpus. There is a considerable deficit in published data exploring ChatGPT models for full text screening and more advanced models will require continued validation to determine which role is best suited to the capabilities of custom GPTs. More research is needed to establish a standardized threshold for practical use.

Keywords: ChatGPT, OpenAI, custom GPT, full text screening, automated text analysis, large language models, systematic review, automation, article filtration

BACKGROUND

Systematic reviews (SR) are critical for developing generalizable evidence that forms the basis of up-to-date best practice and healthcare policy (1). Such a study on studies usually starts broad by identifying all potentially relevant articles. After a thorough screening of title/abstract and full text for inclusion, a narrow subset of individual studies remain. Typically, the screening and data-extraction stages happen in duplicate with two independent reviewers making judgements and reaching consensus when discrepancies occur. This can translate to hundreds, if not thousands, of human work hours depending on the scope of the review (2). Due to the labor-intensive quality of conducting a SR, artificial intelligence (AI) is now intensively explored to take on some human roles.

The application of AI for the purpose of expediting SRs dates back almost two decades (3). An umbrella SR of health science SRs employing AI found that most studies are doing so in the title/abstract screening phases with AI in the role of one of the two reviewers (4). At the full text stage, one study reported that their text mining AI tool missed 10% of includable articles (5). Three out of five studies which reported critical concerns were those employing AI for data extraction. 83% of the included studies reported that the time and workload saving potential of AI in systematic reviewing could translate to hundreds of human work hours saved (4).

One promising AI model to aid in systematic reviewing is ChatGPT. ChatGPT is a large language model (LLM) with the ability to generate human-like text (6). A retrospective study of several different LLMs' ability to identify inclusion criteria in title/abstract screening compared to human performance from three systematic reviews identified ChatGPT4.0 to be the model with the best combined sensitivity and specificity. The other models explored were PaLM, Llama, Gemini, and Claude (7).

In the early version of ChatGPT (ChatGPT3.5) Mehrabanian found a human-AI congruency rate of 76% when given a random selection of pre-screened titles and abstracts (8). Alshami et. al used ChatGPT3.5 for all screening stages as well as data extraction of articles to evaluate the potential of AI to autonomously conduct SRs. In the screening stages, ChatGPT determined inclusion and exclusion in batches of five articles and classified included articles into four fields of interest related to the study. To accomplish this, ChatGPT was trained with a subset of 120 manually sorted articles with known categorization status. In title/abstract screening, ChatGPT was able to identify unrelated articles with a F1 score ranging from 85-of 90% and able to appropriately categorize abstracts with F1 scores of 86-90%. Classification metrics for correct inclusion/exclusion of articles were not reported for full text screening. The major limitations of ChatGPT were identified in data extraction, limited recall of information that was previously inputted, and incomplete responses (9). It is important to consider these limitations within the context of which version of ChatGPT was used. As of November 2023, one can create a custom GPT which can give recommendations on input information based on typed instructions and uploaded files. These custom GPTs will not "forget" instructions in the same way ChatGPT3.5 conversations were observed to do (10). To the best of our knowledge, a scientific evaluation of a custom GPT's performance in full text screening is yet to be published.

Various non-LLM AI tools are already present in SR software developed to assist in title/abstract and full text screening (11). Commonly, these tools provide predictions of article relevancy, however, do not provide determination of inclusion status. Nevertheless, their training parameters may serve as a guide when training new AI models such as ChatGPT:

- *Rayyan* is a leading semi-autonomous, free SR software. It uses machine learning and natural language processing in title/abstract screening. Rayyan trains its AI model with 50 included and 50 excluded articles before predicting the likelihood of non-scored articles to be included by giving them a rating from one to five stars. Rayyan refines its criteria throughout the review process as more studies are scored by reviewers (12).
- *Covidence* is another software used for SRs which recently began employing AI tools in title/abstract screening. Reviewers using Covidence must screen 25 articles as either included or excluded at the title/abstract stage for relevance to be predicted in subsequent articles. Covidence will then rank all remaining articles from most relevant to least relevant, continuing to refine its relevancy prediction and reshuffle as more studies are screened (13).
- *Abstrackr* also generates predictions of relevance but does so with text mining instead of natural language processing. It has been used at both title/abstract and full text screening (5, 14-15). Unlike Rayyan and Covidence, Abstrackr does not specify a set number of articles that must be manually screened before it will begin to generate a model. Like Covidence, it continues to refine its model as more studies are screened (5).
- *ASReviewer* is a free tool launched in 2021 to assist in title/abstract screening. This tool requires only one included and one excluded article before it begins to make predictions of relevance for all other unscreened articles (16).
- *SWIFT-Review* uses unsupervised and supervised learning to assist researchers in prioritizing articles in title/abstract screening. While it can begin to prioritize with as few as one included and one excluded article, performance improves with increasing training set size (17).
- *DistillerSR* uses active machine learning to expedite title/abstract screening through relevancy predictions. A study evaluating the performance of DistillerSR's prioritization algorithms used training subsets ranging from 25 to 200 articles and was able to achieve 95% sensitivity in testing. (18).
- *Nested Knowledge* has the option of replacing a second reviewer entirely in the title/abstract screening with a Robot Screener which can be activated after 40 excluded and 10 included articles (19).

It is generally agreed that the current capabilities of AI tools and models are not in a place to totally replace human reviewers for either title/abstract or full text screening (5, 9, 20-21). While AI may be in an appropriate place to assist or replace a second reviewer during title/abstract screening (19), its function in data extraction and quality appraisal is to validate human-made decisions (20). Additionally, it has been reported that reliability of text mining AI tools is better for intervention studies than for complex health service topics (5, 15, 20, 22).

Performance metrics of AI in screening tasks are typically reported as sensitivity (recall), specificity, false negative rate, proportion missed, positive predictive value (precision), F1-scores, and time savings (5, 9, 15, 21-23). Evaluation of inter-rater reliability between human and AI decisions can also be done using Cohen's kappa statistical analysis (24). AI tools used in screening for reviews tend to have high specificity but lower (<90%) sensitivity (14-15, 23, 25). Performance metrics for binary outcomes are further described with mathematical definitions in table 1 (22-23).

Table 1. Test performance metrics with mathematical definitions for binary outcomes.

METRIC	ACRONYM	DEFINITION	EQUATION[a]
Sensitivity (recall)	SENS	the proportion of correctly included articles to all includable articles	$\frac{TP}{TP+FN}$
Specificity	SPEC	the proportion of correctly excluded articles to all excludable articles	$\frac{TN}{TN+FP}$
False negative rate	FNR	the proportion of incorrectly excluded articles to all includable articles	$\frac{FN}{TP+FN}$
Proportion missed	PM	the proportion of incorrectly excluded articles to all excluded articles	$\frac{FN}{TN+FN}$
Positive predictive value (precision)	PPV	the proportion of correctly included articles to all included articles	$\frac{TP}{TP+FP}$
Observed agreement (accuracy)	--	the proportion of true positives and true negatives to all articles	$\frac{TP+TN}{TP+TN+FP+FN}$
F1	--	the harmonic mean of precision and recall, where 1 indicates perfect precision and recall and 0 indicates total failure of the model.	$\frac{2TP}{2TP+FP+FN}$

[a] True positive (TP): number of articles correctly included, True negative (TN): number of articles correctly excluded, False positive (FP): number of articles incorrectly included, False negative (FN): number of articles incorrectly excluded.

The purpose of this study was to evaluate the reliability and time saving potential of a tailor made custom GPT (cGPT) in full-text screening within a SR. The work was motivated by the need to scrutinize more than a thousand articles in full text. The outcome variables in the SR mainly appeared as intermediary outputs in the includable articles. This required the cGPT to identify defined concepts rather than just the primary studies' main outcome(-s) (26).

METHODS

The articles used in this study were selected from an ongoing SR which began in 2023 (26). Briefly, the SR aims to describe the global meta-mean and dispersion of it's primary (population-level habitual 24-hour urine volume) and secondary (population-level habitual 24-hour creatinine excretion) outcomes, identified in primary studies of adult populations globally. Additionally, the association between several measurable background variables and these outcomes will be explored. The SR's corpus is screened in duplicate at the title/abstract stage and articles which pass the initial screening are again screened in duplicate at the full text stage. The articles selected for training and testing of the cGPT model – the primary corpus - were published between 2020 and 2024 and consisted of a primary training and testing set. Additional articles published between 2015 to 2019 were used to augment the subpopulation of pregnant and breastfeeding women (the PB corpus). The PB corpus was used exclusively for testing.

For the training and testing of the cGPT, articles were coded into their inclusion **status** ('included' as 1, 'excluded' as 0) and **exclusion reason** (coded as 1-9). In the event of an included article, the exclusion reason was coded as 0.

The list of exclusion reasons and their associated numbers used for coding cGPT responses are as follows:

1. *Review* – includes any type of review.
2. *Population < 16 years* - should not include subpopulations less than 16 years old.
3. *Population not free-living* – includes hospitals, prisons, research facilities, and other confined settings.
4. *Population preselected by diseases* - that may distort representation of the general population (27).
5. *Population with a condition that affects fluid turnover* – such as manipulated diet or intravenous fluid administration. Does not include pregnancy/breastfeeding or manipulated 24-hour fluid intake if fluid intake is measured and reported.
6. *Urine volume population n < 30* – the includable population as described in 1-5 must have a minimum of 30 individuals.
7. *Urine volume (24h) not measured* – studies which did not collect all urine produced in 24 hours.
8. *Urine volume not reported in relevant group* – 24h urine volume measured but unreported.
9. *Awaiting classification* – not enough information within the article to make a determination.

The researcher responsible for uploading, recording, and interpreting responses from cGPT was the *cGPT operator*. cGPT's performance was evaluated in three plausible screening roles:

(1) Assistant to the 1st reviewer: The uploaded articles were analyzed by cGPT and the analysis and coding was interpreted by the operator, who decided to code the article in agreement with cGPT or not. (2) Assistant to the 2nd reviewer: the uploaded articles were analyzed by cGPT and the analysis and coding was interpreted by the operator, who decided to code the articles in agreement with cGPT or not. The operator was able to validate cGPT's decisions with any tags or notes attached to the article left by the first human reviewer. (3) Autonomous reviewer: cGPT's analysis of inputted articles was coded without any interpretation by the operator.

cGPT's ability as a reviewer in these different roles was quantified and described as interrater agreement (Cohen's kappa), observed agreement, classification performance metrics, and time savings. All screening took place within the Covidence environment between June and September 2024.

1. AGGREGATION OF TRAINING/TESTING DATA

Beginning in the full text screening stage, 4 human reviewers (E. H., R. D., S. L., and K. G.) kept a log of the following information: article ID number, status, exclusion reason, study design, and field of study. Once a total of 250 articles were classified and catalogued in duplicate by reviewers, a subset of 50 articles (40 excluded, 10 included) was selected as the primary training set. The articles purposefully encompassed all study designs, fields, and exclusion reasons possible to expose cGPT to the greatest variety of scenarios. The remaining articles were relegated to the primary testing set. For practical purposes, one reviewer (E. H.) was the first screener for the entirety of the primary corpus.

The remaining 200 articles for testing were augmented with 38 full-text articles from general populations. They represented all articles screened in duplicate by human reviewers at that point in time. The human/human consensus status of the primary testing set was 63 included and 175 excluded.

Articles with pregnant or breastfeeding populations published between 2015 and 2019 were screened and an additional 64 full texts were imported into the Covidence environment. These articles served as the testing set for assessment of cGPT as assistant to the 1st reviewer and were comprised of 14 included

and 50 excluded articles, identified as the PB testing set. Exclusion reasons for both the PB testing and primary testing sets can be found in Table 2:

Table 2. Distribution of the different categories of exclusion from the human-human consensus.

EXCLUSION REASON		PRIMARY TESTING SET number	PB TESTING SET number
0	Included article	63	14
1	Review	10	2
2	Population < 16 years[a]	0	0
3	Non-free-living population	6	4
4	Diseased population	13	1
5	Condition affecting outcome	8	1
6	N<30	7	2
7	Outcome data not collected	41	19
8	Outcome data not reported	84	20
9	Awaiting classification	6	1
SUM		238	64

[a] Exclusion reason 2 was effectively filtered out during T/A screening and was not present at the full-text stage.

2. TRAINING

The first version of our cGPT was created on 23 August 2024 using the paid feature on OpenAI's ChatGPT platform. The cGPT was powered by ChatGPT4.0 during testing. Initially, cGPT received abridged versions of the protocol and relevant supplements with a short description included in its 'INSTRUCTIONS' text field in the CONGIFURATION tab of the general custom GPT environment. cGPT provided incomplete and inconsistent responses when prompted to assess the training articles, and it was determined that a more structured prompt-response construction would produce more accurate and complete responses. The complete configuration of cGPT can be found in Supplemental 1.

A flowchart was developed to aid cGPT in navigating each article to a final decision (Supplemental 2). cGPT was instructed to 'Navigate this article through the exclusion flowchart' with articles from the training set. Its responses provided explanations for its reasoning for excluding or proceeding at each step of the flowchart. Errors in these responses were addressed by creating 'patches' in the 'INSTRUCTIONS' explicitly advising cGPT on how to proceed when faced with a certain scenario.

The team trialed differing levels of complexity with the flowchart. A persistent issue was cGPT's reluctance to identify includable subpopulations when following the flowchart, resulting in inappropriate exclusion of the study altogether. The simplest solution proved to be the best; cGPT's primary prompt was changed to 'Navigate each population in this article through the exclusion flowchart individually'.

3. TESTING

For the purpose of testing, human-human consensus was considered the 'gold standard'. Testing ran from 25 to 30 September 2024. The cGPT operator was blind to consensus status and exclusion reason throughout testing. A new conversation with cGPT was opened and articles from the primary testing set were uploaded one at a time using the final prompt, described previously. Conversations are separate windows that initially have no history of dialogue, like sending a new email instead of continuing in a thread. No feedback was provided to cGPT, and no other prompts were used for the duration of testing. The cGPT operator coded both cGPT's raw evaluation of the articles as well as the operator's

interpretation of status and exclusion reason based on cGPT's evaluation and supported by notes and tags left by the 1st reviewer. This allowed for the roles of cGPT as an autonomous reviewer and cGPT as assistant to the 2nd reviewer to be coded in parallel. The primary testing set was assessed two separate times by cGPT so that kappa statistics could be calculated between cGPT test 1 and test 2. Test 1 was completed before beginning test 2. Finally, the PB testing set was used to assess cGPT as assistant to the 1st reviewer. The cGPT operator interpreted evaluations from cGPT and coded the operator's interpretation of status and exclusion reason as the 1st reviewer of the PB testing set.

Upon testing completion, the decisions of the operator and cGPT were pooled to identify the "best response" (BR) in the two assistant roles. For the primary testing set, decisions from test 1 were used. Best response was defined as selecting which response, either cGPT or the operator, agreed most closely with the consensus decision. This BR was meant to simulate what would happen in the live environment should a discrepancy occur between the operator and cGPT. In the case of such a discrepancy, the operator would select the decision believed most accurate and leave a note that the article could be classified in one of two ways. If the selected classification were in disagreement with the other human reviewer, consensus decision could be expedited by referring to the second possible classification described in the note.

4. STATISTICAL ANALYSIS

Acceptable performance of cGPT was agreed to be determined by whether the interrater agreement (Cohen's kappa score) between cGPT and the human-human consensus decision existed within the confidence intervals of at least one human/human reviewer pairing. This included interrater agreement at both the status and exclusion reason levels. The kappa score assumes no "true" classification in the interrater comparison. Therefore, we calculated additional metrics using the consensus as the "truth", in accordance with the definitions in the introductory part of this article. These metrics were calculated from the binary variable "status".

When performing kappa statistics for exclusion reason, both excluded articles with exclusion reasons coded 1 through 9 and included articles with exclusion reason coded 0 were included in analysis. Included articles were kept in the exclusion reason analysis since there can at times be a discrepancy in the status decision between cGPT in its various roles and the consensus decision.

R version 4.3.3 for Windows was used for all statistical calculation. Packages used in analysis included "psych" and "irr". The observed agreement and kappa statistic confidence intervals were calculated directly from equations in the R environment.

RESULTS

(1) Experiential observations

An important observation was that cGPT tended to "stumble" , a term coined by this team to describe the phenomenon whereby cGPT provided incomplete responses or failed to respond to the prompt altogether. When this occurred, the operator would reissue the prompt with the article attached. cGPT was observed to be more likely to stumble after it had stumbled a first time. It was also more likely to stumble during longer dialogues, however, there did not appear to be a set length that would precipitate disrupted responses. Longer dialogues also appeared to cause longer response times from cGPT. After observing these trends, it was determined that upon the first instance of cGPT stumbling, the most time-

efficient response was to begin a new conversation powered by cGPT. Test 1 required reinitiating the conversation four times to complete the primary testing set while test 2 required five.

The cGPT was able to reliably divide populations within an article when prompted to do so before navigating populations through the exclusion flowchart. Engineering its prompt in this way allowed for simplification of the flowchart coupled with detailed 'INSTRUCTIONS'. This combination produced responses that were complete and delivered in a prescribed format, with explanations for its decisions that allowed the cGPT operator to identify errors in its logic. In observational studies cGPT most often divided populations based on how the study reported its outcomes, and in intervention studies most often divided populations by intervention and control groups.

(2) cGPT versus operator agreement

The status decision (inclusion/exclusion) between cGPT and the operator during test 1 had an observed agreement of 0.899 (95% CI: 85.4, 93.4%) and a Cohen's kappa of 0.766 (95% CI: 0.639, 0.893). The exclusion reason decision (ten alternatives) between cGPT and the operator during test 2 had an observed agreement of 81.5% (95% CI: 76.6%, 86.4%) and a Cohen's kappa of 0.755 (95% CI: 0.689, 0.821).

(3) cGPT reproducibility

Between test 1 and test 2, cGPT's status decisions had an observed agreement of 0.777 (95% CI: 0.719, 0.829) and a Cohen's kappa of 0.493 (95% CI: 0.368, 0.618). cGPT's exclusion reason decision had an observed agreement of 0.622 (95% CI: 0.56, 0.684) and a Cohen's kappa of 0.495 (0.429, 0.562).

(4) Results of primary test set

In the following, the results are reported in this way: classification performance metrics calculated from the primary test set will be presented using the binary variable "status" (4.1). This will provide a gross overview of how cGPT compares to human reviewers when both are compared against consensus decision (the "gold standard").

Human-human and human-consensus scores will be presented second, both with their status and exclusion reason (4.2, table 4 and 5). These are the benchmarks to which cGPT in all its roles are compared.

Next, status and exclusion reason kappa scores from cGPT as autonomous reviewer will be presented (4.3, table 6 and 7), followed by cGPT as assistant to the second reviewer (4.4, table 8 and 9).

In the final section (5), all results from the pregnancy/breastfeeding (PB) testing set will be reported separately. These results include cGPT in the role of assistant to the 1st reviewer (table 10 and 11).

(4.1) Classification performance metrics

Table 3. Classification performance metrics of status decisions. A: Consensus (CON) between human-human reviewers, best response (BR) of cGPT as assistant to 2nd reviewer and the operator, sensitivity (SENS), specificity (SPEC), false negative rate (FNR), proportion missed (PM), precision (Pr).

REVIEWER		NUMBER OF ARTICLES	STATUS					
A	B		SENS (95% CI)	SPEC (95% CI)	FNR (95% CI)	PM (95% CI)	Pr (95% CI)	F1
CON	Autonomous cGPT	238	**0.778** (0.655, 0.873)	**0.897** (0.842, 0.938)	**0.222** (0.127, 0.345)	**0.082** (0.046, 0.1334)	**0.731** (0.609, 0.832)	**0.754**
CON	BR	238	**0.873** (0.765, 0.944)	**0.897** (0.842, 0.938)	**0.127** (0.056, 0.235)	**0.049** (0.021, 0.093	**0.753** (0.639, 0.847)	**0.809**
CON	E. H.	238	**1.00** (0.943, 1.00)	**0.989** (0.959, 0.999)	**0** (0, 0.057)	**0** (0, 0.021)	**0.969** (0.893, 0.996)	**0.984**
CON	R. D.	83	**0.906** (0.750, 0.980)	**0.980** (0.896, 0.999)	**0.094** (0.020, 0.250)	**0.057** (0.012, 0.157)	**0.967** (0.828, 0.999)	**0.936**
CON	S. L.	76	**1.00** (0.753, 1.00)	**0.968** (0.890, 0.996)	**0** (0, 0.247)	**0** (0, 0.059)	**0.867** (0.595, 0.983)	**0.929**
CON	K. G.	75	**1.00** (0.794, 1.00)	**0.983** (0.909, 1.00)	**0** (0, 0.206)	**0** (0, 0.062)	**0.941** (0.713, 0.999)	**0.970**

(4.2) Human reviewer agreement

The range of human-human reviewer status kappa scores was 0.872 (95% CI: 0.647, 1.00) to 0.961 (95% CI: 0.735, 1.00), table 4.

Table 4. Observed agreement and kappa scores of status decisions for human-human reviewer pairs. Consensus (CON).

REVIEWER		NUMBER OF ARTICLES	STATUS			
A	B		OBS. AG. (95% CI)	COHEN'S KAPPA (95% CI)	P - VALUE[a]	Z - VALUE
E. H.	R. D.	83	**0.940** (0.865, 0.980)	**0.872** (0.658, 1.000)	1.55e-15	7.97
E. H.	S. L.	76	**0.961** (0.889, 0.992)	**0.872** (0.647, 1.000)	2.73e-14	7.61
E. H.	K. G.	75	**0.987** (0.928, 1.000)	**0.961** (0.735, 1.000)	0	8.33
CON	E. H.	238	**0.992** (0.970, 0.999)	**0.979** (0.852, 1.000)	0	15.1
CON	R. D.	83	**0.952** (0.881, 0.987)	**0.897** (0.682, 1.000)	2.22e-16	8.18
CON	S. L.	76	**0.974** (0.908, 0.997)	**0.913** (0.689, 1.000)	1.33e-15	7.99
CON	K. G.	75	**0.987** (0.928, 1.000)	**0.961** (0.735, 1.000)	0	8.33

[a] Test kappa statistics for $H_0 = 0$, equal to random outcome of interrater reliability. A P-value of 0 means less than 1.0e-16.

The range of human-human reviewer exclusion reason kappa scores was 0.713 (95% CI: 0.606, 0.821) to 0.784 (95% CI: 0.656, 0.912), table 5.

Table 5. Observed agreement and kappa scores of exclusion reason decisions for human/human reviewer pairs. Consensus (CON).

REVIEWER		NUMBER OF ARTICLES	EXCLUSION REASON			
A	B		OBS. AG. (95% CI)	COHEN'S KAPPA (95% CI)	P - VALUE[a]	Z - VALUE
E. H.	R. D.	83	**0.783** (0.694, 0.872)	**0.717** (0.611, 0.823)	0	13.3
E. H.	S. L.	76	**0.855** (0.776, 0.934	**0.784** (0.656, 0.912)	0	12
E. H.	K. G.	75	**0.773** (0.679, 0.868)	**0.713** (0.606, 0.821)	0	13
CON	E. H.	238	**0.908** (0.871, 0.944)	**0.878** (0.813, 0.943)	0	26.4
CON	R. D.	83	**0.868** (0.795, 0.940)	**0.829** (0.726, 0.932)	0	15.7
CON	S. L.	76	**0.947** (0.897, 0.998)	**0.923** (0.794, 1.000)	0	14
CON	K. G.	75	**0.867** (0.790, 0.944)	**0.83** (0.719, 0.941)	0	14.7

[a] Test kappa statistics for $H_0 = 0$, equal to random outcome of interrater reliability. A P-value of 0 means less than 1.0e-16.

(4.3) cGPT as an autonomous reviewer

cGPT and consensus decisions had an observed status agreement of 0.866 (95% CI: 0.816, 0.906) and Cohen's kappa of 0.661 (95% CI; 0.534, 0.788). The range of cGPT and individual human reviewer observed status decision was 0.819 (95% CI: 0.720, 0.895) to 0.882 (95% CI: 0.787, 0.944), while Cohen's kappa range was 0.59 (95% CI: 0.364, 0.816) to 0.668 (95% CI: 0.447, 0.889), table 6.

Table 6. Observed agreement and kappa scores of status decisions between cGPT as autonomous reviewer and human reviewer pairs. Consensus (CON).

REVIEWER		NUMBER OF ARTICLES	STATUS			
A	B		OBS. AG (95% CI)	COHEN'S KAPPA (95% CI)	P - VALUE[a]	Z - VALUE
cGPT	CON	238	**0.866** (0.816, 0.906)	**0.661** (0.534, 0.788)	0	10.2
cGPT	E. H.	238	**0.866** (0.816, 0.906)	**0.665** (0.539, 0.792)	0	10.3
cGPT	R. D.	83	**0.819** (0.720, 0.895)	**0.6** (0.385, 0.815)	4.2e-08	5.48
cGPT	S. L.	76	**0.882** (0.787, 0.944)	**0.668** (0.447, 0.889)	3.13e-09	5.92
cGPT	K. G.	75	**0.853** (0.753, 0.924)	**0.59** (0.364, 0.816)	3.15e-07	5.11

[a] Test kappa statistics for $H_0 = 0$, equal to random outcome of interrater reliability. A P-value of 0 means less than 1.0e-16.

cGPT and consensus decisions had an observed exclusion reason agreement of 0.651 (95% CI: 0.591, 0.712) and Cohen's kappa of 0.533 (95% CI; 0.467, 0.599). The range of cGPT and individual human

reviewer observed status decision was 0.6 (95% CI: 0.489, 0.711) to 0.658 (95% CI: 0.551, 0.765), while Cohen's kappa range was 0.465 (95% CI: 0.353, 0.577) to 0.532 (95% CI: 0.464), table 7.

Table 7. Observed agreement and kappa scores of exclusion reason decisions between cGPT as autonomous reviewer and human reviewer pairs. Consensus (CON).

REVIEWER		NUMBER OF ARTICLES	EXCLUSION REASON			
A	B		OBS. AG (95% CI)	COHEN'S KAPPA (95% CI)	P - VALUE[a]	Z - VALUE
cGPT	CON	238	**0.651** (0.591, 0.712)	**0.533** (0.467, 0.599)	0	15.8
cGPT	E. H.	238	**0.655** (0.595, 0.716)	**0.532** (0.464, 0.600)	0	15.3
cGPT	R. D.	83	**0.627** (0.522, 0.731)	**0.521** (0.417, 0.625)	0	9.85
cGPT	S. L.	76	**0.658** (0.551, 0.765)	**0.513** (0.387, 0.639)	1.33e-15	8
cGPT	K. G.	75	**0.6** (0.489, 0.711)	**0.465** (0.353, 0.577)	4.44e-16	8.11

[a] Test kappa statistics for $H_0 = 0$, equal to random outcome of interrater reliability. A P-value of 0 means less than 1.0e-16.

(4.4) cGPT as assistant to the 2nd reviewer

The best response of the cGPT-assisted second reviewer and consensus status decisions had an observed agreement of 0.891 (95% CI: 0.844, 0.927) and Cohen's kappa of 0.733 (95% CI: 0.607, 0.859). The range of human 1st reviewer and cGPT-assisted 2nd reviewer observed status agreement was 0.831 (95% CI: 0.733, 0.905) to 0.882 (95% CI: 0.835, 0.920), while Cohen's kappa range was 0.623 (95% CI: 0.401, 0.845) to 0.715 (95% CI: 0.589, 0.841), table 8.

Table 8. Observed agreement and kappa scores of status decisions for best response (BR) of cGPT in the role of assistant to the 2nd reviewer and human reviewer pairs. Consensus (CON).

REVIEWER		NUMBER OF ARTICLES	STATUS			
A	B		OBS. AG (95% CI)	COHEN'S KAPPA (95% CI)	P - VALUE[a]	Z - VALUE
BR	CON	238	**0.891** (0.844, 0.927)	**0.733** (0.607, 0.859)	0	11.4
BR	E. H.	238	**0.882** (0.835, 0.920)	**0.715** (0.589, 0.841)	0	11.1
BR	R. D.	83	**0.831** (0.733, 0.905)	**0.64** (0.425, 0.855)	5.33e-09	5.84
BR	S. L.	76	**0.882** (0.787, 0.944)	**0.623** (0.401, 0.845)	4e-08	5.49
BR	K. G.	75	**0.88** (0.784, 0.944)	**0.678** (0.453, 0.903)	3.55e-09	5.9

[a] Test kappa statistics for $H_0 = 0$, equal to random outcome of interrater reliability. A P-value of 0 means less than 1.0e-16.

The best response of the cGPT-assisted second reviewer and consensus exclusion reason decisions had an observed agreement of 0.719 (95% CI: 0.661, 0.776) and Cohen's kappa of 0.632 (95% CI: 0.568, 0.696,). The range of human 1st reviewer and cGPT-assisted 2nd reviewer observed exclusion reason agreement was 0.64 (95% CI: 0.531, 0.749) to 0.711 (95% CI: 0.613, 0.808), while Cohen's kappa range was 0.519 (95% CI: 0.395, 0.643) to 0.627 (95% CI: 0.523, 0.731), table 9.

Table 9. Observed agreement and kappa scores of exclusion reason decisions for best response (BR) of cGPT in the role of assistant to the second reviewer and human reviewer pairs. Consensus (CON).

REVIEWER		NUMBER OF ARTICLES	EXCLUSION REASON			
A	B		OBS. AG (95% CI)	COHEN'S KAPPA (95% CI)	P - VALE[a]	Z - VALUE
BR	CON	238	**0.719** (0.661, 0.776)	**0.632** (0.568, 0.696)	0	19.3
BR	E. H.	238	**0.681** (0.621, 0.740)	**0.577** (0.511, 0.643)	0	17.2
BR	R. D.	83	**0.711** (0.613, 0.808)	**0.627** (0.523, 0.731)	0	11.8
BR	S. L.	76	**0.658** (0.551, 0.765)	**0.519** (0.395, 0.643)	2.22e-16	8.22
BR	K. G.	75	**0.64** (0.531, 0.749)	**0.532** (0.422, 0.642)	0	9.46

[a] Test kappa statistics for $H_0 = 0$, equal to random outcome of interrater reliability. A P-value of 0 means less than 1.0e-16.

(5) Pregnancy/breastfeeding testing set

Comparisons between the consensus status decision and the best response of cGPT as assistant to 1st reviewer as well as human response (either K. G. or S. L.) is found in table 10.

Table 10. Classification performance metrics of status decisions.

REVIEWER		NUMBER OF ARTICLES	STATUS					
A	B		SENS (95% CI)	SPEC (95% CI)	FNR (95% CI)	PM (95% CI)	Pr (95% CI)	F1
CON	BR	64	**0.500** (0.230, 0.770)	**0.920** (0.808, 0.978)	**0.500** (0.230, 0.770)	**0.132** (0.055, 0.253)	**0.636** (0.308, 0.891)	**0.560**
CON	HR	64	**1.000** (0.768, 1.000)	**0.980** (0.894, 0.999)	**0.000** (0.000, 0.232)	**0.000** (0.000, 0.073)	**0.933** (0.681, 0.998)	**0.966**

Consensus (CON), human reviewer (HR), best response (BR), sensitivity (SENS), specificity (SPEC), false negative rate (FNR), proportion missed (PM), precision (Pr).

The best response of the cGPT-assisted 1st reviewer and consensus status decisions had an observed agreement of 0.828 (95% CI: 0.713, 0.911) and Cohen's kappa of 0.455 (95% CI: 0.213, 0.697). The human reviewer and consensus status decisions had an observed agreement of 0.984 (95% CI: 0.916, 1.00) and Cohen's kappa of 0.955 (0.710, 1.00). The best response of the cGPT-assisted operator as 1st reviewer and human reviewer was 0.844 (95% CI: 0.731, 0.922) and Cohen's kappa of 0.52 (95% CI: 0.280, 0.160), table 11.

Table 11. Observed agreement and kappa scores of status decisions for best response (BR) of cGPT in the role of assistant to the 1st reviewer and human reviewer pairs. Consensus (CON), human reviewer (HR).

REVIEWER		NUMBER OF ARTICLES	STATUS			
A	B		OBS. AG. (95% CI)	COHEN'S KAPPA (95% CI)	P - VALE[a]	Z – VALUE
BR	HR	64	**0.844** (0.731, 0.922)	**0.52** (0.280, 0.760)	2.23e-5	4.24
BR	CON	64	**0.828** (0.713, 0.911)	**0.455** (0.213, 0.697)	0.0002	3.68
HR	CON	64	**0.984** (0.916, 1.000)	**0.955** (0.710, 1.000)	2e-14	7.65

[a] Test kappa statistics for $H_0 = 0$, equal to random outcome of interrater reliability. A P-value of 0 means less than 1.0e-16.

The best response of the cGPT-assisted 1st reviewer and consensus exclusion reason decisions had an observed agreement of 0.656 (95% CI: 0.540, 0.773) and Cohen's kappa of 0.567 (95% CI: 0.450, 0.685). The human reviewer and consensus status decisions had an observed agreement of 0.938 (95% CI: 0.878, 0.997) and Cohen's kappa of 0.917 (95% CI: 0.785, 1.00). The best response of the cGPT-assisted operator as 1st reviewer and human reviewer was 0.656 (95% CI: 0.540, 0.773) and Cohen's kappa of 0.565 (95% CI: 0.447, 0.683), table 12.

Table 12. Observed agreement and kappa scores of exclusion reason decisions for best response (BR) of cGPT in the role of assistant to the 1st reviewer and human reviewer pairs. Consensus (CON), human reviewer (HR).

REVIEWER		NUMBER OF ARTICLES	EXCLUSION REASON			
A	B		OBS. AG. (95% CI)	COHEN'S KAPPA (95% CI)	P - VALE[a]	Z – VALUE
BR	HR	64	**0.656** (0.540, 0.773)	**0.565** (0.447, 0.683)	0	9.39
BR	CON	64	**0.656** (0.540, 0.773)	**0.567** (0.450, 0.685)	0	9.46
HR	CON	64	**0.938** (0.878, 0.997)	**0.917** (0.785, 1.000)	0	13.6

[a] Test kappa statistics for $H_0 = 0$, equal to random outcome of interrater reliability. A P-value of 0 means less than 1.0e-16.

(6) Time and workload savings

The time required to screen a fixed number of articles was determined by the total number of articles screened divided by the number of articles screened per hour. Time savings was calculated as the difference of time required to screen this set between human reviewers and cGPT. For cGPT, internet quality, type of article, number of populations within the article, and variability in generative speed of responses are unforeseen variables that impacted the number of articles screened per hour. For human reviewers, type of article (including whether the full text could be located), internet quality, and other environmental factors influenced how many articles were screened per hour. Due to this wide variability, time savings are described as minimum and maximum. Out of six total observations, the minimum documented number of articles completed by human reviewers in one hour was 9 while the maximum was 31. Out of four total observations, the minimum number of articles screened by cGPT was 32 and the

maximum was 44. The 2020-2024 time-strata corpus of the present SR contains 1,057 full text articles for screening. Considering our minimum and maximum values, this translates to between 34.1 and 117.4 effective human hours and 24 to 33 cGPT hours to conduct a single screening on this time strata. The time savings then can be described as between 10.1 and 84.4 hours saved when using cGPT instead of a human reviewer. While the human reviewers can only keep their top screening speed for a few hours at a time, cGPT, by its nature, has only the duration limitations of the operator.

DISCUSSION

Consistent with what has been observed in previous studies, cGPT did score below 90% sensitivity with specificity outperforming sensitivity in all reviewer roles. Put plainly, cGPT is more likely to generate false negatives than false positives. The opposite can be observed for human reviewers, who in three out of four instances had higher sensitivity than specificity - meaning when in doubt, they included the article. cGPT in the role of assistant to the 2nd reviewer provided the best combined sensitivity and specificity, a phenomenon that is also reflected in a higher F1 score compared to the two other cGPT roles.

Kappa scores were our primary outcome of interest in determining what role cGPT was best suited for. The kappa score of cGPT as assistant to the 2nd reviewer and the consensus status decision were within the confidence interval of human/human reviewer pairings in two out of three measurements, with the third being borderline. Similarly, the status kappa between cGPT as an autonomous reviewer and consensus were narrowly captured within the confidence interval of two of three human/human reviewer kappa scores. cGPT as assistant to the 1st reviewer and consensus status kappa consistently performed poorer than human-human reviewer kappa and remained outside of the confidence intervals. cGPT as assistant to the 2nd reviewer performed better in all status kappa pairings than cGPT as an autonomous reviewer. This is key as both roles utilized the same dataset, and responses were coded using the same cGPT response. The same cannot be said for cGPT as assistant to the 1st reviewer, who's status kappa scores created wider confidence intervals and lower kappa scores than any other cGPT role. This is probably due to the smaller sample size and highly specialized (pregnancy/breastfeeding) quality of the testing set, a condition that was not as present in the training set. It is therefore with caution that comparisons between cGPT in the role of assistant to the 1st reviewer and the two other reviewer roles can be made.

Turning to the more complex exclusion reason decisions, the kappa score between the cGPT-assisted 2nd reviewer and consensus was within the confidence intervals of two out of three human-human reviewer pairs. The interrater agreement for exclusion reason decisions between cGPT as an autonomous reviewer and consensus was not within the confidence intervals of any of the human-human reviewer pairs, indicating that cGPT would not be appropriate to employ as an autonomous reviewer in this project for determining exclusion reasons. The same can be observed for cGPT in the role as assistant to the 1st reviewer, where exclusion reason decision with consensus did not produce sufficiently high kappa scores to be captured by any human-human reviewer confidence intervals.

Summarized, cGPT in the role of assistant to the 2nd reviewer was the only role in which sufficient interrater agreement scores were produced to meet the minimum requirement for use in this project. Ideally, LLMs like ChatGPT should be able to autonomously take on the role as a full text reviewer, Future models may have the analysis capabilities to do so.

The study of LLMs like ChatGPT in full text screening is still largely uncharted territory, so there is little to compare these results to, pre-prints excluded. Indeed, the singular example of LLMs in full text screening located by the authors of this manuscript did not quantitatively report ChatGPT's performance at this stage of screening but rather described experiences interfacing with the model (9). Further validation studies need to be conducted so results such as those presented here can be compared and a threshold for acceptable use established.

While Nested Knowledge is the only example of an AI tool explicitly designed for full text screening, its validation is reported almost exclusively with binary statistics (recall, precision, and F1), the exception being observed agreement (28). This does not allow for the nuance of exclusion reasons explored in our study, necessitating a more complex metric for interrater agreement. Furthermore, looking at the three reviewer roles cGPT could take on, observed agreement with the consensus status decision, used as the gold standard, ranged from 82.8% to 89.1%, compared to human reviewer and consensus agreement of 95.2% and 99.2%, meaning that for the best human-consensus agreement, ChatGPT will miss one additional article for each ten articles screened. For exclusion reason the range was 65.1% to 71.9% for cGPT and consensus decisions compared to 86.7% and 94.7% for human reviewer and consensus decisions. While the range for the cGPT/consensus status agreement is relatively close to human/consensus, the same cannot be said for the exclusion reason ranges. To compare interrater agreement with ten alternative responses, a Cohen's kappa is necessary. While there are no other studies to point to which have used Cohen's kappa for this specific reason, the authors of this study felt it was the most appropriate metric to compare interrater agreement.

LIMITATIONS & FUTURE DIRECTIONS

A clear limitation to determining the appropriateness of how best to utilize cGPT is the abridged and highly selected dataset used to test cGPT as assistant to the 1st reviewer. A comparable number of articles, and preferably a corpus which is not limited to any particular population (in this case, pregnancy and breastfeeding), would be needed to truly compare cGPT in the role of 1st reviewer's assistant to the other roles described here: The model may not have been sufficiently fit to evaluating pregnant and breastfeeding populations. Additionally, cross-validation was not included in cGPT's training. Rather, 50 articles were selected and the model was patched according to errors noted in responses to those articles. It is possible that cGPT was subsequently "overfit" to the 50 articles used in the primary training set. A future study with a greater abundance of training and testing material would benefit from training with a cross-validation method.

There is a high potential for bias when using human decision as "the gold standard" by which to compare a screening tool. While this bias is partially mitigated by using the consensus decision (the power of two reviewers in agreement) as the gold standard, there nevertheless remain borderline decisions that could be interpreted one of two ways. The kappa score captures this uncertainty within its metrics, while observed agreement does not.

Despite efforts to encourage reproducibility in its configuration, cGPT had a poorer kappa score with itself between test 1 and test 2 of the primary testing set than it did with consensus in either the assistant to 2nd reviewer or autonomous role. Care must be taken by researchers who choose to employ a custom GPT for any purpose due to this inconsistency in its responses. It has been put forward that perhaps using a "majority vote" with 4 different LLMs evaluating articles together may produce improved sensitivity scores and could address the issue of inconsistent responses (7). Alternatively, it is possible that this issue

may be abated by prompting cGPT with the same article five or more times and accepting the majority vote for classification. The present study was not streamlined to such a degree that requesting five iterations of analysis on a single article was feasible, nor were there resources enough to produce 4 different custom LLMs catered to the project. These may be opportunities to explore in the future.

cGPT at times would report logical fallacies within the boundaries of its responses. The three most common logical errors observed were: A population comprised of 30 or more individuals was classified as less than 30, resulting in an inappropriate exclusion reason 6; identifying that one of the outcomes was reported but excluding for reason 8 since the study did not report both of the outcomes; and inclusion based on reporting of urine analytes other than creatinine (calcium, iodine, etc.) in the absence of 24-hour urine volume reporting. These logical fallacies likely contributed to the performance gap between cGPT as autonomous reviewer and cGPT as assistant to the second reviewer, however, it is entirely possible subsequent ChatGPT models will not make such mistakes.

cGPT also expressed a tendency to drift in its exclusion reason assessment, randomly favoriting one exclusion reason over another. The most common "favorite" exclusion reasons were exclusion reason 3 (non-free-living population) and exclusion reason 5 (condition affecting urine volume). This bias is addressed in part by using cGPT as an assistant rather than as an autonomous reviewer, however since cGPT responses terminate when an exclusion reason is reached, the operator would be required to ask follow-up questions to discern the true exclusion reason. Since starting a new conversation with cGPT makes it "forget" this favoritism, a potential solution in the future may be to reinitiate a new conversation at fixed intervals (for example, after every 20th article).

Since screening is done in duplicate, the potential time savings only covers half of the workload of full text screening. Even with the fastest examples of human reviewing compared to the slowest example of cGPT reviewing, cGPT still outperforms human reviewers' screening speed. The question of whether the discrepancies generated by cGPT translate to a significant amount of time savings was not explored quantitatively here. Based on the difference of perfect agreement (100%) and observed agreement, cGPT in its three roles was observed to generate discrepancies at a rate of between roughly 10 to 18% at the status level and 29 to 40% at the exclusion reason level. Resolution of cGPT-generated discrepancies was first investigated by the cGPT operator. If the operator was in agreement with the human reviewer, resolution was reached by the operator without consulting the human reviewer. In practice, this saves time for the human reviewer who is not obliged to perform consensus on all but the most nuanced cases of disagreement. Even with the potential to save human work hours, time savings in vivo will depend on several factors including internet speed, availability and content of articles, and the unpredictability of cGPT generated responses as described throughout this article.

Since the outcomes of interest in our SR are often intermediary steps in the studies being extracted, the reporting of relevant outcomes can be less obvious than, for example, a review of drug outcome studies. If the SR's outcome and the studies' outcomes are aligned, ChatGPT may have performed even better with a kappa score closer to those of the human-human reviewer pairs.

The method used to develop the cGPT was designed to be one that researchers in the public health field who do not have significant computer science background could reproduce or mimic. With a more technical background, a research team could possibly create cGPT as an application programming interface (API) that would pull in articles from a cloud server, thus negating the need for an operator to manually upload one article at a time into the cGPT conversation. Indeed, one study comparing different

LLMs as title/abstract screeners was able to screen 200 title/abstracts in 20 minutes using a variation of this method (7). The purpose of the present study was both to evaluate the legitimacy of cGPT as a full text screener as well as describe the time saving potential of the method used. While its legitimacy is evaluated in detail here, further exploration into optimizing cGPT's timesaving potential is needed.

The present SR is a considerably large study at completion with data from 1990 to 2024 and thousands of articles requiring full text screening. It was feasible to take several hundred articles for the use of testing and training an AI model that can then be employed in the remaining articles. Many SRs will not have screening needs as demanding as those presented here, so there is likely a minimum threshold for size of SRs where the benefits of using a custom GPT outweigh the initial time investment to train and test it.

Using LLMs like ChatGPT in this way does not come without cost. There are environmental impacts, financial costs, social consequences, and the potential for a disruption in what is prioritized in this type of research (29-30). While the production of energy required to power the servers that support LLMs is becoming greener, it is unclear whether the costs outweigh the benefits from a larger systems perspective.

CONCLUSION

cGPT performed best when employed as assistant to the 2nd reviewer, although its use as an autonomous reviewer provided intriguing results. Based on the results of the study, cGPT as assistant to the 2nd reviewer is a suitable role for practical use in the present SR. There is clear time saving potential to utilizing cGPT as described here. A more sophisticated system for inputting articles and coding cGPT responses is needed to maximize the potential time-saving benefit. Investigating cGPT's role as either assistant or autonomous reviewer in a less complex SR may provide time saving potential which outpace the current study.

ETHICS STATEMENT

Not applicable.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

FUNDING STATEMENT

This study did not receive any funding.

DATA AVAILABILITY STATEMENT

All data produced in the present study are available upon reasonable request to the authors.

REFERENCES

1. Gopalakrishnan S, Ganeshkumar P. Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. J Fam Med Prim Care. 2013 Jan;2(1):9–14.
2. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017 Feb 1;7(2):e012545.

3. Yao X, Kumar MV, Su E, Flores Miranda A, Saha A, Sussman J. Evaluating the efficacy of artificial intelligence tools for the automation of systematic reviews in cancer research: A systematic review. Cancer Epidemiol. 2024 Feb 1;88:102511.
4. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods. 2022;13(3):353–62.
5. Giummarra MJ, Lau G, Gabbe BJ. Evaluation of text mining to reduce screening workload for injury-focused systematic reviews. Inj Prev. 2020 Feb 1;26(1):55–60.
6. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI Models: A Preliminary Review. Future Internet. 2023 Jun;15(6):192.
7. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev. 2024 Aug 21;13(1):219.
8. Mehrabanian M. AI-based article screening. Br Dent J. 2023 Dec;235(12):914–5.
9. Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. Systems. 2023 Jul;11(7):351.
10. OpenAI. OpenAI. 2023 [cited 2025 Jan 30]. Introducing GPTs. Available from: https://openai.com/index/introducing-gpts/
11. Ge L, Agrawal R, Singer M, Kannapiran P, De Castro Molina JA, Teow KL, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. Syst Rev. 2024 Oct 25;13(1):269.
12. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016 Dec 5;5(1):210.
13. Walton A. Covidence. 2023 [cited 2025 Jan 30]. Release Notes: December 2022 – Title and Abstract Screening Using Machine Learning. Available from: https://www.covidence.org/blog/release-notes-december-2022-machine-learning/
14. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev. 2015 Jun 15;4(1):80.
15. 15. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev. 2018 Mar 12;7(1):45.
16. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021 Feb;3(2):125–33.
17. Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shah MR, et al. SWIFT-Review: a text-mining workbench for systematic review. Syst Rev. 2016 May 23;5(1):87.
18. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening – impact on reviewer-relevant outcomes. BMC Med Res Methodol. 2020 Oct 15;20(1):256.
19. Nested Knowledge. Robot Screener [Internet]. Nested Knowledge - AutoLit - Screening. 2024 [cited 2025 Jan 30]. Available from: https://about.nested-knowledge.com/docs/robot-screener/

20. Paynter R, Bañez LL, Erinoff E, Lege-Matsuura J, Potter S. Commentary on EPC methods: an exploration of the use of text-mining software in systematic reviews. J Clin Epidemiol. 2017 Apr 1;84:33–6.
21. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. 2023 Jul 1;13(7):e072254.
22. Valizadeh A, Moassefi M, Nakhostin-Ansari A, Hosseini Asl SH, Saghab Torbati M, Aghajani R, et al. Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. BMC Med Res Methodol. 2022 Jun 2;22(1):160.
23. dos Reis AHS, de Oliveira ALM, Fritsch C, Zouch J, Ferreira P, Polese JC. Usefulness of machine learning softwares to screen titles of systematic reviews: a methodological study. Syst Rev. 2023 Apr 15;12(1):68.
24. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159–74.
25. Pham B, Jovanovic J, Bagheri E, Antony J, Ashoor H, Nguyen TT, et al. Text mining to support abstract screening for knowledge syntheses: a semi-automated workflow. Syst Rev. 2021 May 26;10(1):156.
26. Heen E, Madar A, Meyer H, Henjum S, List SS, Davis RC, et al. 24-hour urine volume and 24-hour urine creatinine excretion in general adult populations, a systematic review and meta-analysis. OSF [Internet]. 2024 Dec 17 [cited 2025 Jan 30]; Available from: https://osf.io/gqza3/
27. Heen E, Madar A, Meyer H, Henjum S, List SS, Davis RC, et al. 24-hour urine volume and 24-hour urine creatinine excretion in general adult populations, a systematic review and meta-analysis. Supplementary material IV – revision I 12.06.2024. OSF [Internet]. 2024 Dec 17 [cited 2025 Jan 30]; Available from: https://osf.io/zehpu
28. Sauca M. AI for Screening: How is it used, and how can we tell how accurate it is? [Internet]. Nested Knowledge. [cited 2025 Jan 30]. Available from: https://about.nested-knowledge.com/2024/08/02/ai-for-screening-how-is-it-used-and-how-can-we-tell-how-accurate-it-is/
29. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency [Internet]. Virtual Event Canada: ACM; 2021 [cited 2025 Jan 30]. p. 610–23. Available from: https://dl.acm.org/doi/10.1145/3442188.3445922
30. Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and Benefits of Large Language Models for the Environment. Environ Sci Technol. 2023 Mar 7;57(9):3464–6.