Title: Prompting is all you need: LLMs for systematic review screening

Author list:

Christian Cao[1, 2, *, §], Jason Sang[3, *], Rohit Arora[4], Robbie Kloosterman[1], Matt Cecere[1], Jaswanth Gorla[1], Richard Saleh[1], David Chen[1], Ian Drennan[1, 5, 6], Bijan Teja[7, 8], Michael Fehlings[9], Paul Ronksley[10], Alexander A Leung[11], Dany E Weisz[12], Harriet Ware[2], Mairead Whelan[2], David B Emerson[13], Rahul Arora[2, **], Niklas Bobrovitz[2, 14, **]

Affiliations

1. Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
2. Centre for Health Informatics, Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada
3. Stripe, Inc., San Francisco, CA, United States
4. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
5. Department of Emergency Services and Sunnybrook Research Institute, Sunnybrook Health Science Centre
6. Orange Air Ambulance and Critical Care Transport
7. Department of Anesthesiology and Pain Medicine, University of Toronto, Toronto, ON, Canada
8. Department of Anesthesia and Critical Care Medicine, St. Michael's Hospital, Toronto, Ontario, Canada
9. Department of Surgery, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
10. Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
11. Department of Medicine and Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, AB, Canada
12. Department of Newborn and Developmental Paediatrics, Sunnybrook Health Sciences Centre, Toronto, Canada
13. Vector Institute, Toronto, ON, Canada
14. Department of Emergency Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

*These authors contributed equally

**These authors jointly supervised

Corresponding Author:

§Christian Cao

Medical Sciences Building, University of Toronto

1 King's College Cir, Toronto, Ontario, Canada M5S 1A8.

Email address: christian.cao@mail.utoronto.ca

Abstract

Systematic reviews (SRs) are the highest standard of evidence, shaping clinical practice guidelines, policy decisions, and research priorities. However, their labor-intensive nature, including an initial rigorous article screen by at least two investigators, delays access to reliable information synthesis. Here, we demonstrate that large language models (LLMs) with intentional prompting can match human screening performance. We introduce Framework Chain-of-Thought, a novel prompting approach that directs LLMs to systematically reason against predefined frameworks. We evaluated our prompts across ten SRs covering four common types of SR questions (i.e., prevalence, intervention benefits, diagnostic test accuracy, prognosis), achieving a mean accuracy of 93.6% (range: 83.3-99.6%) and sensitivity of 97.5% (89.7-100%) in full-text screening. Compared to experienced reviewers (mean accuracy 92.4% [76.8-97.8%], mean sensitivity 75.1% [44.1-100%]), our full-text prompt demonstrated significantly higher sensitivity in four reviews ($p<0.05$), significantly higher accuracy in one review ($p<0.05$), and comparable accuracy in two of five reviews ($p>0.05$). While traditional human screening for an SR of 7000 articles required 530 hours and $10,000 USD, our approach completed screening in one day for $430 USD. Our results establish that LLMs can perform SR screening with performance matching human experts, setting the foundation for end-to-end automated SRs.

Main

Systematic reviews (SR) are rigorous forms of knowledge synthesis that involve the gathering, critical appraisal, and analysis of evidence. Recognized as the gold standard in evidence-based practice, SRs bolster decision-making across various domains including medicine, business, and agriculture, among others.[1] SRs are resource-intensive, typically requiring one year and upwards of $100,000 to complete.[2,3] These costs stem from the comprehensive processes of conducting detailed searches, screening articles, extracting data, analyzing findings, and report writing.[2–4] The screening phase is particularly demanding and typically involves two investigators working independently, and in duplicate, to identify articles that meet predefined eligibility criteria.[1,5] Investigators begin with an initial sensitive title and abstract screen, followed by an accurate screen of article full-texts.[1,5] Despite a growing catalog of tools and resources,[6] SR automation is elusive as existing tools only supplement human workflows and lack the performance required for independent decision making.[7]

The rise of large language models (LLMs), such as GPT, with advanced generative capabilities creates new horizons for streamlining and automating SR processes.[8,9] MedPrompt,[10] a collection of prompting techniques to optimize GPT4 performance on medical benchmarks, has demonstrated that generalist foundation models can surpass traditional model fine-tuning methods simply through better prompting. However, evaluations of LLMs in the medical domain have faced issues such as irreproducibility due to the use of web browser applications with hidden prompts, unrecorded model versions and settings, and the transient nature of chatbot histories. Furthermore, basic zero-shot prompting–where LLMs are only given task instructions without examples–likely underestimates model capabilities. Consequently, past zero-shot assessments of LLMs for SR screening have demonstrated low sensitivity/recall, failing to capture relevant studies and compromising their use for automation.[11–14]

Here, we present a comprehensive evaluation of LLM performance in SR screening and prompting innovations that enhance screening efficacy. First, we address a critical gap in SR automation and create *BenchSR*, a robust set of 10 SR datasets covering diverse medical question types and domains. We then introduce two prompting innovations: 'Framework Chain of Thought (CoT),' a novel prompting approach to guide models towards systematic reasoning against predetermined frameworks, and 'Instruction Structure Optimized (ISO)' prompting, designed to address LLM context-loss. We propose *Abstract ScreenPrompt* and *ISO-ScreenPrompt* for abstract and full-text screening, and demonstrate that their performance can match human reviewers. We evaluate our strategies across seven generalist LLMs (GPT3.5, GPT4-0125-preview, GPT4-Turbo-0409, GPT4o-0513, Gemini Pro, Mixtral-8x22, Mistral-Large) and demonstrate that our findings are model agnostic. Our work highlights the importance of applying rigorous prompting techniques to fully leverage the capabilities of LLMs, and serves as a guide for future researchers interested in performing LLM evaluations for SR screening

and beyond. Collectively, our datasets and findings illustrate the feasibility of LLM-automated SR screening and lay the groundwork for fully automating the SR workflow.

Results

Datasets and BenchSR

We curated *BenchSR*, a collection of 10 SR datasets comprising over 170,000 articles, spanning four of six Oxford Center for Evidence-Based Medicine (CEBM) question types,[15,16] and nine different clinical domains[17] (Table 1, Fig. 1). This compilation includes SR metadata (inclusion/exclusion criteria, study objectives), and the complete set of labeled articles (included, excluded) for each review.

Abstract Prompt Engineering

We evaluated the performance of three standard prompting methodologies on our training split dataset (SeroTracker [ST] train split; Methods, Testing Methodology) (Fig. 1, Fig. 2a, Table 2). Unless otherwise stated, we prompted the GPT4-0125-preview model with the ST train split. Initial tests with zero-shot prompting approaches (Methods Prompt Engineering, Table 2), adapted from Guo et al.,[11] revealed suboptimal performance, with an accuracy of 65.0%, 30% sensitivity, and 100% specificity (Fig. 2b, Supplementary Table 1). These results were improved with random few-shot prompting (78% accuracy, 56% sensitivity, 100% specificity) (Fig. 2b, Supplementary Table 1). Rephrasing prompts to be more inclusive also improved sensitivity and performance (+3.3% accuracy, +6.5% sensitivity; Supplementary Table 2, Table 2), and we applied this consideration when prompting thereafter. Finally, Zero-shot CoT prompting[18] was associated with notable improvements in accuracy (86.3% accuracy, 74.5% sensitivity, 98% specificity) (Fig. 2b, Supplementary Table 1), although model accuracy and sensitivity were relatively low for independent SR screening.

We analyzed zero-shot CoT outputs and identified inconsistencies in the generated reasoning structures for article inclusion (Supplementary Table 3). This insight inspired us to experiment with directing the LLM to articulate its reasoning according to the predefined inclusion and exclusion criteria. This novel zero-shot prompting approach was termed ‘Framework CoT’ and reported notable performance improvements (91.3% accuracy, 86.5% sensitivity, 96% specificity) (Fig. 2b, Supplementary Table 1, Methods Abstract Prompt Engineering).

Further analysis of incorrect zero-shot framework CoT model outputs revealed that the model generated incorrect inferences of review study objectives based on the inclusion and exclusion criteria (Supplementary Table 3). To counteract this, we inserted study objectives directly from the SR protocol or manuscript, and defined our prompting strategy as ‘ScreenPrompt’ (Table 2). ScreenPrompt was further refined by adding context about the limitations of abstract content and importance of article inclusion (‘+ abstract’, Supplementary Fig. 1a,

Supplementary Table 4). This final abstract-specific prompting strategy was referred to as 'Abstract ScreenPrompt' (Table 2, Supplementary Fig. 1a, Supplementary Table 4).

We found that our Abstract ScreenPrompt prompting approach had the best performance on our ST training dataset (94.3% accuracy, 94.5% sensitivity, 94% specificity) (Fig. 2b, Supplementary Table 1). To check for risk of overfitting, we evaluated our Abstract ScreenPrompt approach on our ST validation dataset, and found comparable performance (94.3% accuracy, 96% sensitivity, 92.5% specificity) (Fig. 2b, Supplementary Table 1), confirming that our prompt engineering process did not overfit the training dataset.

While few-shot prompting (adding additional labeled examples) is traditionally believed to enhance LLM performance,[19] our balanced few-shot GPT-CoT Abstract ScreenPrompt prompt was associated with decreased performance (85.8% accuracy, 98.0% sensitivity, 73.5% specificity). We hypothesized that we could modulate model specificity and sensitivity by adjusting the ratio of included and excluded few-shot examples (Supplementary Fig. 1b, Supplementary Table 5). Surprisingly, we found that our sensitivity analysis with varied ratios (9:1 inclusion- or exclusion-favored GPT-CoT Abstract ScreenPrompt few-shot examples) did not notably alter performance for inclusion-favored few-shot prompting (inclusion-favored: 85% accuracy, 98% sensitivity, 72% specificity), and was associated with a decrease in performance for exclusion-favored few-shot prompting (exclusion-favored: 83.8% accuracy, 98.5% sensitivity, 69% specificity).

Abstract Screening Performance across LLMs

We conducted a comparative analysis across GPT4-0125-preview, GPT4-Turbo-0409, GPT4o-0513, GPT-3.5, Gemini Pro, Mixtral-8x22, and Mistral-Large LLM models to assess performance differences. We utilized the same ST training dataset and Abstract ScreenPrompt prompting strategy (Fig. 2b, Supplementary Table 6). Our cross-model evaluation found that the GPT4-0125-preview model was associated with the greatest performance (94.3% accuracy, 94.5% sensitivity, 94% specificity), but also had the highest cost ($12.54, 400 abstracts). In contrast, the GPT3.5 model was associated with the lowest overall performance, but had high sensitivity (66.9% accuracy, 90.5% sensitivity, 43% specificity). The Gemini Pro model had relatively low performance (77.5% accuracy, 67.5% sensitivity, 84.8% specificity), but was free to run. We found a moderate drop in accuracy and sensitivity with the newer GPT4-Turbo-0409 (89.8% accuracy, 83.5% sensitivity, 96% specificity) and GPT4o-0513 models (89.0% accuracy, 79.5% sensitivity, 98.5% specificity). To confirm that our prompting innovations were model agnostic, we compared Zero-Shot and Abstract ScreenPrompt performance in our lowest performing (GPT3.5) and next best performing (GPT4-Turbo-0409) models (Supplementary Table 7). We found that Abstract ScreenPrompt was associated with improvements in accuracy in both models, with the greatest sensitivity gains in GPT4-Turbo-0409 (+70% sensitivity).

Generalizability of Abstract ScreenPrompt

We hypothesized that our Abstract ScreenPrompt could be readily adapted for abstract screening beyond the ST dataset. To test this, we applied our prompt to 10 distinct SR datasets within our *BenchSR* (Fig. 2c, Supplementary Table 8), using original study objectives and eligibility criteria for each dataset without modification. Our Abstract ScreenPrompt demonstrated high sensitivity across reviews (97.0% [range: 86.7-100.0%]), in contrast to zero-shot prompting (53.4% mean sensitivity [16.7-87.6%]). Both Abstract ScreenPrompt (88.4% [73.5-95.5%]) and zero-shot prompting (89.8% [71.5-98.0%]) produced comparable accuracy.

Our previous evaluations were performed with a single sampled generation. However, due to the stochasticity of LLM generations, the model may produce varying reasoning traces for a given prompt. Therefore, we applied 11-vote self-consistency (SC)[20] to Abstract ScreenPrompt on the ST test dataset, and the Reinfection dataset, which had the lowest specificity (Supplementary Fig. 1c, Supplementary Table 9). We found that Abstract ScreenPrompt-SC was associated with an overall gain in performance in both datasets across all metrics (ST: +2.8% accuracy, +2% sensitivity, +3.5% specificity; Reinfection: +1.9% accuracy, +1.7% sensitivity, +1.8% specificity). Additionally, we could modulate sensitivity and specificity by setting different self-consistency thresholds (number of votes needed) for article inclusion/exclusion (Supplementary Fig. 1d, Methods Prompt Engineering). Collectively, these findings suggest that our Abstract ScreenPrompt prompting strategy is readily generalizable and can perform well across different SR contexts.

Full-text Prompt Engineering

Building on the performance of Abstract ScreenPrompt, we evaluated the capabilities of the LLMs for full-text article screening (Fig. 3a, Supplementary Table 10-11, Table 2). Unless otherwise stated, we prompted the GPT4-0125-preview model with the full-text ST training dataset. Our original Abstract ScreenPrompt approach, tailored for inclusion, demonstrated high sensitivity (84.2% accuracy, 99% sensitivity, 69.2% specificity) (Fig. 3a). However, we sought to prioritize accuracy in downstream optimization steps.

While abstracts are concise summaries of study findings, full-text articles can span thousands of words. Therefore we hypothesized that adjustments to our prompt structure—without making semantic changes to prompt content—could aid full-text screening. This hypothesis was based on the 'lost-in-the-middle' phenomenon,[21,22] which refers to the variable information retrieval rates of LLMs within lengthy texts. Our experiments positioning the pre-prompt, inclusion criteria, and instruction prompt modules before the full-text article was associated with modest performance gains (Init: 92.5% accuracy, 94.5% sensitivity, 90.5% specificity), while placing them after the full-text articles reduced performance (Fin: 82.2% accuracy, 98% sensitivity, 66.3% specificity) (Supplementary Fig. 2a, Supplementary Table 11). With ‘init’

prompting, the model occasionally provided a single token decision with no additional reasoning (Supplementary Table 12). Appending additional instructions at the end of the prompt helped mitigate this inconsistency ('Init + Fin-Instructions', Supplementary Table 12). Further improvements were observed with our Framework CoT (init + fin) prompt, where we appended the complete set of prompt modules before and after the full-text article (94.8% accuracy, 95% sensitivity, and 94.5% specificity) (Fig. 3a, Supplementary Fig. 2a, Supplementary Table 11). Applying the same prompting strategy to abstract screening did not notably improve results, (Supplementary Fig. 2b, Supplementary Table 13) possibly because abstracts are too short to manifest the lost-in-the-middle phenomenon.

When analyzing the model outputs, we also found that LLM responses occasionally lacked sufficiently granular reasoning against each inclusion and exclusion criterion (Supplementary Table 12). We refined our Framework CoT prompt to elicit detailed CoT reasoning for each individual sub-criterion by numbering them, while preserving the original criteria content. We term this revised prompt as 'Numbered Framework CoT' and similarly found improved performance (95.2% accuracy, 98% sensitivity, 92.4% specificity) (Fig. 3b, Supplementary Table 10).

Merging 'Numbered Framework CoT' with our optimal prompt structure (init + fin) resulted in a prompting strategy dubbed '*Instruction-Structure-Optimized (ISO) ScreenPrompt*'. ISO-ScreenPrompt exhibited the best overall performance on our training dataset (95.5% accuracy, 94% sensitivity, 98% specificity) (Fig. 3b, Supplementary Table 10). Evaluation on the separate ST validation dataset confirmed the robustness of our approach, showing comparable performance (96.3% accuracy, 97.5% sensitivity, 95% specificity) and confirmed that we did not overfit our training dataset.

Full-text Screening Performance across LLMs

Consistent with our abstract findings, our comparative analysis between LLM models revealed that Gemini Pro and GPT3.5 models had poor performance across all metrics (Supplementary Table 14). Interestingly, Mixtral-8x22 (93.7% accuracy, 91.9% sensitivity, 95.4% specificity) outperformed Mistral-Large (86.9% accuracy, 77.8% sensitivity, 96% specificity). GPT4-0125-preview, GPT4-Turbo-0409, and GPT4o-0513 models all had similar performance (95.3-95.8% accuracy, 93.0-93.5% sensitivity, 97.5-98.5% specificity). We then compared Zero-Shot and ISO-ScreenPrompt performance in our lowest performing (GPT3.5) and best performing model (GPT4-Turbo-0409) to assess model agnosticism (Supplementary Table 16). We found that ISO-ScreenPrompt was associated with improvements in accuracy across all models, with the greatest sensitivity gains in GPT4-Turbo-0409 (+79.5% sensitivity).

Generalizability of ISO-ScreenPrompt

We assessed the generalizability of our ISO-ScreenPrompt approach for full-text screening across datasets from *BenchSR* (Fig. 3c, Supplementary Table 16). We found that ISO-ScreenPrompt demonstrated high sensitivity across reviews, with 97.5% (89.7-100.0%) mean sensitivity, in contrast to zero-shot prompting (65.6% [11.8-93.8 %]). ISO-ScreenPrompt (93.6% [83.2-99.6%]) and zero-shot prompting (93.1% [73.9-99.0%]) demonstrated comparable accuracy, but ISO-ScreenPrompt had slightly higher accuracy with lower variance. When compared to Abstract ScreenPrompt, ISO-ScreenPrompt had a modest improvement in accuracy while maintaining sensitivity.

We applied self-consistency (SC) to ISO-ScreenPrompt on the ST test dataset (Methods, Testing Methodology) and the Reinfection dataset (Supplementary Fig. 2c, Supplementary Table 17). We found that ISO-ScreenPrompt-SC was associated with an overall gain in performance in both datasets across all metrics, except sensitivity in ST (ST: +1% accuracy, -1% sensitivity, +3% specificity; Reinfection: +4.3% accuracy, +2.8% sensitivity, +5.1% specificity). We could also modulate sensitivity and specificity by setting different self-consistency thresholds for article inclusion/exclusion (Supplementary Fig. 2d). Collectively, these findings suggest that our ISO-ScreenPrompt prompting strategy is generalizable and can perform well across different SR contexts.

Real World Implementation Assessment and Benefit

Dual human screening is considered the gold-standard approach for SR screening workflows.[23,24] In this process, at least two reviewers independently perform title and abstract screening, resolving any discrepancies through consensus or decision by a third reviewer (Fig. 4a). Relevant abstracts are then moved to a full-text screen, and the process is repeated to culminate a 'full dual screen' that identifies the final set of articles for inclusion (Fig. 4a).

To evaluate the real-world applicability of our approach, we performed head-to-head comparisons between our Abstract ScreenPrompt and ISO-ScreenPrompt with the traditional gold-standard dual human screening workflow (Fig. 4a). We selected reviews using stratified probability sampling for each Oxford CEBM question type, with two spots for intervention benefit due to their higher prevalence in the dataset. The article inclusion/exclusion labels set by the original study authors were used as the gold-standard.

We recruited a team of four researchers with previous SR experience and performed a 'calibration' exercise to evaluate their baseline screening performance, finding acceptable performance (Supplementary Table 18, Methods Head-to-Head Comparisons). Reviewers then screened the randomly selected reviews according to the standard dual screening workflow. We found that human reviewers performed well with dual abstract screening (mean accuracy 94.6% [89.5-97.8%]), mean sensitivity 90.9% [84.1-100%]), but performance dropped with full

dual screening (mean accuracy 92.4% [76.8-97.8%], mean sensitivity 75.1% [44.1-100%]) (Supplementary Table 19-20).

Head-to-head comparisons between Abstract ScreenPrompt and dual abstract screening revealed comparable sensitivity across three reviews (difference in binomial proportions, two-sided $p>0.05$). However, human reviewers exhibited significantly higher accuracy in three reviews ($p<0.05$) (Fig. 4b, Supplementary Fig. 3a, Supplementary Table 19), while Abstract ScreenPrompt demonstrated higher sensitivity than humans in two reviews ($p<0.05$). When compared to single human-reviewer abstract screening, Abstract ScreenPrompt exhibited significantly higher sensitivity across three reviews ($p<0.05$), and comparable accuracy across four reviews ($p>0.05$) (Supplementary Fig. 3b-c). Further, head-to-head comparisons between ISO-ScreenPrompt and full dual screening revealed that ISO-ScreenPrompt had significantly higher sensitivity to humans across four reviews ($p<0.05$), significantly higher accuracy in one review ($p<0.05$), and comparable accuracy across two reviews ($p>0.05$) (Fig. 4c, Supplementary Fig. 3d, Supplementary Table 20).

We derived cost- and time-estimates for our LLM and human screening workflows (Supplementary Table 21-22; Methods, Cost Analysis). Traditional dual screening costs ranged from $1,385.67 to $67,872.00 USD (2,257-130,436 articles), while ISO-ScreenPrompt costs ranged from $196.43 to $12,661.30 USD at a compensation rate of $20 USD per hour. For Abstract ScreenPrompt, costs ranged from $55.25 to $4,122.49 USD, compared to $376.17 to $21,739.33 USD for single human-reviewer abstract screening (Supplementary Table 22; Methods Cost Analysis). Our LLM approach was also substantially faster. Where dual screening took approximately 69.3 to 3393.6 hours, our ISO-ScreenPrompt approach completed screening in 0.9 to 52.2 hours (Supplementary Table 22). Similarly, single human-reviewer abstract screening was estimated to take 18.8 to 1087.0 hours, whereas Abstract ScreenPrompt completed screening in 16 minutes to 15.2 hours (Supplementary Table 22). Moreover, implementing the new OpenAI batch API further reduced costs by 50% for both prompting methods (ISO-ScreenPrompt: $98.22 to $6,330.65 USD; Abstract ScreenPrompt: $27.63 to $2061.25 USD), and reduced the maximum screening time to under 24 hours.

Discussion

Systematic review workflows are encumbered by resource- and time-intensive screening processes. Despite efforts to automate SR screening, existing tools demonstrate inadequate performance and are unable to independently screen abstracts and full-texts.[7,25,26] Here, we leverage the capabilities of text-based LLMs, such as GPT4, for SR abstract and full-text screening. Our experiments spotlight the importance of strategic prompting, demonstrating substantial performance improvements over basic zero-shot prompting approaches, and serve as a valuable resource for researchers interested in performing LLM evaluations. In this context, we introduce Framework CoT and ISO-prompting as effective and generalizable strategies for enhancing SR screening. Our findings reveal that our Abstract ScreenPrompt and ISO-ScreenPrompt can achieve performance levels matching human reviewers, and in some cases can even surpass humans. Finally, we introduce *BenchSR*, a curated collection of SR datasets that may be used to evaluate the effectiveness of screening automation tools and support advancements in SR automation.

Previous studies evaluating LLM performance for abstract screening have primarily utilized zero-shot prompting approaches with unsatisfactory results. Guo et al.[11] reported 59.3-100% sensitivity with GPT4 models, and Gargari et al.[13] reported 38-69% sensitivity with GPT3.5 models. These observations demonstrate that zero-shot prompting, akin to assessing a race car's performance without shifting gears, likely underestimates the true capabilities of LLM models for downstream tasks. Zero-shot prompting has been frequently used for other tasks such as medical question answering,[27] medical code mapping,[28] and disease risk stratification.[29] Furthermore, these evaluations are often accompanied by other problematic practices, such as web browser-based evaluations (i.e., ChatGPT).[27,30–32] Unknown system prompts, variable GPT versions, and the transient nature of chatbot context in these settings severely limit research reproducibility. Future studies should adopt more sophisticated prompting techniques, such as reasoning elicitation techniques (i.e., CoT), few-shot, and task-specific prompting. Additionally, researchers should use API calls to support reliable and reproducible research. Our analysis across various model versions demonstrated that model updates do not always improve performance, and can occasionally lead to performance drops. This variability further highlights the importance of transparent model and parameter reporting for LLM evaluations.

Our work offers several novel prompt engineering insights. CoT reasoning instructs LLMs to break down questions into intermediate reasoning steps before generating answers. However, the unstructured nature of freeform rationales can result in errors.[33] For instance, we observed that some CoT responses did not elicit reasoning against exclusion criteria. In response, we introduced ‘Framework CoT,’ a novel prompting approach that leverages predefined criteria or ‘frameworks’ for LLMs to reason against. This facilitates a structured analysis that mimics

human cognitive processes, inducing the model to systematically consider each criterion before making a final decision.

Furthermore, adjusting the proportion of GPT-CoT few-shot label distributions (i.e., balanced, exclusion-favored, inclusion-favored) did not influence model accuracy. We hypothesize that the role of labeled examples in few-shot prompting is complex and likely dependent on the task, model, and prompt design.[34,35] The benefits of in-context learning may be attributed to the format of few-shot responses, which guide the model to better structure its output response.[34] Few-shot GPT-Cot also resulted in worse performance than zero-shot methods. We speculate that the decreased accuracy was due to semantic contamination,[36] where the model misinterprets the example reasoning as directly relevant to the task. Consequently, we observed heightened sensitivity at the cost of reliable reasoning, and highlight the nuanced challenge in applying few-shot prompting.

Next, we address the 'lost-in-the-middle' phenomenon,[21,22,37]where as context length increases, much like stretching dough, the LLMs' ability to adhere to instructions wanes as 'gaps' or 'holes' begin to appear within the context. This limitation of autoregressive LLMs is particularly evident in lengthy documents such as full-text articles. Our ScreenPrompt (init) prompt occasionally produced single-token outputs, possibly because the instructions to 'think step-by-step' were lost in context. In response, we developed *ISO-ScreenPrompt*, and repeated our prompt modules before and after the full-text content to capitalize on the LLMs' context retrieval strength at these points.[22] Additionally, we numerically labeled each sub-criterion to leverage the proficiency of LLMs with structured input formats.[38] This adjustment facilitated reasoning against each specific sub-criterion, rather than broader meta-criteria. Finally, we demonstrate that self-consistency can introduce additional performance improvements at the cost of additional generations. We find these simple methods are effective strategies for bolstering LLM performance with long-context documents and anticipate that our prompting techniques can extend to other text classification domains with structured decision frameworks, such as identifying patients eligible for clinical trials[39] and medical code mapping using electronic health records.[28]

Our research highlights that our *Abstract ScreenPrompt* and *ISO-ScreenPrompt* prompts can excel in abstract and full-text screening, surpassing previous tools such as Abstrackr, Rayyan, and RobotAnalyst.[25,26] Both prompts achieve high sensitivity, with a mean value of 97.0% for Abstract ScreenPrompt and 97.5% for ISO-ScreenPrompt, across ten different reviews. Our prompts also maintain robust accuracy, with 88.4% mean accuracy for Abstract ScreenPrompt and 93.6% mean accuracy for ISO-ScreenPrompt. Furthermore, we did not modify or attempt to optimize the eligibility criteria or study objectives from each review, highlighting that our approach can be readily implemented. Our findings ultimately underscore the efficacy of

prompt design in real-world applications and sets a new standard for SR screening automation efforts, which have historically demonstrated unsatisfactory performance.[25,26]

We address a critical gap within the realm of SR screening automation by providing one of the most comprehensive reproducibility evaluations for gold-standard dual human screening. Previous evaluations have been limited to title/abstract screening phases across three or fewer SRs.[23] In contrast, our study performed full dual screen evaluations across five reviews covering four common Oxford CEBM question types. Our reviewers were calibrated for screening proficiency, and revealed that dual human screening can be error prone, with many articles erroneously excluded at the full-text stage.

Crucially, our ISO-ScreenPrompt approach is the first to demonstrate that SR automation efforts can match, and in some instances, outperform gold-standard human dual screening. ISO-ScreenPrompt had higher sensitivity than dual screening in all reviews, with significantly higher sensitivity in four of five reviews. Furthermore, it maintained comparable or greater accuracy in three reviews. Barring full-text accessibility issues, this superior sensitivity and comparable accuracy could revolutionize the SR process by potentially eliminating the need for an initial human screening phase of abstracts. Instead, reviewers could initiate data extraction on the subset of articles deemed included by ISO-ScreenPrompt, and iteratively remove 'false-positive' articles.

Abstract ScreenPrompt, which displayed superior sensitivity relative to dual abstract screening, offers another viable alternative for immediate implementation. While it may not replace dual abstract screening workflows, it achieved comparable accuracy to single human reviewers in four reviews and exceeded human sensitivity in all five reviews, with significantly higher sensitivity in three reviews. These findings suggest that Abstract ScreenPrompt could reliably serve as a single human reviewer vote at the abstract stage without compromising quality. Finally, both ISO-ScreenPrompt and Abstract ScreenPrompt are associated with substantial cost- and time-savings. Where traditional SR screening can take weeks to months, our approach completed screening in minutes to hours, freeing reviewers for deeper scientific evaluations and accelerating synthesis of study conclusions.

Finally, our work addresses a critical gap in the realm of SR screening by introducing *BenchSR*, a growing collection of 10 SRs that span four types of medical questions across nine different clinical domains. Existing benchmarks for SR screening such as CLEF,[40] Seed Collection,[41] and Cohen[42] have important limitations. The CLEF and Seed Collection benchmark lack critical study metadata, such as detailed inclusion and exclusion criteria; and feature incomplete datasets with only article identifiers, necessitating additional efforts to gather article abstracts and/or full-texts. Furthermore, the Cohen benchmark is no longer publicly accessible.

BenchSR provides (i) a comprehensive set of all included and excluded abstracts (with article identifiers) found during the initial search, (ii) essential study metadata not typically reported (detailed eligibility criteria), sourced directly from the original study authors, and (iii) a non-online hosting solution to mitigate concerns about LLM pre-training data contamination. Study metadata is currently limited to eligibility criteria and study objectives, but will be expanded significantly in future work (i.e., data extraction templates). Future research should focus on enhancing transparency and completeness in reporting SRs, including detailed study metadata and data extraction templates, to better support and refine automation efforts. We invite other researchers to contribute to our growing *BenchSR* benchmark.

Our study has several limitations. Although we apply our analysis across a wide range of SRs, the generalizability of our findings to other clinical questions, non-English SRs and other review methodologies (e.g. scoping reviews) requires further study. However, our prompt templates would likely perform well in reviews with PICO-structured inclusion/exclusion criteria. Furthermore, our work may extend to non-English SRs as LLM models continually improve language accessibility. Our full-text analysis was limited by the availability of freely accessible texts in the PubMed Central (PMC) database. While we made efforts to enhance our dataset by manually scraping 'included' full-texts for select reviews, we did not manually gather all 'excluded' texts due to accessibility and resource constraints. Our LLM screening was also solely conducted on text content. Including figures and tables could potentially enhance performance and accuracy and warrants future investigation. Moreover, our analysis explored Gemini Pro, GPT3.5, and GPT4-0125-preview, GPT4-Turbo-0409, and GPT4o-0513, but not the Claude 3 LLMs due to regional availability issues. Due to GPT4 context length limitations (128k tokens), we only tested few-shot prompting on abstracts. Furthermore, while we surveyed a wide range of prompting techniques, we did not explore every possible method. The influence of subtle changes (i.e., word choice) may further optimize models. Our focus on developing ready-to-use and generally applicable prompts meant that we did not modify original SR inclusion and exclusion criteria. However, we suspect that additional optimizations could improve performance.

In conclusion, our study underscores the importance of deliberate prompting to achieve human-level performance for SR screening. The promising results from ISO-ScreenPrompt and Abstract ScreenPrompt marks a major advancement in the automation of SR processes. We encourage further research into the capabilities of LLMs for other SR tasks, such as data extraction and meta-analysis. Fully-automated SRs will revolutionize evidence-based practice and offer indispensable value to medicine and beyond.

Tables

Table 1: Descriptive Overview of *BenchSR* Datasets

CEBM Type	Dataset	Number of Articles (number of full-texts)	Clinical Domain (WoS)[17]
Prevalence *How common is the problem?*	SeroTracker	130436 (3659)	Infectious Diseases
Diagnostic Test Accuracy *Is this diagnostic or monitoring test accurate?*	PA-Testing	8000 (248)	Endocrinology & Metabolism
	Spinal	2233 (296)	Neurosciences and neurology
Prognosis *What will happen if we do not add a therapy?*	SVCF	2257 (95)	Pediatrics Cardiovascular System & Cardiology
	Infant-NO	1317 (53)	Pediatrics Respiratory System
Intervention Benefits *Does this intervention help?*	Reinfection	6724 (1256)	Infectious Diseases
	Sepsis	5034 (59)	General & Internal Medicine
	Meds-HA	9707 (1563)	Pharmacology & Pharmacy
	Calcium-HA	1939	Cardiovascular System & Cardiology

		(37)	
	PA-Outcomes	5376 (74)	Endocrinology & Metabolism Surgery
Intervention Harms *What are the COMMON/RARE harms?*	N/A	N/A	N/A
Screening *Is this (early detection) test worthwhile?*	N/A	N/A	N/A

Table 2: Abstract and Full-text Screening Prompts

Prompting Technique	Prompt Structure. Modules are enclosed in {} and were not inserted in the prompt.[11]
Zero-shot *Adapted from Guo et al.[11]	{Pre-Prompt} *You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article.* {Inclusion Criteria} {Exclusion Criteria} {Abstract in investigation} {Instructions} *Only type "YYY" for included articles or "XXX" for excluded articles to indicate your decision. Do not type anything else.* {Model output}
Random few-shot (k=10)	{Pre-Prompt} *You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article.* **{Inclusion Criteria}** **{Exclusion Criteria}** {Example Include n= 5} {Example Exclude n= 5} {Abstract in investigation} {Instructions} *Only type "YYY" for included articles or "XXX" for excluded articles*

	to indicate your decision. Do not type anything else. {Model output}
Zero-shot CoT	{Pre-Prompt} *You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article.* {Inclusion criteria} {Exclusion Criteria} {Abstract in investigation} {Instructions} ***Let's think step by step for why an article should be included or excluded.*** *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
Zero-shot Framework CoT	{Pre-prompt} *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} {Abstract in investigation} {Instructions} # *Instructions* ***We now assess whether the paper should be included from***

	the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met. *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
ScreenPrompt	{Pre-prompt, including **objectives**} ***Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.*** *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} {Abstract in investigation} {Instructions, including task considerations} *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving*

	reasons for why they are met or not met. *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
Abstract ScreenPrompt	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} {Abstract in investigation} {Instructions, including abstract considerations} # *Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* ***Studies that may not fully align with the primary focus of our inclusion criteria but provide data or insights potentially relevant to our review deserve thoughtful consideration. Given***

	the nature of abstracts as concise summaries of comprehensive research, some degree of interpretation is necessary. *Our aim should be to inclusively screen abstracts, ensuring broad coverage of pertinent studies while filtering out those that are clearly irrelevant. We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
Few-shot GPT-CoT Abstract ScreenPrompt (k=10)	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} **{GPT-CoT Include n= 5}** **{GPT-CoT Exclude n= 5}** {Abstract in investigation} {Instructions, including abstract considerations} *# Instructions* *We now assess whether the paper should be included from the*

	systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met. *Studies that may not fully align with the primary focus of our inclusion criteria but provide data or insights potentially relevant to our review deserve thoughtful consideration. Given the nature of abstracts as concise summaries of comprehensive research, some degree of interpretation is necessary.* *Our aim should be to inclusively screen abstracts, ensuring broad coverage of pertinent studies while filtering out those that are clearly irrelevant. We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
Full-text	
Numbered ScreenPrompt:	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {**Numbered** Inclusion criteria} ***1.*** *Humans of any age* ...

	11. *Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country* {**Numbered** Exclusion Criteria} ***1.*** *Non-human (e.g., in silico, animal, in vitro)* ... ***10.*** *Does not report the location at which the study took place* {Abstract in investigation} {Instructions} # *Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
ScreenPrompt (Init):	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:*

	{Inclusion criteria} {Exclusion Criteria} **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Abstract in investigation} {Model output}
ScreenPrompt (Fin):	{Abstract in investigation} **{Pre-prompt, including objectives}** *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* **{Inclusion criteria}** **{Exclusion Criteria}**

	{Instructions} *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
ScreenPrompt (Init + Fin-Instructions):	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or*

	excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met. *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Abstract in investigation} **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
ScreenPrompt (Init + Fin):	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria}

	{Exclusion Criteria} **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Abstract in investigation} **{Pre-prompt, including objectives}** *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* **{Inclusion criteria}** **{Exclusion Criteria}** **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every*

	predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met. *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.*
ISO-ScreenPrompt:	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {**Numbered** Inclusion criteria} {**Numbered** Exclusion Criteria} {Instructions} *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.*

	{Abstract in investigation} **{Pre-prompt, including objectives}** *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* **{Numbered Inclusion criteria}** **{Numbered Exclusion Criteria}** **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}

Figures

Figure 1: Infographic of study design. ScreenPrompt achieves SOTA performance for abstract and full-text SR screening. The LLM assessment component shows evaluation of the SeroTracker (ST) dataset (n=400) used in head-to-head human and LLM testing. Error bars represent 95% CIs for binomial proportions.

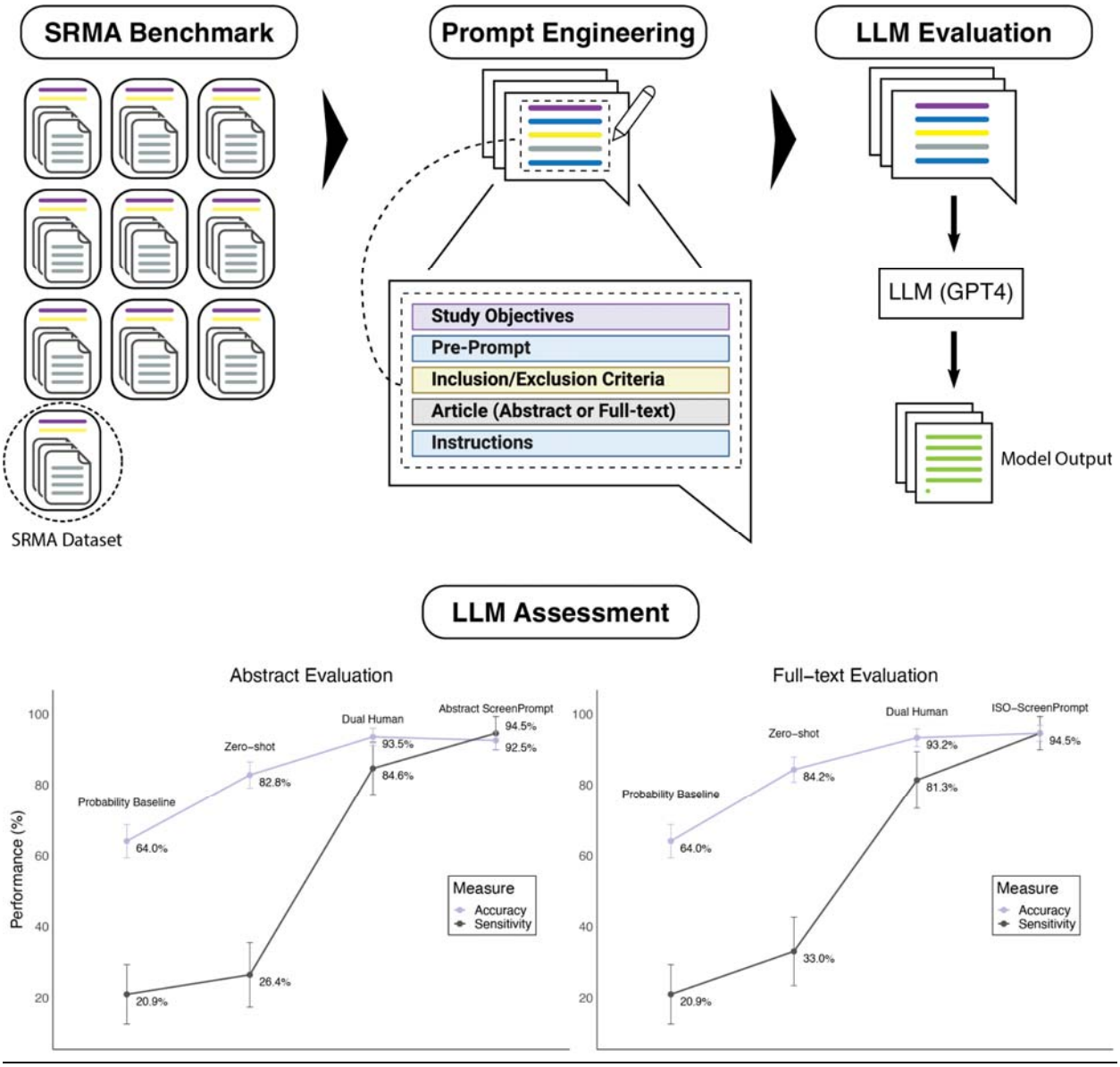
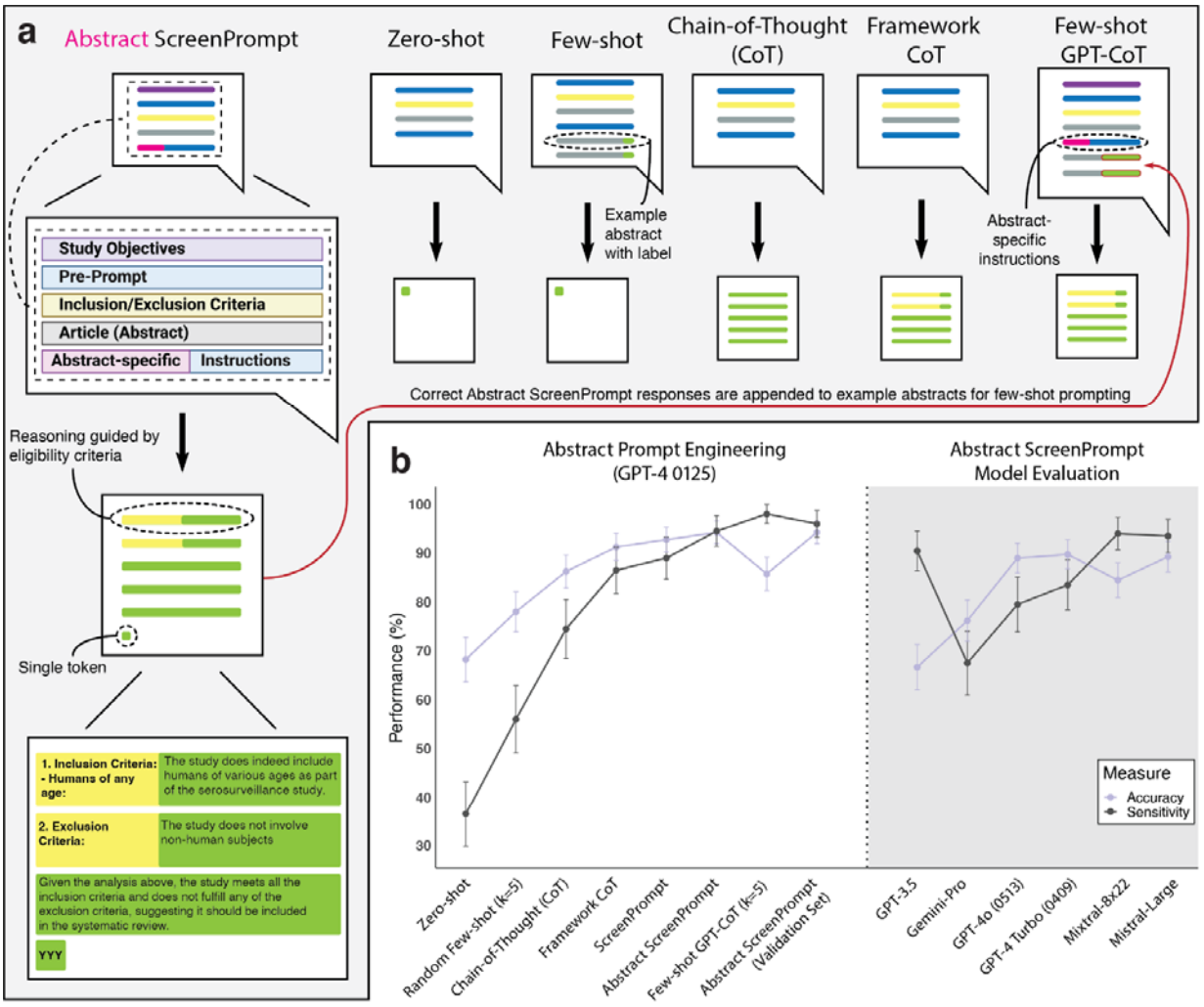


Figure 2: Abstract ScreenPrompt achieves SOTA performance for abstract screening, generalizing across studies a) Diagram of different prompting strategies used for abstract screening, including Zero-shot, Few-shot, Chain-of-Thought (CoT), Framework CoT, and Few-shot GPT-CoT. The Abstract ScreenPrompt approach integrates study objectives, inclusion/exclusion criteria, and abstract-specific instructions to guide reasoning. b) Performance comparison of different abstract prompting methodologies on the SeroTracker training, showing accuracy and sensitivity. Abstract ScreenPrompt is evaluated within the ST training dataset (n=400) across GPT4-0125-preview, GPT-3-5, Gemini-Pro, GPT4-o-0513, GPT4-Turbo-0409, Mixtral-8x22, Mistral-Large. Abstract ScreenPrompt is also separately evaluated on the ST validation dataset (n=400). Error bars represent 95% CIs for binomial proportions. c) Barplot displaying generalizability of zero-shot Abstract ScreenPrompt sensitivity and accuracy across 10 different systematic review datasets from *BenchSR*. Error bars represent 95% CIs for exact proportions with the Clopper-Pearson method.



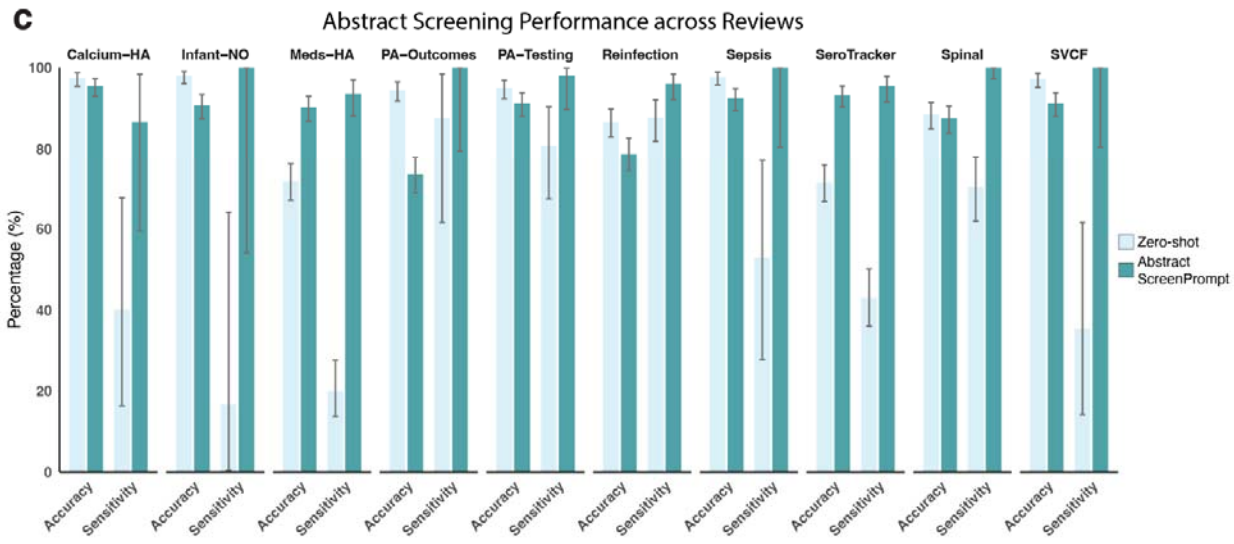


Figure 3: ISO-ScreenPrompt achieves SOTA performance for full-text screening, generalizing across studies a) Diagram illustrating the various prompting strategies used for full-text screening, including Zero-shot, Few-shot, Chain-of-Thought (CoT), Framework CoT, and Few-shot GPT-CoT. The ISO-ScreenPrompt approach incorporates repeated prompt modules (i.e., objectives, pre-prompt, eligibility criteria, instructions) and numbering to enhance performance. b) Performance comparison of different full-text prompting methodologies on the SeroTracker (ST) training dataset (n=400), showing accuracy and sensitivity. ISO-ScreenPrompt is evaluated across multiple models with the ST training dataset: GPT4-0125-preview, GPT-3.5, Gemini-Pro, GPT4o-0513, GPT4-Turbo-0409, Mixtral-8x22, and Mistral-Large. ISO-ScreenPrompt is also separately evaluated on the ST validation dataset (n=400). Error bars represent 95% CIs for binomial proportions. c) Barplot displaying the generalizability of the zero-shot ISO-ScreenPrompt sensitivity and accuracy across 10 different systematic review datasets from *BenchSR*. Error bars represent 95% CIs for exact proportions with the Clopper-Pearson method.

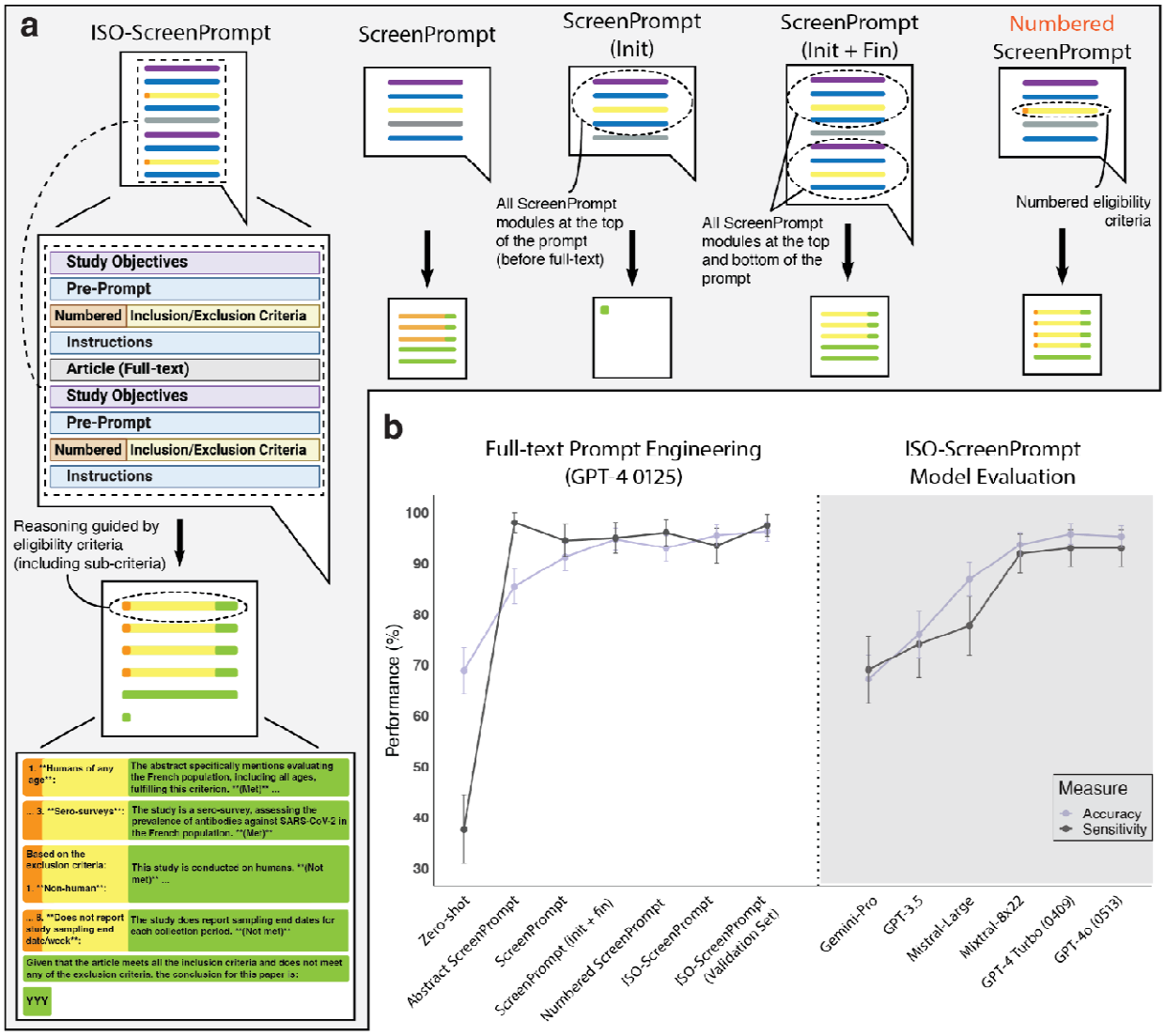
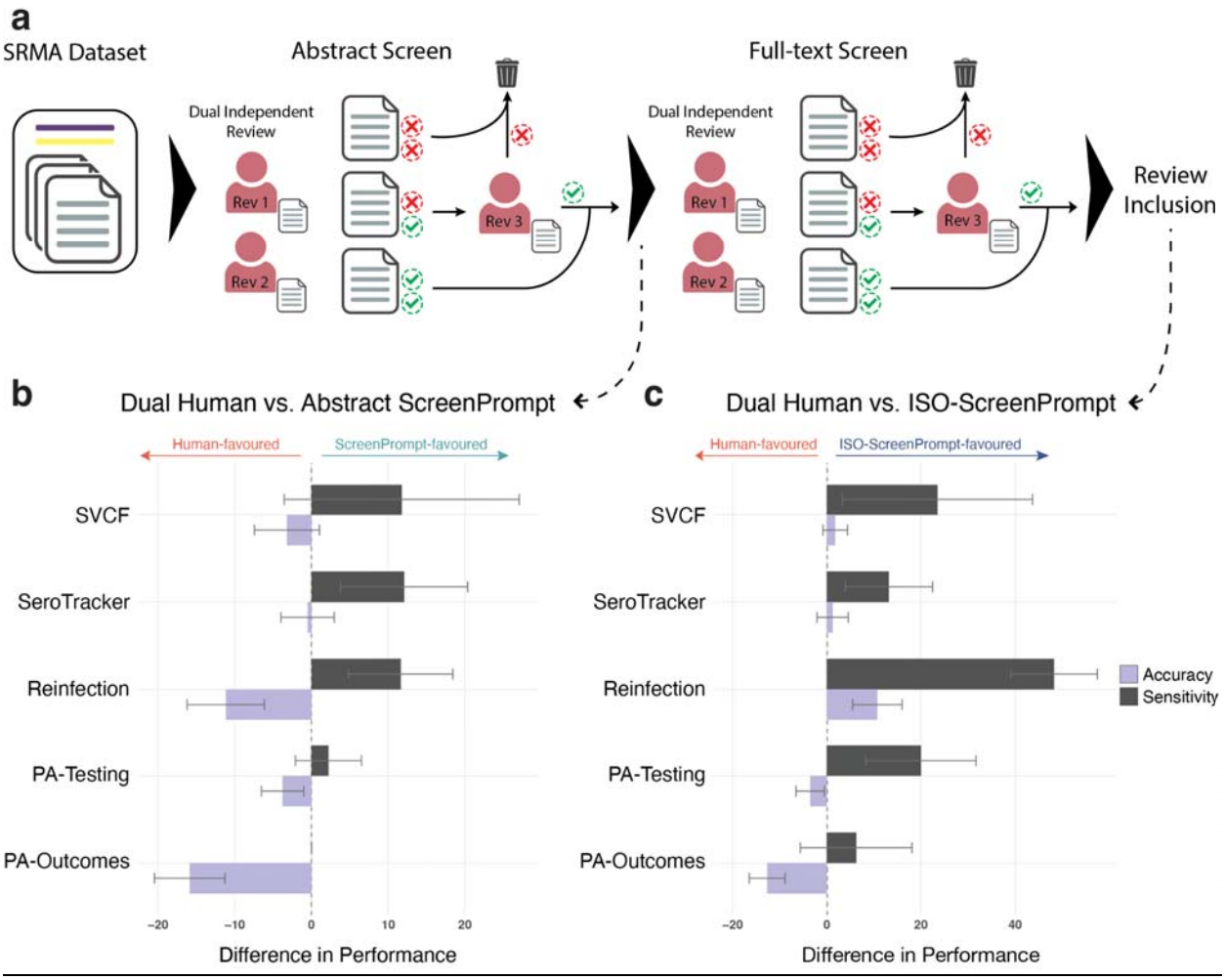


Figure 4: ScreenPrompt and ISO-ScreenPrompt perform comparably, or better than dual human review a) Overview of the dual human review process for systematic reviews, illustrating dual independent review stages for both abstract and full-text screening, leading to final review inclusion. Conflicts between reviewers are resolved by a third reviewer. b) Difference in performance between Dual Human Review and Abstract ScreenPrompt across five systematic reviews (SVCF n=167, SeroTracker n=400, Reinfection n=400, PA-Testing n=400, PA-Outcomes n=400). The barplot shows differences in accuracy and sensitivity, with human-favored differences as negative (left) and ScreenPrompt-favored differences as positive (right). Error bars represent 95% CIs for the difference in binomial proportions. c) Difference in performance between Dual Human Review and ISO-ScreenPrompt across five systematic reviews (SVCF, SeroTracker, Reinfection, PA-Testing, PA-Outcomes). The barplot shows differences in accuracy and sensitivity, with human-favored differences as negative (left) and ISO-ScreenPrompt-favored differences as positive (right). Error bars represent 95% CIs for the difference in binomial proportions.



Methods

Datasets and data acquisition

We acquired abstract screening data from 10 distinct SRs spanning eight unique clinical domains through purposeful convenience sampling based on Oxford CEBM SR question types. To compile these datasets, we engaged with systematic review investigators at the University of Calgary and the University of Toronto. We extracted study information concerning review objectives from the published manuscript or PROSPERO protocol, and contacted study authors for internal reviewer inclusion/exclusion criteria.

In brief, 'SeroTracker' (ST) was a living SR of prevalence, exploring observational cohort studies that reported single-estimate prevalence in the context of COVID-19.[43] 'Reinfection' was an SR of intervention benefits that assessed the comparative effectiveness of vaccination and past COVID-19 infection relative to past COVID-19 infection alone in observational studies reporting associations.[44] 'PA-Outcomes' was an SR of intervention benefits comparing clinical outcomes between surgery and medication treatments in patients with primary aldosteronism.[45] 'PA-Testing' was an SR of diagnostic test accuracy evaluating guideline-recommended confirmatory tests (i.e., saline infusion test, salt loading test, fludrocortisone suppression test, and captopril challenge test) relative to a reference standard.[46] 'Sepsis' was an SR of intervention benefits that assessed the comparative effectiveness and safety of fludrocortisone plus hydrocortisone, hydrocortisone alone and placebo/usual care in adults with septic shock.[47] 'Spinal' was an SR of diagnostic test accuracy that evaluated the efficacy of intraoperative neurophysiological monitoring among patients undergoing spine surgery for any indication.[48] 'Calcium-HA' was an SR of intervention benefits that compared the outcomes of routine calcium administration to no calcium administration for cardiac arrest in adults or children.[49] 'Infant-NO' was an SR of prognosis that assessed whether an immediate response to inhaled nitric oxide therapy was associated with reduced mortality in preterm infants with hypoxemic respiratory failure and pulmonary hypertension.[50] 'Meds-HA' was an SR of SRs, covering intervention benefits, that identified medications that affected hospital admissions.[51] 'SVCF' was an SR of prognosis that evaluated the association of low SVC flow, diagnosed in the first 48 hours after birth echocardiography, with neurological morbidity and mortality, among very preterm neonates.[52]

Included and excluded abstracts were downloaded from Covidence (Veritas Health Innovation, Melbourne, Australia), a systematic review screening software. Abstracts were stored in csv files. We downloaded all 'Included' (included articles at full-text screening), 'Excluded' (excluded articles at full-text screening), and 'Irrelevant' (excluded articles at abstract screening) articles. Excluded and irrelevant articles were collated to form our excluded article dataset. To obtain full-text articles, we utilized the PMC ID Converter API to convert abstract DOIs into Pubmed IDs (PMIDs), followed by the BioC API for PMC to obtain XML full-text files

from each abstract PMID. Full-texts with over 150,000 characters were excluded (i.e., abstract conference proceedings). Human authors (CC, JS, DC, RA, NB) manually scraped all 'included' full-text articles not captured by the BioC API for SRs selected in our head-to-head human-screening performance evaluation in plain text files.

The datasets described here including labeled abstract sets, and all associated metadata, are available at (link made available upon publication). Full-text articles are not provided due to copyright concerns. We invite other researchers to contribute to our growing *BenchSR*.

Abstract Prompt Approaches

During the evaluation of GPT4 abstract screening performance, we applied eight distinct prompting methodologies, discussed in more detail below. A brief illustration and the structure for each prompt is shown in Fig. 2a and Table 2.

Zero-shot prompting: Models are prompted with only instructions for completing the task at hand, including the required context directly related to the task, i.e. abstract text to analyze. No additional context or examples are provided. For our zero-shot prompts, we adapted a prompt from Guo et al.[11] that evaluated GPT4 and GPT3.5 SR screening performance. We similarly specified that the LLM returns only a single token output with its final decision.

Random few-shot prompting: Models are prompted with the instruction-task at hand, with the addition of (k=n) labeled examples relevant to the task at hand. For our purposes, the examples are randomly selected and accurately labeled as included ('YYY') or excluded ('XXX'). We set k=10 (5 included, 5 excluded), in agreement with MedPrompt and general prompting guidelines.[10]

Zero-shot Chain-of-Thought (CoT): Models are prompted with the instruction-task at hand, along with additional natural language statements, such as "*Let's think step by step*" to encourage the model to generate intermediate reasoning steps before generating a final answer.[18]

Zero-shot Framework CoT: We devised a new prompting approach, termed 'Framework CoT,' wherein we deliberately prompt the model to reason against each criterion. Similar to zero-shot prompting, no additional context or examples are provided.

Zero-shot ScreenPrompt: We included additional well-defined study objectives (adapted from published manuscripts) to our Framework CoT prompt to better orient our prompt to screening tasks.

Zero-shot Abstract ScreenPrompt: We further incorporated additional context that acknowledged the inherent content limitations of abstracts and goals of inclusivity to our ScreenPrompt prompt to better orient our prompt to the task of abstract screening.

Few-shot GPT-CoT Abstract ScreenPrompt: We prompted our model with *Abstract ScreenPrompt*, but also included additional examples that contained *Abstract ScreenPrompt* GPT-generated reasoning. We discarded answers that did not match the ground truth label to uphold 'correctness' in example reasoning. This approach was adapted from MedPrompt, which has suggested that GPT-generated CoT reasoning can outperform human experts, as

well as automate the CoT example process. We set k=10 (5 included, 5 excluded), in agreement with MedPrompt and general prompting guidelines.[10]

Self-consistency: To address variability in model outputs due to stochasticity, we conducted repeat evaluations using different seed parameters for the same prompt. The final answer is the decision to 'include' or 'exclude' according to majority vote. We set the number of self-ensembles to 11, in accordance with literature recommendations.[20] For our AUROC analysis, we adjusted the self-consistency threshold, which dictates the number of votes required for an article's inclusion or exclusion, within a range from 0 to 12 votes (0 always include; 12 always exclude). This method enables us to fine-tune sensitivity and specificity by leveraging the consensus of multiple model generations.

Full-text Prompt Approaches

During the evaluation of GPT4 full-text screening performance, we applied six distinct prompting methodologies, discussed in more detail below. A brief illustration and the structure for each prompt is shown in Fig. 3a, and Table 2. Our standard 'prompt structure' used for abstract screening consisted of: {Pre-Prompt}, {Inclusion Criteria}, {Exclusion Criteria}, {Article}, {Instructions}. Prompt elements refer to all modules except {Article}

ScreenPrompt (Init): We modify our 'prompt structure' by appending all of our prompt elements ({Objectives}, {Inclusion criteria}, {Exclusion criteria}, {Instructions}) to the start of the prompt, before the {Article} text content.

ScreenPrompt (Fin): We modify our 'prompt structure' by appending all of our prompt elements to the end of the prompt, after the {Article} text content.

ScreenPrompt (Init + Fin-Instructions): We modify our 'prompt structure' by appending all of our prompt elements to the start of the prompt, before the {Article} text content. We additionally append {Instructions} after the {Article} text content.

ScreenPrompt (Init + Fin): We modify our 'prompt structure' by appending all of our prompt elements to the start of the prompt (before the {Article} text content) and end of the prompt (after the {Article} text content).

Numbered ScreenPrompt: We preserve the semantic content of our prompt and add number symbols for each individual inclusion and exclusion sub-criterion. We removed meta-criteria headings (i.e., population, intervention, etc.) where applicable.

ISO-ScreenPrompt: We apply a combination of our Numbered ScreenPrompt where each individual inclusion/exclusion sub-criterion is numbered, and our (Init + Fin) prompt structure, where all prompt elements are appended to the start and end of the prompt.

Prompt Testing Methodology

To avoid concerns of prompting 'overfitting' during training and testing our prompt engineering process, we apply principles of sound testing methodology for machine learning studies and randomly sampled train/validation/test splits for our downstream analysis. Our iterative prompt optimization process was only performed on train splits, and we validated the performance of our optimized prompting strategy on validation splits. Test splits were held out from prompt engineering consideration until the final testing phase to assess real world, or 'eyes-off' performance.

To determine the minimum sample size for abstract evaluation, we used the Cochran's sample size formula.[53] We set our desired confidence level to 95% (p=0.05), Margin of Error (MoE) to 5%, and model performance to 50% (assuming 50% chance of inclusion/exclusion labels), resulting in a minimum sample size of 385 abstracts.

Due to its substantial dataset, we utilized data from the SeroTracker SR (n=130k excluded abstracts, n=3000 included abstracts) to derive balanced sets of included and excluded abstracts for our train, validation, and test splits (Train: n=200 included, n=200 excluded; Validation: n=200 included, n=200 excluded; Test: n=200 included, n=200 excluded;). Furthermore, we randomly sampled sets of included and excluded abstracts for our GPT-CoT prompting to prevent cross contamination (GPT-CoT: n=100 included, n=100 excluded).

Following our prompt-optimization procedures, we also tested the performance of our prompting strategy on the ST Test split, and 9 other SR abstract datasets to model its real-world performance and generalizability across different SR domains. As these datasets were held-out from any prompt engineering steps, we included all 'included' articles, and randomly sampled 'excluded' articles to reach a test sample size of 400 articles.

We replicated the same procedures for full-text evaluations. In brief, we randomly sampled the SeroTracker SR (n=5137 excluded, n=1797 included PMC-scraped full-texts) to derive balanced training and validation datasets (Train: n=200 included, n=200 excluded; Validation: n=200 included, n=200 excluded). Furthermore, we randomly sampled sets of included and excluded abstracts for our GPT-CoT prompting to prevent cross contamination (GPT-CoT: n=100 included, n=100 excluded). The test dataset was composed of all remaining articles (Test: n=1297 included, n=4637 excluded), but we opted to randomly sample a smaller subset (n=200 included, n=200 excluded) in the interest of cost and rate limits.

Following our prompt-optimization procedures, we tested the performance of our *ISO-ScreenPrompt* prompting strategy for full-text screening on the ST Test split, and 9 other SR datasets. For each SR evaluation, we set our sample size to n=400, and incorporated all included articles that were available from the PMC web scrape. For instances where we were unable to obtain a sample size of 400 (i.e., spinal), we evaluated all available 'included' and 'excluded articles. As previously mentioned, human researchers performed additional manual scraping of all 'included' articles for SRs chosen in our head-to-head evaluation of human screening performance.

Head-to-head comparisons with human screening

We assembled a panel of four researchers with past SR experience (1 BSc, 3 MSc) to perform end-to-end SR screening while adhering to canonical dual reviewer screening protocols.[1] The reviewers blinded from our study objectives (comparing GPT performance with human reviewers), and were only provided with the list of references for screening, internal inclusion/exclusion criteria, and SR objectives (Supplementary Note 1). Researchers were instructed to perform standard SR screening, with a sensitive abstract screen followed by an accurate full-text screen. Screening for each review was performed independently by two reviewers in duplicate. Any conflicts during screening were resolved by a 3rd independent reviewer. We analyzed the performance of single human-reviewer abstract screening, dual human abstract screening (based on independent votes from two reviewers and resolved conflicts at the abstract stage), and complete dual screening (based on independent full-text votes and resolved conflicts at the full-text stage). For comparisons of single human-reviewer abstract screening with Abstract ScreenPrompt, we randomly selected the performance of one reviewer from the two reviewers for each review.

We calibrated our reviewers according to screening proficiency by having prospective reviewers first screen a 'calibration set' of abstracts. This set was sourced from a prior study by the SeroTracker group,[54] which assessed the performance of dual human reviewer workflows. Notably, the SeroTracker researchers were experienced SR screeners, having contributed to the SeroTracker living SR for over a year, and represent a high-performing baseline for screening accuracy. The results from this prior study, and calibration set were used in head-to-head comparison analysis for the SeroTracker (ST) dataset. Reviewers were recruited for subsequent screening tasks if their accuracy was within 5% of the performance benchmarks set by the SeroTracker group. The performance metrics of our reviewers were highly similar to those by the SeroTracker group and are detailed in Supplementary Table 18.

For a representative comparison of GPT4-0125-preview and human screening performance, we used stratified random probability sampling across the four Oxford CEBM review questions, selecting one SR for each type. Our sample included various datasets: the SeroTracker dataset for reviews of prevalence (adapted from Perlman-Arrow et al.[54]), the Reinfection dataset for

reviews of intervention benefits, the PA-Testing dataset for reviews of diagnostic test accuracy, and SVCF dataset for reviews of prognosis. Due to the high volume of SR studies focusing on intervention benefits, we also included the PA-Outcomes dataset in our sample through additional random sampling.

Time and Cost Analysis

To derive cost estimates for traditional human dual screening for each review, we began by determining the total number of abstracts and full-texts screened at each stage. For abstracts, we used the total number of articles for a given review. For full-texts, we counted the total number of 'Included' and 'Excluded' articles from Covidence. Based on previous studies, the time required to screen a single abstract ranges from 20-461 seconds,[5,54,55] and 4.3-20 minutes for a single full-text article.[55,56] We aligned with Perlman-Arrow et al.[54] due to our use of the same ST dataset and set the screening time at 30 seconds per abstract and 10 minutes per full-text.

We calculated time estimates for human reviewers by multiplying the total number of screened abstracts by 30 seconds and the total number of screened full-texts by 10 minutes. The resulting value was then doubled to derive the cost of dual human screening (in duplicate). For single human abstract screening, we only took the total number of abstracts. We assumed a compensation rate of $20 USD per hour for human reviewers to calculate costs. It is important to note that our cost estimates do not consider the additional costs of conflict screening (additional abstracts and full-texts screened due to conflicts between reviewer decisions), and likely represents a conservative estimate of human screening costs.

For our ISO-ScreenPrompt and Abstract ScreenPrompt approach using the GPT4-0125-preview, we calculated the exact cost for each full-text and abstract run used in our head-to-head comparison analysis ($10 per million input tokens, $30 per million output tokens, OpenAI Pricing). We then derived the cost per article by dividing the total cost of our runs by the number of articles in each run. We multiplied this cost per article by the total number of articles for each review to estimate the cost of our ISO-ScreenPrompt and Abstract ScreenPrompt approach for each review. To derive estimates for our two approaches following implementation of the OpenAI Batch API, we divided the estimated cost by two (Batch API offers a 50% discount). We derived time estimates by obtaining the time per article (ISO-ScreenPrompt: 1.44s/full-text; Abstract ScreenPrompt: 0.42s/abstract), and multiplied this estimate against the total number of articles for each review.

LLM API and LLM Evaluations

We used GPT3.5-Turbo-0125 (GPT3.5), GPT4-0125-preview, GPT4-Turbo-0409, GPT4o-0513, Gemini Pro, open Mixtral-8x22-0424, and Mistral-Large-0224 models and compared their performance for abstract and full-text screening with our optimized ‘Abstract ScreenPrompt’

and 'ISO-ScreenPrompt prompts', respectively. For all models, we set the maximum tokens to 2048, and used the advised default settings. GPT model settings were set to temperature=1, top_p=1, frequency_penalty=0, presence_penalty=0. The Gemini pro model settings were set to temperature=1, top_p=1. The Mixtral-8x22-0424, and Mistral-Large-0224 model settings were set to temperature=0.7, top_p=1. We note that for Gemini Pro, the default temperature was changed to temperature=0.9 as of May 2024 (after our testing). We set the seed for all of our LLM models at 0. For our self-consistency analysis (n=11), we set different seed parameters from 0 to (n-1) for each repeat evaluation.

To evaluate the model responses, we required an output containing an 'evaluation token': either "XXX" (exclude) or "YYY" (include). When responses contained both "XXX" and "YYY", we checked the last 500 characters of the output and used the final instance of the token.

We conducted our initial runs synchronously, sending requests directly to the API and waiting for responses in the same network call. This approach allowed us to easily retry individual calls to a model's API. For each abstract, if a response was interrupted by a technical error (rate limit, disconnect, timeout) or was missing an 'evaluation token', we retried the instance up to three times. If all three attempts failed, we deemed the request as 'unparseable' and disqualified the request from the total count, occasionally reducing the set of 400 articles. During the writing of this paper, OpenAI released the Batch API, which significantly reduced costs and improved run times. Shifting to this infrastructure meant we could no longer efficiently retry individual abstracts, but the new system greatly reduced technical errors and rarely resulted in missing "evaluation tokens".

Data analysis

We assessed the performance of our prompts by analyzing accuracy, sensitivity, and specificity metrics, and calculated binomial 95% confidence intervals (CIs) for sensitivity and accuracy. "Unparseable requests" were discarded and the metrics computed only for those responses that were parseable. To compare the performance of our Abstract ScreenPrompt and ISO-ScreenPrompt with zero-shot methods across all ten reviews, we calculated 95% CIs using the Clopper-Pearson method,[57] employing the binom package in R. This approach was necessary due to the small sample size of included articles in some reviews ($n<10$). We determined statistical significance and 95% CIs with a two-tailed test computing the difference between two independent binomial proportions in our comparative analysis of human versus LLM screening performance.

DATA AVAILABILITY

Researchers can access our data via the following github repository (link made available upon publication).

CODE AVAILABILITY

All code used for experiments in this study can be found in a github repository (link made available upon publication).

References

1. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Editorial Unit, editor. Cochrane Database Syst Rev [Internet]. 2019 Oct 3 [cited 2024 Jun 1]; Available from: https://doi.wiley.com/10.1002/14651858.ED000142
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;7(2):e012545.
3. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun. 2019 Dec;16:100443.
4. Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. Syst Rev. 2012 Dec;1(1):10.
5. Chai KEK, Lines RLJ, Gucciardi DF, Ng L. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. Syst Rev. 2021 Dec;10(1):93.
6. Johnson EE, O'Keefe H, Sutton A, Marshall C. The Systematic Review Toolbox: keeping up to date with tools to support evidence synthesis. Syst Rev. 2022 Dec 1;11(1):258.
7. O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst Rev. 2019 Dec;8(1):143.
8. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst Rev. 2023 Apr 29;12(1):72.
9. Nashwan AJ, Jaradat JH. Streamlining Systematic Reviews: Harnessing Large Language Models for Quality Assessment and Risk-of-Bias Evaluation. Cureus [Internet]. 2023 Aug 6 [cited 2024 Jun 1]; Available from: https://www.cureus.com/articles/178248-streamlining-systematic-reviews-harnessing-large-language-models-for-quality-assessment-and-risk-of-bias-evaluation
10. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine [Internet]. arXiv; 2023 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2311.16452
11. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 2024 Jan 12;26:e48996.
12. Syriani E, David I, Kumar G. Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews [Internet]. arXiv; 2023 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2307.06464
13. Kohandel Gargari O, Mahmoudi MH, Hajisafarali M, Samiee R. Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. BMJ Evid-Based Med. 2024 Feb;29(1):69–70.
14. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. 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 Mar 14;jrsm.1715.
15. Burns PB, Rohrich RJ, Chung KC. The Levels of Evidence and Their Role in Evidence-Based Medicine: Plast Reconstr Surg. 2011 Jul;128(1):305–10.
16. OCEBM Levels of Evidence Working Group. The Oxford Levels of Evidence 2 [Internet]. Oxford Centre for Evidence-Based Medicine. Available from: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence
17. Clarivate. Web of Science: List of Subject Classifications for All Databases [Internet].

Available from: https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Citing-Web-of-Science-data?language=en_US

18. Kojima T, Gu S (Shane), Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in neural information processing systems [Internet]. Curran Associates, Inc.; 2022. p. 22199–213. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf
19. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners [Internet]. arXiv; 2020 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2005.14165
20. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models [Internet]. arXiv; 2022 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2203.11171
21. An S, Ma Z, Lin Z, Zheng N, Lou JG. Make Your LLM Fully Utilize the Context [Internet]. arXiv; 2024 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2404.16811
22. Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, et al. Lost in the Middle: How Language Models Use Long Contexts. Trans Assoc Comput Linguist. 2024 Feb 23;12:157–73.
23. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med Res Methodol. 2019 Dec;19(1):132.
24. Mahtani KR, Heneghan C, Aronson J. Single screening or double screening for study selection in systematic reviews? BMJ Evid-Based Med. 2020 Aug;25(4):149–50.
25. Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Syst Rev. 2019 Dec;8(1):278.
26. Matyas N, Gartlehner G, Ravaud P, Atal I. Comparing the performance of three tools for semi-automated abstract screening when conducting systematic reviews: Abstrackr, Rayyan and RobotAnalyst. In: Cochrane Colloquium Abstracts [Internet]. Available from: https://abstracts.cochrane.org/2019-santiago/comparing-performance-three-tools-semi-automated-abstract-screening-when-conducting
27. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmol. 2023 Jun 1;141(6):589.
28. Soroush A, Glicksberg BS, Zimlichman E, Barash Y, Freeman R, Charney AW, et al. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. NEJM AI [Internet]. 2024 Apr 25 [cited 2024 Jun 1];1(5). Available from: https://ai.nejm.org/doi/10.1056/AIdbp2300040
29. Han C, Kim DW, Kim S, Chan You S, Park JY, Bae S, et al. Evaluation of GPT-4 for 10-year cardiovascular risk prediction: Insights from the UK Biobank and KoGES data. iScience. 2024 Feb;27(2):109022.
30. Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. BMC Digit Health. 2024 Jan 11;2(1):4.
31. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, et al. Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. Clin Pract. 2023 Nov 20;13(6):1460–87.
32. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: A systematic review and a meta□analysis. BJOG Int J Obstet Gynaecol.

2024 Feb;131(3):378–80.
33. Nguyen MV, Luo L, Shiri F, Phung D, Li YF, Vu TT, et al. Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs [Internet]. arXiv; 2024 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2402.11199
34. Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, et al. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? [Internet]. arXiv; 2022 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2202.12837
35. Wei J, Wei J, Tay Y, Tran D, Webson A, Lu Y, et al. Larger language models do in-context learning differently [Internet]. arXiv; 2023 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2303.03846
36. Reynolds L, McDonell K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm [Internet]. arXiv; 2021 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2102.07350
37. Xiao G, Tian Y, Chen B, Han S, Lewis M. Efficient Streaming Language Models with Attention Sinks [Internet]. arXiv; 2023 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2309.17453
38. Sui Y, Zhou M, Zhou M, Han S, Zhang D. Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study [Internet]. arXiv; 2023 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2305.13062
39. Wornow M, Lozano A, Dash D, Jindal J, Mahaffey KW, Shah NH. Zero-Shot Clinical Trial Patient Matching with LLMs [Internet]. arXiv; 2024 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2402.05125
40. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2019 technology assisted reviews in empirical medicine overview. In: Conference and labs of the evaluation forum [Internet]. 2019. Available from: https://api.semanticscholar.org/CorpusID:263629854
41. Wang S, Scells H, Clark J, Koopman B, Zuccon G. From Little Things Big Things Grow: A Collection with Seed Studies for Medical Systematic Review Literature Search. 2022 [cited 2024 Jun 1]; Available from: https://arxiv.org/abs/2204.03096
42. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. J Am Med Inform Assoc. 2006 Mar 1;13(2):206–19.
43. Bobrovitz N, Arora RK, Cao C, Boucher E, Liu M, Donnici C, et al. Global seroprevalence of SARS-CoV-2 antibodies: A systematic review and meta-analysis. Khudyakov YE, editor. PLOS ONE. 2021 Jun 23;16(6):e0252617.
44. Bobrovitz N, Ware H, Ma X, Li Z, Hosseini R, Cao C, et al. Protective effectiveness of previous SARS-CoV-2 infection and hybrid immunity against the omicron variant and severe disease: a systematic review and meta-regression. Lancet Infect Dis. 2023 May;23(5):556–67.
45. Samnani S, Cenzer I, Kline GA, Lee SJ, Hundemer GL, McClurg C, et al. Time to Benefit of Surgery vs Targeted Medical Therapy for Patients With Primary Aldosteronism: A Meta-analysis. J Clin Endocrinol Metab. 2024 Feb 20;109(3):e1280–9.
46. Leung AA, Symonds CJ, Hundemer GL, Ronksley PE, Lorenzetti DL, Pasieka JL, et al. Performance of Confirmatory Tests for Diagnosing Primary Aldosteronism: a Systematic Review and Meta-Analysis. Hypertension. 2022 Aug;79(8):1835–44.
47. Teja B, Berube M, Pereira TV, Law AC, Schanock C, Pang B, et al. Effectiveness of Fludrocortisone Plus Hydrocortisone versus Hydrocortisone Alone in Septic Shock: A Systematic Review and Network Meta-Analysis of Randomized Controlled Trials. Am J Respir Crit Care Med. 2024 May 15;209(10):1219–28.
48. Alvi MA, Kwon BK, Hejrati N, Tetreault LA, Evaniew N, Skelly AC, et al. Accuracy of Intraoperative Neuromonitoring in the Diagnosis of Intraoperative Neurological Decline in the Setting of Spinal Surgery—A Systematic Review and Meta-Analysis. Glob Spine J.

2024 Mar;14(3_suppl):105S-149S.
49. Hsu CH, Couper K, Nix T, Drennan I, Reynolds J, Kleinman M, et al. Calcium during cardiac arrest: A systematic review. Resusc Plus. 2023 Jun;14:100379.
50. Baczynski M, Jasani B, De Castro C, Dani C, Subhedar NV, Chandrasekharan P, et al. Association between immediate oxygenation response and survival in preterm infants receiving rescue inhaled nitric oxide therapy for hypoxemia from pulmonary hypertension: A systematic review and meta-analysis. Early Hum Dev. 2023 Sep;184:105841.
51. Bobrovitz N, Heneghan C, Onakpoya I, Fletcher B, Collins D, Tompson A, et al. Medications that reduce emergency hospital admissions: an overview of systematic reviews and prioritisation of treatments. BMC Med. 2018 Dec;16(1):115.
52. Mascarenhas D, Weisz D, Jasani B, Persad N, Main E. Premedication for rapid sequence intubation in neonates - a network meta-analysis. PROSPERO 2022 CRD42022384259 [Internet]. PROSPERO. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022384259
53. Cochran WG. Sampling techniques, 3rd edition. John Wiley; 2002.
54. Perlman□Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real□world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods. 2023 Jul;14(4):608–21.
55. Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjo M, et al. Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol. 2021 Nov;139:287–96.
56. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large□evidence systematic reviews and meta□analyses. Res Synth Methods. 2019 Sep;10(3):330–42.
57. Clopper CJ, Pearson ES. THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL. Biometrika. 1934;26(4):404–13.

Additional Information

Acknowledgements

We would like to thank Dr. Michelle Baczynski for her support in providing access to the Infant-NO dataset. We would like to thank Dr. Mohammed Ali Alvi for his support in providing access to the Spinal dataset.

Author Information

These authors contributed equally: C.C., J.S. These authors jointly supervised the work: R.K.A., N.B.

Author contribution statements

C.C., J.S., N.B., R.K.A contributed to the conception and design of the work. C.C., J.S., R.A., R.K., M.C., J.G., R.S., D.C., I.D., B.T., M.F., P.R., A.A.L., D.E.W., H.W., M.W., N.B., contributed to the data acquisition and curation. C.C., J.S., contributed to original investigation and formal analysis. C.C., J.S., R.A., D.E.W., N.B., R.K.A., interpreted results and provided feedback on the study. C.C., J.S., R.A., D.E.W., contributed to validation of the data. C.C., J.S., H.W., M.W., N.B., R.K.A., contributed to project supervision. C.C., J.S., R.A., N.B., prepared the original draft of the manuscript with input form all co-authors. All authors were responsible for review and editing of the manuscript. All authors debated, discussed, edited, and approved the final version of the manuscript.

Ethics Declaration

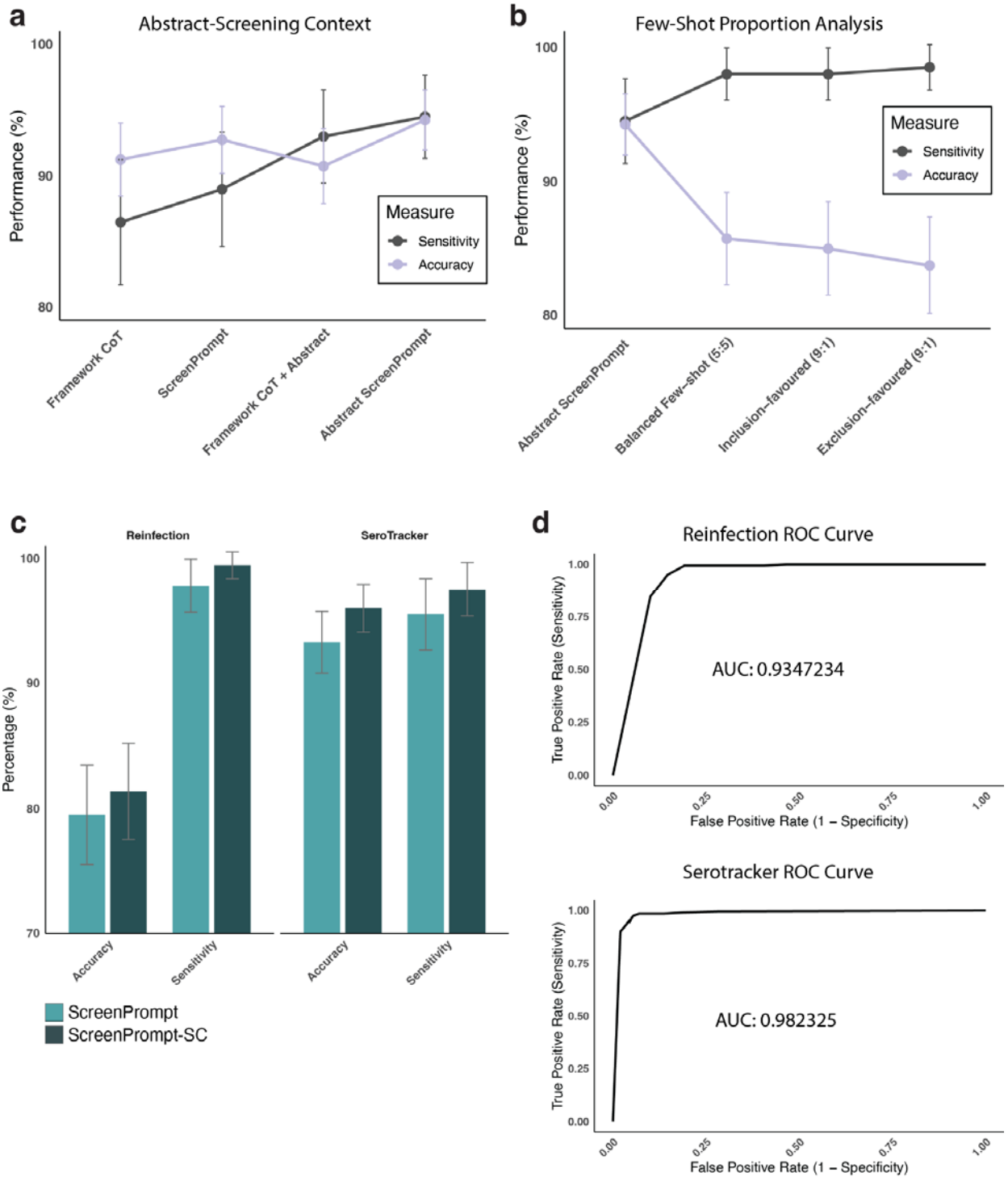
Competing Interests

C.C., H.W., M.W., R.K.A., and N.B., report grants from the Public Health Agency of Canada through Canada's COVID-19 Immunity Task Force, the World Health Organization Health Emergencies Programme, the Robert Koch Institute, and the Canadian Medical Association Joule Innovation Fund. No funding source had any role in the design of this study, its execution, analyses, interpretation of the data, or decision to submit results. This manuscript does not necessarily reflect the views of the World Health Organization or any other funder.

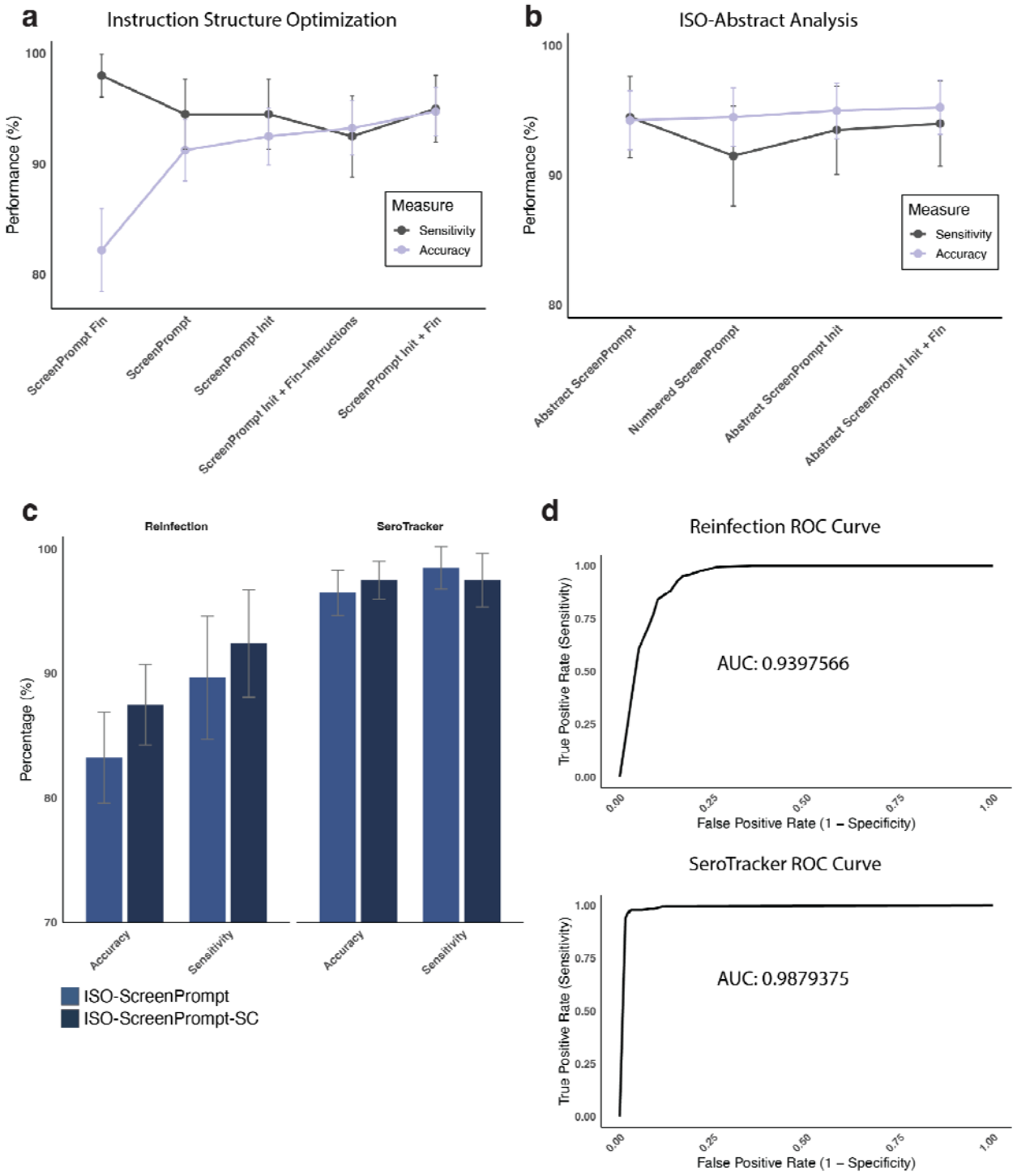
R.K.A. is employed at OpenAI and may own stock as part of the standard compensation package. R.K.A., was also previously a Venture Fellow at Flagship Pioneering, minority shareholder of Alethea Medical, and has received funding from the Rhodes Trust and Open Philanthropy.

Supplementary Information

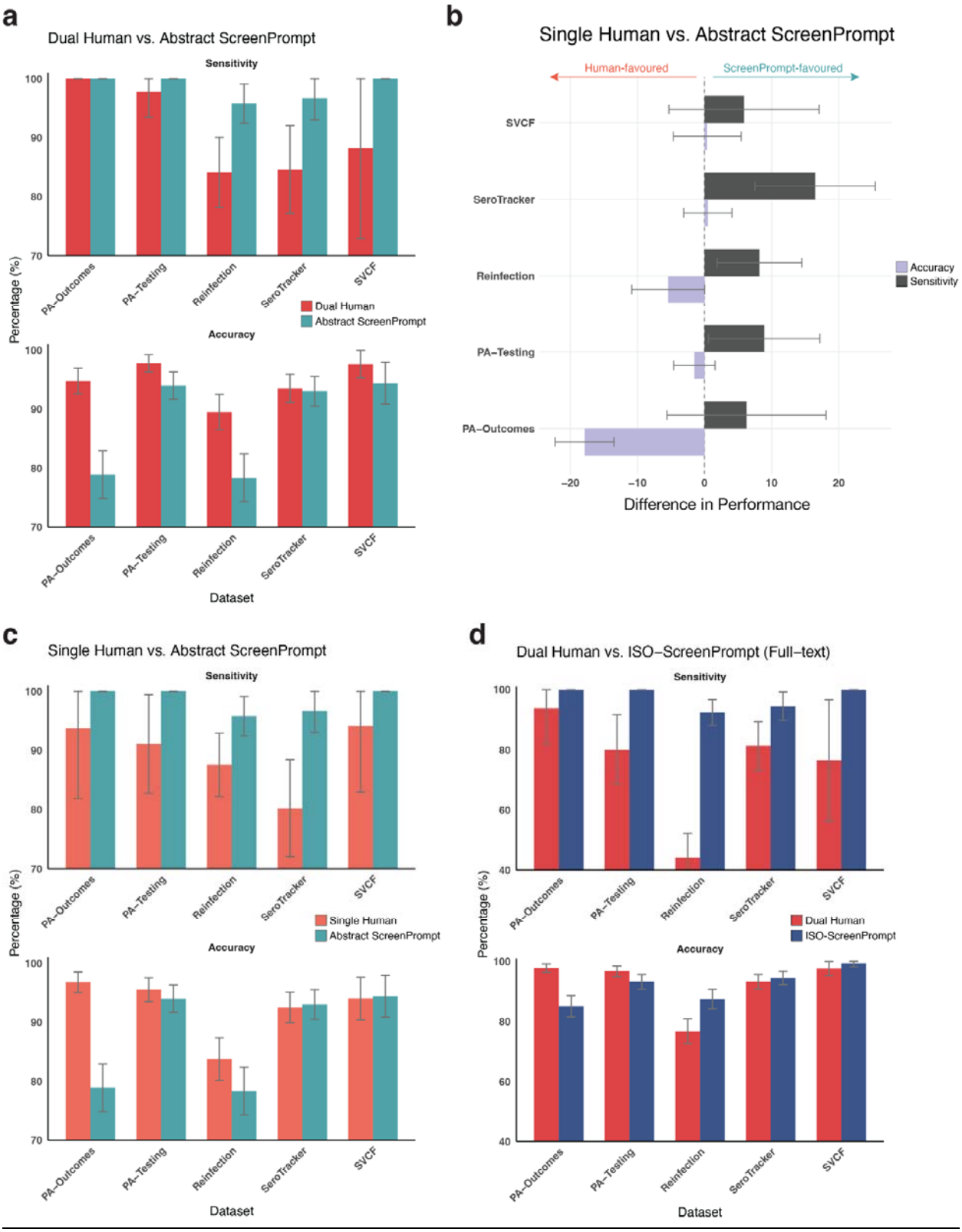
Supplementary Figure 1: Abstract screening prompt optimization a) Performance comparison of different abstract prompting methodologies relating to the addition of study objectives and abstract-specific considerations on the SeroTracker (ST) training dataset, (n=400) showing accuracy and sensitivity. Error bars represent 95% CIs for binomial proportions. b) Performance comparisons of different few-shot prompting methodologies on the ST training dataset, showing accuracy and sensitivity. Prompts compare differing proportions of inclusion-labeled and exclusion-labeled few-shot examples. Error bars represent 95% CIs for binomial proportions. c) Barplot displaying the performance of Abstract ScreenPrompt sensitivity and accuracy with and without self-consistency (SC) in SeroTracker and Reinfection test datasets (n=400). Error bars represent 95% CIs for binomial proportions. d) Receiver operating characteristics (ROC) curves generated from differing self-consistency thresholds (number of votes needed, 0-12) for article inclusion in SeroTracker and Reinfection test datasets.



Supplementary Figure 2: Full-text screening prompt optimization. a) Performance comparison of different full-text prompting methodologies with modifications in prompt structure on the ST training dataset (n=400), showing accuracy and sensitivity. Error bars represent 95% CIs for binomial proportions. b) Performance comparison of abstract screening performance with modifications in prompt structure on the ST training dataset, showing accuracy and sensitivity. Error bars represent 95% CIs for binomial proportions. c) Barplot displaying the performance of ISO-ScreenPrompt sensitivity and accuracy with and without self-consistency (SC) in SeroTracker and Reinfection test datasets (n=400). Error bars represent 95% CIs for binomial proportions. d) Receiver operating characteristics (ROC) curves generated from differing self-consistency thresholds (number of votes needed, 0-12) for article inclusion in SeroTracker and Reinfection test datasets.



Supplementary Figure 3: ScreenPrompt and ISO-ScreenPrompt vs. human screening. a) Barplot displaying the performance of Abstract ScreenPrompt and dual human abstract screening sensitivity and accuracy. Error bars represent 95% CIs for binomial proportions. b) Difference in performance between single human-reviewer and Abstract ScreenPrompt across five systematic reviews (SVCF n=167, SeroTracker n=400, Reinfection n=400, PA-Testing n=400, PA-Outcomes n=400). The barplot shows differences in accuracy and sensitivity, with human-favored differences as negative (left) and ScreenPrompt-favored differences as positive (right). Error bars represent 95% CIs for the difference in binomial proportions. c) Barplot displaying the performance of Abstract ScreenPrompt and single-human reviewer screening sensitivity and accuracy. Error bars represent 95% CIs for binomial proportions. d) Barplot displaying the performance of ISO-ScreenPrompt and full dual human screening sensitivity and accuracy. Error bars represent 95% CIs for binomial proportions.



Supplementary Table 1 - Abstract prompt engineering performance

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Zero-shot -	65.0% 260/400	30.0% 60/200	**100.0%** **200/200**
	Random Few-shot Balanced: 5 include 5 exclude	78.0% 312/400	56.0% 112/200	100.0% 200/200
	Zero-shot Chain of Thought (CoT)	86.3% 345/400	74.5% 149/200	98.0% 196/200
	Zero-shot framework CoT	91.3% 365/400	86.5% 173/200	96.0% 192/200
	ScreenPrompt	92.8% 371/400	89.0% 178/200	96.5% 193/200
	Abstract ScreenPrompt	**94.3%** **377/400**	94.5% 189/200	94.0% 188/200
	Few-shot GPT-CoT ScreenPrompt Balanced: 5 include 5 exclude	85.8% 343/400	**98.0%** **196/200**	73.5% 147/200
Validation (n=200 include, n=200 exclude)	Abstract ScreenPrompt	94.3% 377/400	96.0% 192/200	92.5% 185/200

Supplementary Table 2: Inclusive Language Prompting

Dataset	Zero-shot prompt	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	... Only type “YYY” for included articles or “XXX” for excluded articles to indicate your decision. Do not type anything else.	65.0% 260/400	30.0% 60/200	**100.0%** **200/200**
	... We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.	68.3% 273/400	36.5% 73/200	100.0% 200/200

Supplementary Table 3: Exemplar Wrong Outputs during Abstract Prompt Engineering

Prompt	Example LLM Output	Author Feedback
Zero-shot CoT:	{Model output} **Population**: 1. The study involves humans of any age. [Meets criteria] 2. The subjects include those with suspected disease without requiring them to have been deliberately sampled for being antigen positive. [Meets criteria] **Study Design**: 1. This is a cross-sectional serologic survey. It fits the inclusion criteria for study design as it is one of the specified designs interested in measuring seroprevalence. [Meets criteria] 2. Serology measurements were presumably taken at a single time point as common in cross-sectional designs. [Meets criteria] **Special Design**: 1. This paper is not a systematic review or meta-analysis but rather a primary seroprevalence study. However, non-systematic reviews/meta-analyses are not listed as exempt from consideration; the inclusion criteria express a specific interest in systematic reviews and meta-analyses without excluding primary studies. [Meets criteria] **Sampling**: 1. The method of sampling is not entirely detailed in the abstract, but given that it's a cross-sectional study among a defined number of participants (888), it suggests a deliberate approach to sampling, which is compliant with the inclusive approach to any sampling method. [Meets criteria] **Types of Evidence**: 1. Assuming this is a published academic article (given its detailed results and conclusions, along with the referencing of copyright by Wiley & Sons Ltd.), it meets the criteria for the types of evidence being sought. [Meets criteria] **Outcome Measures**: 1. The study reports a seroprevalence estimate, with 30% of the study volunteers having detected neutralizing antibodies. [Meets criteria] 2. The number of participants enrolled in the study is clearly stated as 888. [Meets criteria] 3. While the study sampling end date/week is not explicitly mentioned, this detail might be present in the full text; however, for abstract screening, its absence is noted but its potential availability in the full report could still mean the study might meet the criteria upon full review. [Unclear] 4. The study clearly reports the location as being in Egypt, fitting the requirement to categorize the location of the study. [Meets criteria] Given that all but one of the clear inclusion criteria are met with only the study sampling end date/week not explicitly stated in the abstract, there's a stronger leaning towards inclusion, with the assumption that further details necessary to fully satisfy all criteria might be found in the full text. **Conclusion: YYY**	Output did not assess/evaluate exclusion criteria → Led to Framework CoT
Zero-Shot Framework CoT	Let's evaluate the abstract provided against the inclusion and exclusion criteria systematically. ### Inclusion Criteria Evaluation:	Erroneously inferred study objectives

	1. **Population:** The study includes humans of any age, which meets the criterion. It doesn't specify excluding only those with suspected or confirmed COVID-19, suggesting broader inclusion, thus meeting this requirement. 2. **Study design:** The study design focuses on the qualitative assay of immunoglobulins G and M in diagnosed patients, aiming to assess transmissibility and the possibility of reinfection. This can be interpreted as part of a broader sero-survey effort, but it appears more focused on specific diagnostic outcomes rather than prevalence estimation. It lacks a clear mention of being a cross-sectional, repeated cross-sectional, or cohort study as defined by our criteria. 3. **Special design:** The study does not present itself as a systematic review or meta-analysis of seroprevalence studies. 4. **Sampling:** The method of sampling is not explicitly described, but the study included a total of 553 patients, indicating a form of sampling was done. 5. **Types of evidence:** Presumably, this is published or unpublished academic literature, considering it has authorship and an abstract, thus meeting this criterion. 6. **Outcome measures:** The abstract reports on IgG and IgM levels in patients, implying an assessment of seroprevalence to some extent. However, it does not clearly state it reports a seroprevalence estimate in terms of the proportion of the population with detectable antibodies. The number of participants is reported, but it lacks clear information on study sampling end date/week or detailed location categorization. ### Exclusion Criteria Evaluation: 1. **Population:** This study is conducted in humans, thus not meeting any of the population-based exclusion criteria directly. 2. **Study design:** The study could arguably fall under the exclusion due to its focus on evaluating immunological responses rather than explicitly being a cross-sectional or cohort design aimed at seroprevalence estimation. 3. **Types of evidence:** There's no mention of using excluded multimedia sources or unassociated dashboards. 4. **Outcome measures:** The abstract mainly discusses the detection of IgG and IgM antibodies but does not only report on antigen prevalence. However, it does not fully comply with reporting requirements around the study sampling end date/week or location detail. ### Conclusion: **Given the specific focus on evaluating immunological responses in patients rather than estimating the seroprevalence across a population,** and the absence of clear compliance with several crucial inclusion criteria (study design specifics, complete outcome measure reporting), the study appears to lean more towards exclusion based on its design and reported outcomes. However, considering the nature of abstract reviews, the decision leans on available information which might not fully capture the study's relevance to the systematic review aims. Given the potential relevance of immunological response data to the broader topic of seroprevalence and immunity estimation, there is a fine line of interpretation. Considering the primary focus on diagnostic outcomes as opposed to population-level antibody prevalence estimation and the critical unfulfilled aspects of both inclusion and exclusion criteria, the conclusion leans towards: XXX	→ Led to ++ Objectives

Supplementary Table 4: Abstract ScreenPrompt Ablation Prompting

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Zero-shot framework CoT	91.3% 365/400	86.5% 173/200	96.0% 192/200
	Zero-shot framework CoT ++ Study Objectives (ScreenPrompt)	92.8% 371/400	89.0% 178/200	**96.5%** **193/200**
	Zero-shot framework CoT ++ Abstract Consideration	90.8% 363/400	93.0% 186/200	88.5% 177/200
	Zero-shot framework CoT ++ Abstract Consideration ++ Study Objectives (Abstract ScreenPrompt)	**94.3%** **377/400**	**94.5%** **189/200**	94.0% 188/200

Supplementary Table 5: Few-shot GPT-CoT Label Proportion Analysis

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Few-shot GPT-CoT ScreenPrompt Balanced: 5 include 5 exclude	**85.8%** **343/400**	98.0% 196/200	**73.5%** **147/200**
	Few-shot GPT-CoT ScreenPrompt Inclusion-favored: 9 include 1 exclude	85.0% 340/400	98.0% 196/200	72.0% 144/200
	Few-shot GPT-CoT ScreenPrompt Exclusion-favored: 1 include 9 exclude	83.8% 335/400	**98.5%** **197/200**	69.0% 138/200

Supplementary Table 6: Comparative analysis of Abstract ScreenPrompt across LLM models

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity	Cost (CAD)
ST Training n=200 include n=200 exclude	GPT4-0125-preview Abstract ScreenPrompt	**94.3%** **377/400**	**94.5%** **189/200**	94.0% 188/200	$12.54
	GPT4-Turbo-0409 Abstract ScreenPrompt	89.8% 359/400	83.5% 167/200	96.0% 192/200	$12.16
	GPT3.5 Abstract ScreenPrompt	66.7% 266/399	90.5% 180/199	43.0% 86/200	$0.37
	Gemini Pro Abstract ScreenPrompt	76.2% 301/395	67.5% 133/197	84.8% 168/198	**Free**
	Mistral-Large Abstract ScreenPrompt	89.3% 357/400	93.5% 187/200	85.0% 170/200	$2.65
	Mixtral-8x22 Abstract ScreenPrompt	84.5% 337/399	94.0% 188/200	74.9% 149/199	$4.15
	GPT4o-0513 Abstract ScreenPrompt	89.0% 356/400	79.5% 159/200	**98.5%** **197/200**	$7.37

Supplementary Table 7 - Zero-shot vs. Abstract ScreenPrompt across LLM Models

Dataset	Model	Prompt	Accuracy	Sensitivity	Specificity
Training n=200 include n=200 exclude	GPT4-0125-preview	Abstract ScreenPrompt	**94.3%** **377/400**	**94.5%** **189/200**	94.0% 188/200
		Zero-shot	65.0% 260/400	30.0% 60/200	**100.0%** **200/200**
	GPT4-Turbo-0409	Abstract ScreenPrompt	**89.8%** **359/400**	**83.5%** **167/200**	96.0% 192/200
		Zero-shot	56.8% 227/400	13.5% 27/200	**100.0%** **200/200**
	GPT3.5	Abstract ScreenPrompt	**66.7%** **266/399**	90.5% 180/199	**43.0%** **86/200**
		Zero-shot	64.5% 258/400	**97.0%** **193/200**	32.5% 65/200

Supplementary Table 8 - Generalizability of Abstract ScreenPrompt across SRs

Dataset	# of Abstracts n = total (include/ exclude)	Prompt	Accuracy	Sensitivity	Specificity
SeroTracker (Test Dataset)	400 (200/200)	Zero-shot	71.5% 286/400	43.0% 86/200	100.0% 200/200
		Abstract ScreenPrompt	93.3% 373/400	95.5% 191/200	91.0% 182/200
Reinfection	400 (181/219)	Zero-shot	86.9% 345/397	87.6% 156/178	86.3% 189/219
		Abstract ScreenPrompt	79.5% 318/400	97.8% 177/181	64.4% 141/219
PA-Testing	400 (52/348)	Zero-shot	95.0% 380/400	80.8% 42/52	97.1% 338/348
		Abstract ScreenPrompt	91.3% 365/400	98.1% 51/52	90.2% 314/348
PA-Outcomes	400 (16/384)	Zero-shot	94.5% 378/400	87.5% 14/16	94.8% 364/384
		Abstract ScreenPrompt	73.5% 294/400	100.0% 16/16	72.4% 278/384
Meds-HA	400 (140/260)	Zero-shot	71.8% 287/400	20.0% 28/140	99.6% 259/260
		Abstract ScreenPrompt	90.3% 361/400	93.6% 131/140	88.5% 230/260

Sepsis	400 (17/383)	Zero-shot	97.8% 391/400	52.9% 9/17	99.7% 382/383
		Abstract ScreenPrompt	92.5% 370/400	100.0% 17/17	92.2% 353/383
Spinal	400 (135/265)	Zero-shot	88.5% 354/400	70.4% 95/135	97.7% 259/265
		Abstract ScreenPrompt	87.5% 350/400	100.0% 135/135	81.1% 215/265
Infant-NO	400 (6/394)	Zero-shot	98.0% 392/400	16.7% 1/6	99.2% 391/394
		Abstract ScreenPrompt	90.8% 363/400	100.0% 6/6	90.6% 357/394
Calcium-HA	400 (15/385)	Zero-shot	97.5% 390/400	40.0% 6/15	99.7% 384/385
		Abstract ScreenPrompt	95.5% 382/400	86.7% 13/15	95.8% 369/385
SVCF	400 (17/383)	Zero-shot	97.3% 389/400	35.3% 6/17	100.0% 383/383
		Abstract ScreenPrompt	91.3% 365/400	100.0% 17/17	90.9% 348/383

Supplementary Table 9 - Abstract ScreenPrompt Self-Consistency Analysis

Dataset	Method	Accuracy	Sensitivity	Specificity	Cost (USD)
SeroTracker	Abstract ScreenPrompt	93.3% 373/400	95.5% 191/200	91.0% 182/200	**$12.56**
	Abstract ScreenPrompt + Self-Consistency	**95.9%** **383/399**	**97.5%** **195/200**	**94.5%** **188/199**	$88.02 Batch API
Reinfection	Abstract ScreenPrompt	79.5% 318/400	97.8% 177/181	64.4% 141/219	**$14.77**
	Abstract ScreenPrompt + Self-Consistency	**81.4%** **323/397**	**99.4%** **180/181**	**66.2%** **143/216**	$89.18 Batch API

Supplementary Table 10 - Full-text prompt engineering performance

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Zero-shot	68.9% 275/399	37.7% 75/199	100.0% 200/200
	Abstract ScreenPrompt	85.5% 342/400	**98.0%** **196/200**	73.0% 146/200
	ScreenPrompt	91.3% 365/400	94.5% 189/200	88.0% 176/200
	ScreenPrompt Init + Fin	94.8% 379/400	95.0% 190/200	94.5% 189/200
	Numbered ScreenPrompt	95.3% 381/400	96.0% 192/200	94.5% 189/200
	ISO-ScreenPrompt	**95.5%** **382/400**	93.5% 187/200	**97.5%** **195/200**
Validation (n=200 include, n=200 exclude)	ISO-ScreenPrompt	96.3% 385/400	97.5% 195/200	95% 190/200

Supplementary Table 11: Full-text Prompt Structure Testing

Dataset	Prompting Strategy (Structure)	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	ScreenPrompt	91.3% 365/400	94.5% 189/200	88.0% 176/200
	ScreenPrompt Init	92.5% 370/400	94.5% 189/200	90.5% 181/200
	ScreenPrompt Fin	82.2% 328/399	98.0% 196/200	66.3% 132/199
	ScreenPrompt Init + Fin-Instructions	93.3% 373/400	92.5% 185/200	94.0% 188/200
	ScreenPrompt Init + Fin	94.8% 379/400	95.0% 190/200	94.5% 189/200

Supplementary Table 12: Exemplar Wrong Outputs during Full-text Prompt Engineering

Prompt	Example LLM Output	Author feedback
ScreenPrompt Init	XXX	Single token output → Led to experiments testing adding instructions and whole prompts at the bottom of the prompt
ScreenPrompt Init + Fin-Instructions	### Decision Process for Inclusion or Exclusion To assess the paper's suitability for inclusion in the systematic review, we must methodically examine it against each criterion listed under the inclusion and exclusion categories. This approach ensures that the decision is objective, based on predefined standards and takes the study's characteristics into account. Let's delve into each criterion, evaluating how the paper stands against them. ### Inclusion Criteria 1. **Population:** The study focuses on humans of any age across districts in India, which meets the first inclusion criterion. 2. **Study design:** The paper appears to employ a cross-sectional analysis within a defined population over a certain period, aiming to correlate BMI with the severity and spatial variation of COVID-19 in India. This aligns with the inclusion criterion related to study designs accepted. 3. **Special design:** While not a systematic review or meta-analysis itself, the paper's systematic approach to evidence synthesis can indirectly contribute to understanding seroprevalence studies' landscape. 4. **Sampling:** The paper does not detail its sampling method distinctly, but its analysis based on district-level data suggests an organized systematic approach, likely consistent with acceptable sampling methods for inclusion. 5. **Types of evidence:** The manuscript is an academic literature piece. Even though it does not mention its publication status, given the context, it's assumed to meet the inclusion criteria related to evidence types. **6. **Outcome measures:** The paper reports on several outcome measures, including confirmed cases and severity indicators of COVID-19, which indirectly inform on the pandemic's spread, though not seroprevalence estimates directly.** ### Exclusion Criteria 1. **Population:** The study solely includes human subjects, thus not meeting any exclusion criteria related	Inclusion/exclusion criteria evaluated by meta-criteria headings, reducing important granularity. For example, seroprevalence was not reported, but the LLM inferred that outcome measures were correctly reported, thus deeming the criteria (**incorrectly**) satisfied → Led us to enhance our instruction format and model outputs with 'Numbered framework CoT'

	to the population. 2. **Study design:** The paper embodies a cross-sectional study design, thus not falling under any excluded designs like case reports or evaluations of serological tests. 3. **Types of evidence:** The manuscript does not rely on multimedia, dashboards unrelated to a defined serology study, or unattributable slide deck presentations, thereby not meeting the exclusion criteria regarding evidence types. 4. **Outcome measures:** Although the paper does not directly report seroprevalence estimates, its focus on COVID-19's intensity and spatial variation based on BMI associations indirectly contributes to understanding the pandemic's spread. Hence, it doesn't directly fulfill any of the specific outcome measure exclusion criteria. ### Conclusion Considering the analysis above, the paper does not fit neatly into the exclusion criteria and aligns with several critical inclusion criteria, notably around population, study design, and types of evidence. Despite its focus not being directly on seroprevalence estimates, its insight into the pandemic's spread through indirect measures like BMI's impact offers valuable, albeit tangential, contributions to seroprevalence discussions. The decision, thus, veers towards inclusion based on the information provided and the indirect relevance of its findings to understanding aspects of the pandemic's spread, which is of interest in the broader context of seroprevalence studies. YYY	
Numbered ScreenPrompt	### Decision-Making Process To decide whether to include or exclude the abstract, we'll evaluate it against the inclusion and exclusion criteria systematically. ### Inclusion Criteria Assessment 1. **Humans of any age.** [Met] - The study discusses the impacts of COVID-19 policies on the Swedish population, implying human subjects of unspecified ages. 2. **Including COVID-19 antigen positive persons and those with suspected disease if not deliberately sampled.** [Not Met] - The abstract focuses on policy outcomes and pandemic metrics rather than sampling individuals based on antigen positivity or suspected disease. 3. **Sero-surveys.** [Not Met] - This study does not mention sero-surveys or the collection and testing of serum to estimate SARS-CoV-2 antibody prevalence. 4. **Study Designs.** [Not Met] - The study follows a longitudinal observational design, evaluating pandemic variables over time but does not fit	Model now correctly evaluates each individual inclusion/exclusion sub-criterion

	the specified designs focused on serology measurements. 5. **Include systematic reviews and meta-analysis of seroprevalence studies.** [Not Met] - The study does not review or meta-analyze seroprevalence studies; it examines policy outcomes and pandemic data. 6. **Any sampling method.** [Not Met] - Since it doesn't involve direct sero-surveys, it doesn't specify a sampling method for such purposes. 7. **Literature type.** [Met] - The abstract seems part of academic literature, which is eligible. 8. **Reports a seroprevalence estimate.** [Not Met] - No seroprevalence or proportion of the population with antibodies is reported or estimated. 9. **Reports the number of participants enrolled (denominator).** [Not Met] - The study does not involve enrolling participants in the traditional sense nor reports such a denominator. 10. **Reports study sampling end date/week.** [Met] - Provides detailed timelines regarding the pandemic data collection period. 11. **Reports the locations at which the study takes place.** [Met] - The study clearly takes place in Sweden, satisfying the geographical reporting requirement. ### Exclusion Criteria Assessment 1. **Non-human studies.** [Not Met] - The study involves human data. 2. **Study only included individuals with confirmed COVID-19 cases.** [Not Applicable] - Does not exclusively focus on individuals with confirmed or suspected COVID-19 cases in a clinical assessment sense but evaluates general pandemic outcomes. 3. **Study only included vaccinated individuals.** [Not Applicable] - Vaccination status is not the focus. 4. **Study designs other than specified.** [Met] - The study does not adhere to the specified cross-sectional, cohort, or serology measurement designs. 5. **Multimedia sources or unverifiable slides.** [Not Met] - The abstract is presented as part of academic literature. 6. **Dashboards not associated with a defined serology study.** [Not Applicable] - The study is not about a dashboard per se but uses data from national health registries. 7. **Only reports incidence or prevalence of antigen.** [Not Met] - The study reports on broader pandemic variables without focusing solely on antigen incidence or prevalence. 8. **Does not report sampling end date/week.** [Not	

	Met] - It reports the period of data collection. 9. **Does not report the number of participants.** [Met] - As it is not a sero-survey-based study, it does not report on specific participant numbers in that context. 10. **Does not report the location.** [Not Met] - Location is reported. ### Conclusion The paper is predominantly focused on evaluating the effects of policy measures and their outcomes in terms of general pandemic variables over time in Sweden. It lacks a direct focus on sero-surveys, seroprevalence estimates regarding SARS-CoV-2 antibodies, and does not fit the required study design for seroprevalence studies as dictated by the inclusion criteria. Therefore, the paper does not meet the inclusion criteria essential for the systematic review focused on SARS-CoV-2 antibodies prevalence, variations, and underestimation of pandemic spread through acute infection surveillance. Consequently, based on the systematic evaluation: XXX	

Supplementary Table 13: ISO-Prompting in Abstract Screening

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Abstract ScreenPrompt (Regular structure)	94.3% 377/400	94.5% 189/200	94.0% 188/200
	Numbered Abstract ScreenPrompt	94.5% 378/400	91.5% 183/200	97.5% 195/200
	Abstract ScreenPrompt Init	95.0% 380/400	93.5% 187/200	96.5% 193/200
	Abstract ScreenPrompt Init + Fin	95.3% 381/400	94.0% 188/200	96.5% 193/200

Supplementary Table 14 - Comparative analysis of ISO-ScreenPrompt across LLM models

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity	Cost (CAD)
Training (n=200 include, n=200 exclude)	GPT4-0125-preview ISO-ScreenPrompt	95.5% 382/400	**93.5%** **187/200**	97.5% 195/200	$51.90
	GPT4-Turbo-0409 ISO-ScreenPrompt	**95.8%** **383/400**	93.0% 186/200	**98.5%** **197/200**	$51.06
	GPT3.5 ISO-ScreenPrompt	76.1% 251/330	74.1% 129/174	78.2% 122/156	$1.76
	Gemini Pro ISO-ScreenPrompt	67.3% 255/379	69.1% 134/194	65.4% 121/185	**Free**
	Mistral-Large ISO-ScreenPrompt	86.9% 345/397	77.8% 154/198	96% 191/199	$19.77
	Mixtral-8x22 ISO-ScreenPrompt	93.7% 370/395	91.9% 182/198	95.4% 188/197	$9.93
	GPT4o-0513 ISO-ScreenPrompt	95.3% 381/400	93.0% 186/200	97.5% 195/200	$27.12

Supplementary Table 15 - Zero-shot vs. ISO-ScreenPrompt across LLM Models

Dataset	Model	Prompt	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	GPT4-0125-preview	**ISO-ScreenPrompt**	**95.5%** **382/400**	**93.5%** **187/200**	97.5% 195/200
		Zero-shot	68.9% 275/399	37.7% 75/199	**100.0%** **200/200**
	GPT4-Turbo-0409	**ISO-ScreenPrompt**	**95.8%** **383/400**	**93.0%** **186/200**	98.5% 197/200
		Zero-shot	56.8% 227/400	13.5% 27/200	**100.0%** **200/200**
	GPT3.5	**ISO-ScreenPrompt**	76.1% 251/330	74.1% 129/174	**78.2%** **122/156**
		Zero-shot	72.1% 253/351	**92.8%** **168/181**	50.0% 85/170

Supplementary Table 16 - Generalizability of ISO-ScreenPrompt across SRs

Dataset	# of full-text n = total (include/ exclude)	Prompt	Accuracy	Sensitivity	Specificity
SeroTracker SR (Test Dataset)	400 (200/200)	Zero-shot	73.9% 294/398	48.0% 95/198	99.5% 199/200
		ISO-ScreenPrompt	96.5% 384/398	98.5% 195/198	94.5% 189/200
Reinfection SR	400 (141/255)	Zero-shot	85.0% 340/400	71.0% 103/145	92.9% 237/255
		ISO-ScreenPrompt	83.3% 333/400	89.7% 130/145	79.6% 203/255
PA-Testing SR	400 (45/355)	Zero-shot	97.3% 389/400	88.9% 40/45	98.3% 349/355
		ISO-ScreenPrompt	93.2% 372/399	100.0% 45/45	92.4% 327/354
PA-Outcomes	400 (16/384)	Zero-shot	90.8% 363/400	93.8% 15/16	90.6% 348/384
		ISO-ScreenPrompt	84.3% 337/400	100.0% 16/16	83.6% 321/384
Meds-HA SR	218 (17/200)	Zero-shot	95.5% 382/400	11.8% 2/17	99.2% 380/383
		ISO-ScreenPrompt	95.8% 383/400	100.0% 17/17	95.6% 366/383

Sepsis SR	400 (16/384)	Zero-shot	99.0% 383/387	75.0% 12/16	100.0% 371/371
		ISO-ScreenPrompt	95.3% 369/387	93.8% 15/16	95.4% 354/371
Spinal SR	244 (13/231)	Zero-shot	96.7% 236/244	84.6% 11/13	97.4% 225/231
		ISO-ScreenPrompt	90.2% 220/244	100.0% 13/13	89.6% 207/231
Infant-NO	197 (6/191)	Zero-shot	98.5% 194/197	66.7% 4/6	99.5% 190/191
		ISO-ScreenPrompt	98.0% 193/197	100.0% 6/6	97.9% 187/191
Calcium-HA	238 (14/224)	Zero-shot	97.5% 232/238	57.1% 8/14	100.0% 224/224
		ISO-ScreenPrompt	99.6% 237/238	92.9% 13/14	100.0% 224/224
SCVF	167 (17/150)	Zero-shot	95.8% 160/167	58.8% 10/17	100.0% 150/150
		ISO-ScreenPrompt	99.4% 166/167	100.0% 17/17	99.3% 149/150

Supplementary Table 17 - ISO-ScreenPrompt Self-Consistency Analysis

Dataset	Method	Accuracy	Sensitivity	Specificity	Cost (USD)
SeroTracker NLP Test (Prevalence)	ISO-ScreenPrompt	96.5% 384/398	**98.5%** **195/198**	94.5% 189/200	$51.13
	ISO-ScreenPrompt SC	**97.5%** **390/400**	97.5% 195/200	**97.5%** **195/200**	$142.45 Batch API
Reinfection (Intervention Benefits)	ISO-ScreenPrompt	83.3% 333/400	89.7% 130/145	79.6% 203/255	$51.16
	ISO-ScreenPrompt SC	**87.5%** **350/400**	**92.4%** **134/145**	**84.7%** **216/255**	$130.87 Batch API

Supplementary Table 18 - Human Screening Calibration

Screening	Method	Accuracy	Sensitivity	Specificity
Dual Abstract	SeroTracker Human	93.5% 374/400	84.6% 77/91	96.1% 297/309
	Team 1 (Rev 1 + 2 duplicate) (Rev 3 conflicts)	93.5% 374/400	82.4% 75/91	96.8% 299/309
	Team 2 (Rev 3 + 4 duplicate) (Rev 1 conflicts)	93.5% 374/400	85.7% 78/91	95.8% 296/309
Dual full-text	SeroTracker Human	93.3% 373/400	81.3% 74/91	96.8% 299/309
	Team 1 (Rev 1 + 2 duplicate) (Rev 3 conflicts)	93.3% 373/400	79.1% 72/91	97.4% 301/309
	Team 2 (Rev 3 + 4 duplicate) (Rev 1 conflicts)	94.3% 377/400	81.3% 74/91	98.1% 303/309

Supplementary Table 19 - Abstract ScreenPrompt vs Human Dual Abstract Screening

Dataset	Method	Accuracy	Sensitivity	Specificity
SeroTracker NLP Test (Prevalence)	ISO-ScreenPrompt	93.0% 372/400	**96.7%** **88/91**	91.9% 284/309
	Dual Human (SeroTracker Human)	**93.5%** **374/400**	84.6% 77/91	**96.1%** 297/309
Reinfection (Intervention Benefits)	ISO-ScreenPrompt	78.3% 311/397	**95.8%** **136/142**	68.6% 175/255
	Dual Human	**89.5%** **358/400**	84.1% 122/145	**92.5%** **236/255**
PA-Outcomes (Intervention Harms)	ISO-ScreenPrompt	78.9% 310/393	100.0% 16/16	78.0% 294/377
	Dual Human	**94.8%** **379/400**	100.0% 16/16	**94.5%** **363/384**
PA-Testing (Diagnostic Test Accuracy)	ISO-ScreenPrompt	94.0% 375/399	**100.0%** **44/44**	93.2% 331/355
	Dual Human	**97.8%** **391/400**	97.8% 44/45	**97.7%** **347/355**
SVCF (Prognosis)	ISO-ScreenPrompt	94.4% 152/161	**100.0%** **17/17**	93.8% 135/144
	Dual Human	**97.6%** **163/167**	88.2% 15/17	**98.7%** **148/150**

Supplementary Table 20 - ISO-ScreenPrompt vs Full Dual Screening

Dataset	Method	Accuracy	Sensitivity	Specificity
SeroTracker NLP Test (Prevalence)	ISO-ScreenPrompt	**94.5%** **378/400**	**94.5%** **86/91**	94.5% 292/309
	Dual Human (SeroTracker Human)	93.3% 373/400	81.3% 74/91	**96.8%** **299/309**
Reinfection (Intervention Benefits)	ISO-ScreenPrompt	**83.3%** **333/400**	**89.7%** **130/145**	79.6% 203/255
	Dual Human	76.8% 307/400	44.1% 64/145	**95.3%** **243/255**
PA-Outcomes (Intervention Harms)	ISO-ScreenPrompt	84.3% 337/400	**100%** **16/16**	83.6% 321/384
	Dual Human	**97.8%** **391/400**	93.8% 15/16	**97.9%** **376/384**
PA-Testing (Diagnostic Test Accuracy)	ISO-ScreenPrompt	93.2% 372/399	**100.0%** **45/45**	92.4% 327/354
	Dual Human	**96.8%** **387/400**	80.0% 36/45	**98.9%** **351/355**
SVCF (Prognosis)	ISO-ScreenPrompt	**99.4%** **166/167**	**100.0%** **17/17**	99.3% 149/150
	Dual Human	97.6% 163/167	76.5% 13/17	**100.0%** **150/150**

Supplementary Table 21 - Time and Cost Savings Analysis for ISO-ScreenPrompt and Full-dual Screening

Dataset	Number of Articles (full-texts)	Estimated time - Human (hours)	Estimated review cost	Estimated time - GPT (hours)	ISO-ScreenPrompt GPT4-0125-preview cost (USD) (Hours)
SeroTracker	130436 (3659)	3393.6	$67872.0	52.2	$12661.30
Reinfection	6724 (1256)	530.7	$10614.67	2.7	$859.97
PA-Testing	8000 (248)	216.0	$4320.00	3.2	$917.76
PA-Outcome	5376 (74)	114.3	$2285.33	2.2	$600.63
SVCF	2257 (95)	69.3	$1385.67	0.9	$196.43

Supplementary Table 22 - Time and Cost Savings Analysis for Abstract ScreenPrompt and Single Human-Reviewer Screening

Dataset	Number of Articles	Estimated time - Human (hours)	Estimated review cost Single Reviewer	Estimated time - GPT (hours)	Abstract ScreenPrompt GPT4-0125-preview cost (USD)
SeroTracker	130436	1086.97	$21739.33	15.2	$4122.49
Reinfection	6724	56.03	$1120.67	0.8	$244.93
PA-Testing	8000	66.67	$1333.40	0.9	$266.64
PA-Outcome	5376	44.80	$896.00	0.6	$162.55
SVCF	2257	18.81	$376.17	0.3	$55.25

Supplementary Note 1

Below represents the screening documents provided to human reviewers for each selected SR. The documents contain the same study objectives, inclusion and exclusion criteria used in prompting ('simplified objectives and inclusion/exclusion criteria'), and any additional internal protocol documents provided by the original study authors for each SR.

SeroTracker

Simplified Objectives + Inclusion/Exclusion Criteria

Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):

Population

1. Humans of any age
2. Including COVID-19 antigen positive persons and those with suspected disease if not deliberately sampled.

Study design

1. Sero-surveys – defined as the collection and testing of serum (or proxy such as oral fluid) specimens from a sample of a defined population over a specified period of time to estimate the prevalence of antibodies against SARS-CoV-2 as an indicator of immunity
2. Cross-sectional, repeated cross sectional, and cohort study designs, with serology measurements at single time points or repeated at multiple time points

Special design

1. Include systematic reviews and meta-analysis of seroprevalence studies for the purpose of tracking evidence synthesis efforts

Sampling

1. Any sampling method

Types of evidence

1. Published or unpublished academic literature, grey literature (government or institutional reports), or media reports. Slide deck presentations were included if we could identify the person giving the presentation and the date of the presentation

Outcome measures

1. Reports a seroprevalence estimate (proportion of the population with detectable antibodies)
2. Reports the number of participants enrolled in the study (denominator)
3. Reports study sampling end date/week
4. Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country

Exclusion Criteria (if any met then exclude):

Population

1. Non-human (e.g., in silico, animal, in vitro)
2. The study only included individuals with suspected, active, or previously diagnosed with COVID-19 using PCR, antigen testing, clinical assessment, or self-assessment
3. The study only included individuals vaccinated against SARS-CoV2

Study design

1. Study designs other than cross-sectional or cohort design: case reports, case-control studies, evaluations of serological tests, study protocols

Types of evidence

1. Multimedia sources of data (audio clips, video clips) were excluded due to the feasibility of extracting. Slide deck presentations were excluded if we could not identify the person giving the presentation and the date of the presentation
2. Dashboards not associated with a defined serology study

Outcome measures

1. Only reports incidence or prevalence of SARS-CoV-2 antigen (as opposed to antibody)
2. Does not report study sampling end date/week
3. Does not report the number of participants included in the study (sample denominator)
4. Does not report the location at which the study took place

Original author protocol

SeroTracker Inclusion/Exclusion Criteria

Criteria for including evidence (must meet all the criteria to be included)

Characteristics	Criteria for inclusion
Population	· Humans of any age o Including COVID-19 antigen positive persons and those with suspected disease if not deliberately sampled.
Study design	· Sero-surveys – defined as the collection and testing of serum (or proxy such as oral fluid) specimens from a sample of a defined population over a specified period of time to estimate the prevalence of antibodies against SARS-CoV-2 as an indicator of immunity · Cross-sectional, repeated cross sectional, and cohort study designs, with serology measurements at single time points or repeated at multiple time points
*Special design	· Include systematic reviews and meta-analysis of seroprevalence studies for the purpose of tracking evidence synthesis efforts
Sampling	· Any sampling method
Types of evidence	· Published or unpublished academic literature, grey literature (government or institutional reports), or media reports. Slide deck presentations were included if we could identify the person giving the presentation and the date of the presentation
Outcome measures	· Reports a seroprevalence estimate (proportion of the population with detectable antibodies) · Reports the number of participants enrolled in the study (denominator) · Reports study sampling end date/week · Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country
Languages	· Any

Criteria for excluding evidence (if any met then exclude)

Characteristics	Criteria for exclusion

Population	· Non-human (e.g., *in silico*, animal, *in vitro*) · The study only included individuals with suspected, active, or previously diagnosed with COVID-19 using PCR, antigen testing, clinical assessment, or self-assessment · The study only included individuals vaccinated against SARS-CoV2
Study design	· Study designs other than cross-sectional or cohort design: case reports, case-control studies, evaluations of serological tests, study protocols
Sampling	· N/A
Types of evidence	· Multimedia sources of data (audio clips, video clips) were excluded due to the feasibility of extracting. Slide deck presentations were excluded if we could not identify the person giving the presentation and the date of the presentation · Dashboards not associated with a defined serology study
Outcome measures	· Only reports incidence or prevalence of SARS-CoV-2 antigen (as opposed to antibody) · Does not report study sampling end date/week · Does not report the number of participants included in the study (sample denominator) · Does not report the location at which the study took place
Language	· N/A

Reinfection

Simplified Objectives + Inclusion/Exclusion Criteria

We aimed to systematically review the evidence for the magnitude and duration of the effectiveness of (i) previous infection and (ii) hybrid immunity against multiple clinical outcomes of SARS-CoV-2 infection caused by the omicron variant. We also aimed to examine the comparative protection of hybrid immunity relative to previous infection only, vaccination only, and hybrid immunity with fewer vaccine doses.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):

Population

1. Humans of any age, in any geographical setting.

Exposure Group

1. Confirmed case of SARS-CoV-2 infection with or without COVID-19 vaccination.

A. SARS-CoV-2 infection will be defined as a confirmed case according to the following criteria, adapted from WHO case definitions (positive nucleic acid amplification test (NAAT) according to laboratory records or self report, positive SARS-CoV-2 antigen rapid diagnostic test (AgRDT)[a] with high accuracy according to laboratory records or self report, or a positive serology test from a lab-based assay (i.e. CLIA/ELISA) or an antibody-detecting rapid diagnostic test (Ab-RDT) with high accuracy[a]).

2. Studies will be included if they report on individuals with previously confirmed infection that have documented vaccination (partially, fully, or boosted), as defined in the randomized controlled trials for each vaccine.

A. Partial vaccination will be defined as ≥14 days after a single dose of Pfizer/BioNTech-Comirnaty, <7 days from the second dose for Pfizer/BioNTech-Comirnaty (BNT162b2), ≥14 days after a single dose of AstraZeneca-Vaxzevria, <14 days from the second dose for AstraZeneca-Vaxzevria, ≥14 days after a single dose of Moderna-mRNA-1273, <14 days from the second dose of Moderna-mRNA-1273, <14 days from the first dose of Janssen-Ad26.COV2.S, and ≥14 days after a single dose of Sinovac-CoronaVac.

B. Full vaccination will be defined as >7 days from the second dose for Pfizer/BioNTech-Comirnaty, >14 days from the first dose of Janssen-Ad26.COV2.S, >14 days from the second dose for AstraZeneca-Vaxzevria, Moderna-mRNA-1273, or Sinovac-CoronaVac.

C. Booster vaccination one will be defined as >=7 days from an additional dose after full vaccination.

D. Booster vaccination two will be defined as >=7 days from an additional dose after booster vaccination one.

Comparison Group

1. no previous vaccinations and no previously confirmed SARS-CoV-2 infection defined using WHO criteria;
2. previously confirmed SARS-CoV-2 infection defined using WHO criteria;
3. partial vaccination (defined above);
4. full vaccination (defined above);
5. booster vaccination (defined above).

Outcome

1. SARS-CoV-2 reinfection defined as a possible, probable, or confirmed reinfection case according to the following criteria, adapted from WHO case definitions.

A. Possible reinfection case will be defined as NAAT or AgRDT SARS-CoV-2 positive case with a history of a primary SARS-CoV-2 infection diagnosed by serology, with at least 60 days between the positive serology test and the subsequent positive NAAT or AgRDT.

B. Probable reinfection case will be defined as NAAT or AgRDT SARS-CoV-2 positive case with a history of a primary SARS-CoV-2 infection diagnosed by NAAT or AgRDT, with at least 90 days between the episodes. Alternatively, genomic evidence for the second episode is available and includes lineage that was not submitted to SARS-Cov-2 genomic databases at the time of first infection.

C. Confirmed reinfection case will be defined as two PCR positive episodes supported by viral genomic data from both episodes of infection revealing different Pango lineages. If viral genomic data reveal two distinct Pango lineages this will qualify as adequate evidence to confirm reinfection, regardless of the time elapsed between the two episodes.

Study Design

1. Test-negative case-control, traditional case-control, cross-sectional, cohort, non-randomized controlled trials, and randomized controlled trials.

Type of literature

1. Published peer-reviewed research articles, preprints, and grey literature in any language. We will prioritize peer-reviewed versions of articles for inclusion and analysis in instances where pre-print versions of peer-reviewed articles are available.

<u>Exclusion Criteria (if any met then exclude):</u>

Population

1. N/A

Exposure Group

1. No evidence of prior confirmed case. No information on the timing, brand, or dose number for the vaccination in hybrid immunity studies.

Comparison Group

1. N/A

Outcome

1. Prior infection studies not reporting the period of time between primary infection and reinfection such that determining reinfection according to the inclusion criteria is not possible.

2. Hybrid immunity studies not reporting the period of time between either the determination of primary infection or vaccination.

Study Design

1. Case reports, case series, incomplete randomized controlled trials, and review papers.

Type of literature

1. Media, news stories, and conference abstracts.

Original author protocol

SARS-CoV-2 protective effectiveness of prior infection and hybrid immunity: a systematic review protocol

2.2 Study inclusion and exclusion criteria

Table 1. Inclusion criteria	
Population	Humans of any age, in any geographical setting.

Exposure group	Confirmed case of SARS-CoV-2 infection with or without COVID-19 vaccination. SARS-CoV-2 infection will be defined as a **confirmed case** according to the following criteria, adapted from WHO case definitions[1] (positive nucleic acid amplification test (NAAT) according to laboratory records or self report, positive SARS-CoV-2 antigen rapid diagnostic test (AgRDT)[a] with high accuracy according to laboratory records or self report, or a positive serology test from a lab-based assay (i.e. CLIA/ELISA) or an antibody-detecting rapid diagnostic test (Ab-RDT) with high accuracy[a]). Studies will be included if they report on individuals with previously confirmed infection that have documented vaccination (partially, fully, or boosted), as defined in the randomized controlled trials for each vaccine. **Partial vaccination** will be defined as ≥14 days after a single dose of Pfizer/BioNTech-Comirnaty, <7 days from the second dose for Pfizer/BioNTech-Comirnaty (BNT162b2), ≥14 days after a single dose of AstraZeneca-Vaxzevria, <14 days from the second dose for AstraZeneca-Vaxzevria, ≥14 days after a single dose of Moderna-mRNA-1273, <14 days from the second dose of Moderna-mRNA-1273, <14 days from the first dose of Janssen-Ad26.COV2.S, and ≥14 days after a single dose of Sinovac-CoronaVac. **Full vaccination** will be defined as >7 days from the second dose for Pfizer/BioNTech-Comirnaty, >14 days from the first dose of Janssen-Ad26.COV2.S, >14 days from the second dose for AstraZeneca-Vaxzevria, Moderna-mRNA-1273, or Sinovac-CoronaVac. **Booster vaccination one** will be defined as >=7 days from an additional dose after full vaccination. **Booster vaccination two** will be defined as >=7 days from an additional dose after booster vaccination one.

Comparison group	Five comparison groups will be eligible: (1) no previous vaccinations and no previously confirmed SARS-CoV-2 infection defined using WHO criteria; (2) previously confirmed SARS-CoV-2 infection defined using WHO criteria; (3) partial vaccination (defined above); (4) full vaccination (defined above); (5) booster vaccination (defined above).
Outcome	SARS-CoV-2 reinfection defined as a possible, probable, or confirmed reinfection case according to the following criteria, adapted from WHO case definitions. **Possible reinfection case** will be defined as NAAT or AgRDT SARS-CoV-2 positive case with a history of a primary SARS-CoV-2 infection diagnosed by serology, with at least 60 days between the positive serology test and the subsequent positive NAAT or AgRDT. **Probable reinfection case** will be defined as NAAT or AgRDT SARS-CoV-2 positive case with a history of a primary SARS-CoV-2 infection diagnosed by NAAT or AgRDT, with at least 90 days between the episodes. Alternatively, genomic evidence for the second episode is available and includes lineage that was not submitted to SARS-Cov-2 genomic databases at the time of first infection. **Confirmed reinfection case** will be defined as two PCR positive episodes supported by viral genomic data from both episodes of infection revealing different Pango lineages. If viral genomic data reveal two distinct Pango lineages this will qualify as adequate evidence to confirm reinfection, regardless of the time elapsed between the two episodes.
Study design	Test-negative case-control, traditional case-control, cross-sectional, cohort, non-randomized controlled trials, and randomized controlled trials.

Type of literature	Published peer-reviewed research articles, preprints, and grey literature in any language. We will prioritize peer-reviewed versions of articles for inclusion and analysis in instances where pre-print versions of peer-reviewed articles are available.

[a]For AgRDT ≥80% sensitivity and ≥97% specificity of test, compared to NAAT, in suspected cases of infection; For AbRDT, ≥90% sensitivity (>14 days post symptom onset) and ≥97% specificity, compared to a reference lab-based test, in suspected cases of infection.

Table 2. Exclusion criteria

Population	N/A
Exposure group	No evidence of prior confirmed case. No information on the timing, brand, or dose number for the vaccination in hybrid immunity studies.
Comparison group	N/A
Outcome	Prior infection studies not reporting the period of time between primary infection and reinfection such that determining reinfection according to the inclusion criteria is not possible. Hybrid immunity studies not reporting the period of time between either the determination of primary infection or vaccination.
Study design	Case reports, case series, incomplete randomized controlled trials, and review papers.
Type of literature	Media, news stories, and conference abstracts.

PA-Testing

Simplified Objectives + Inclusion/Exclusion Criteria

The purpose of this study was to assess the characteristics of confirmatory tests for primary aldosteronism (PA) and to interpret these in the context of study design and potential risks of bias.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):

1. The study is about primary aldosteronism (PA). Please note that primary hyperaldosteronism is a synonym.
2. The study examined at least one of the guideline-recommended confirmatory tests for PA. We are interested in the saline infusion test (SIT), oral salt loading test (SLT), fludrocortisone suppression test (FST), and the captopril challenge test (CCT). Please note some may have synonyms (e.g., SIT = intravenous saline suppression test [IVSS]).
3. Research articles reporting original data
4. The study is conducted in humans of any age
5. The study is in the English language
6. The performance of the confirmatory test(s) was compared with an independent reference standard A reference standard needs to be present to verify disease status. These may include: (i) clinical response to treatment (adrenalectomy and/or medical therapy), (ii) adrenal vein sampling results, (iii) histopathology, (iv) or another confirmatory test.
7. The confirmatory test was used to diagnose PA, rather than exclusively for subtyping. We are interested in knowing how confirmatory testing works for diagnosing PA, not just subtyping. If the confirmatory test was for subtyping, it is still possible to compare how many people had PA vs. non-PA.
8. The data (as published) is extractable for a 2x2 table (TP, FP, FN, TN). If the data are reported in any of the following formats, a 2x2 table can be reconstructed:(i) 2x2 table given, (ii) TP, FP, FN, TN rates given, (iii) Total number of patients (with disease) and total number of study subjects (with and without disease) is known, and corresponding sensitivity and specificity are given.

Exclusion criteria (if any met then exclude):

1. No mention about primary aldosteronism (PA)
2. Use of confirmatory tests that fall outside of guide-line recommendations
3. Non-human (e.g., in silico, animal, in vitro).
4. Conference abstracts, reviews (systematic reviews and narrative reviews), editorials, protocols, and secondary publications (data already published in another study).

5. No comparison of confirmatory test performance with a reference standard.
6. Confirmatory test was used only for subtyping, and unable to compare cases of PA vs. non-PA
7. Data not extractable for 2x2 table

Original author protocol

Confirmatory Testing in Primary Aldosteronism Systematic Review Reference Sheet

(Version: May 28, 2021)

Primary Screen (Title/Abstract)

Question 1: *Is the study about primary aldosteronism? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes
b. No
c. Unclear

Diagnosis of interest	We are interested in primary aldosteronism (PA). Please note that primary hyperaldosteronism is a synonym.

Question 2: *Does the study examine at least one of the guideline-recommended confirmatory tests for PA? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes
b. No
c. Unclear

Index test	We are interested in the saline infusion test (SIT), oral salt loading test (SLT), fludrocortisone suppression test (FST), and the captopril challenge test (CCT). Please note some may have synonyms (e.g., SIT = intravenous saline suppression test [IVSS]).

Question 3: *Is this a research study reporting original data? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes
b. No
c. Unclear

Index test	We are only interested in original studies. We will exclude conference abstracts, reviews, editorials, and protocols.

Question 4: *Is this article potentially relevant but in one of the following formats?*

a. Conference abstract

b. Systematic review

c. Narrative review

d. Secondary publication (data already published in another study)

Other reports	Although these articles will not be abstracted for our data analysis, we will still save them and later review them for background information.

Question 5: *Is this a human study? [*Exclude if "No." Include for full-text review if "Yes"]*

a. Yes

b. No

Population	We will only be considering human studies.

Question 6: *Is this an English study? [*Exclude if "No." Include for full-text review if "Yes"]*

a. Yes

b. No

Population	We will only be considering English studies.

Secondary Screen (Full-text)

Question 1: *Was the performance of the confirmatory test(s) compared with an independent reference standard? [*Exclude if "No." Include for final analysis if "Yes"]*

a. Yes

b. No

Reference test	A reference standard needs to be present to verify disease status. These may include: · Clinical response to treatment (adrenalectomy and/or medical therapy) · Adrenal vein sampling results · Histopathology · Another confirmatory test

Question 2: *Was the confirmatory test used to diagnose PA, rather than exclusively for subtyping? [*Exclude if "No." Include for final analysis if "Yes"]*

a. Yes: either for diagnosing PA; or if it was for subtyping, it is still possible to compare how many people had PA vs. non-PA

b. No: for subtyping only, and unable to compare cases of PA vs. non-PA

Diagnosis of interest	We are interested in knowing how confirmatory testing works for diagnosing PA, not just subtyping.

Question 3: *Are the data (as published) extractable for 2x2 table (TP, FP, FN, TN)? [*Exclude if "No." Include for final analysis if "Yes"]*

a. Yes

b. No

Outcome	If the data are reported in any of the following formats, a 2x2 table can be reconstructed: · 2x2 table given · TP, FP, FN, TN rates given · Total number of patients (with disease) and total number of study subjects (with and without disease) is known, and corresponding sensitivity and specificity are given

PIRD Framework:
Population = patients suspected of having primary aldosteronism (PA)
Index test = confirmatory test for PA
Reference standard = clinical response to targeted treatment (gold), adrenal vein sampling lateralization (surrogate), histopathology (surrogate), or another confirmatory test (surrogate)
Diagnosis of interest = PA

PA-Outcomes

Simplified Objectives and Inclusion/Exclusion Criteria

We aimed to conduct a meta-analysis to examine the clinical outcomes of surgery vs medical therapy with respect to mortality, composite major adverse cardiovascular events (MACE, and its individual components), progression to chronic kidney disease, and incident diabetes mellitus.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):

1. The study is about primary aldosteronism (PA). Please note that primary hyperaldosteronism is a synonym.
2. Reports surgery (adrenalectomy) or medication (mineralocorticoid receptor antagonists, spironolactone, eplerenone) treatment for primary aldosteronism.
3. Reports mortality, MACE, ACS, stroke, arrhythmia, heart failure, chronic kidney disease, and/or incident diabetes clinical outcomes after treatment.
4. Research articles reporting original data.
5. Randomized clinical trials, cohort studies, and cross-sectional studies.
6. The study is conducted in humans of any age.
7. The study is in the English language.

Exclusion criteria (if any met then exclude):

1. Does not report about primary aldosteronism (PA).
2. Does not report treatment for primary aldosteronism.
3. Does not report mortality, major adverse cardiovascular events (MACE), acute coronary syndrome (ACS), stroke, arrhythmia, heart failure, chronic kidney disease, and/or incident diabetes clinical outcomes after treatment.
4. Non-human (e.g., in silico, animal, in vitro).
5. Conference abstracts, case reports, case series, reviews (systematic reviews and narrative reviews), editorials, protocols, and secondary publications (data already published in another study).

Original author protocol

Primary Aldosteronism Treatment Response: Systematic Review Reference Sheet

(Version: June 1, 2022 revised)

Primary Screen (Title/Abstract)

Question 1: *Is the study about primary aldosteronism? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes

b. No

c. Unclear

Diagnosis of interest	We are interested in primary aldosteronism (PA). Please note that primary hyperaldosteronism is a synonym.

Question 2: *Does this study describe treatment for primary aldosteronism? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes

b. No

c. Unclear

Intervention (exposure) of interest	We are interested in treatment outcomes for primary aldosteronism (PA), comparing surgery (adrenalectomy) vs. medications (mineralocorticoid receptor antagonists, spironolactone, eplerenone).

Question 3: *Does this study describe hard clinical outcomes? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes

b. No

c. Unclear

Outcomes of interest	We are interested in hard clinical outcomes for primary aldosteronism (PA), such as mortality, MACE, ACS, stroke, arrhythmia, heart failure, chronic kidney disease, and/or incident diabetes.

Question 4: *Is this a research study reporting original data? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes

b. No

c. Unclear

Design	We are only interested in original studies. We will exclude conference abstracts, reviews, editorials, and protocols.

Question 5: *Is this article potentially relevant but in one of the following formats?*

a. Conference abstract

b. Systematic review

c. Narrative review

d. Secondary publication (data already published in another study)

Other reports	Although these articles will not be abstracted for our data analysis, we will still save them and later review them for background information.

Question 6: *Is this a human study? [*Exclude if "No." Include for full-text review if "Yes"]*

a. Yes

b. No

Population	We will only be considering human studies.

Question 7: *Is this an English study? [*Exclude if "No." Include for full-text review if "Yes"]*

a. Yes

b. No

Language	We will only be considering English studies.

Secondary Screen (Full-text)

Question 1: *Does the study report on an outcome of interest? [*Exclude if "No." Include for final analysis if "Yes"]*

a. Yes

b. No

Outcome	The study must report on an outcome of interest with a quantitative metric (either as a summary statistic, like a HR, RR, or OR; or as a survival/cumulative incidence curve). These may include: · Mortality, all-cause death · Cardiovascular death · Major adverse cardiovascular events / composite cardiovascular events · Acute coronary syndrome / myocardial infarction / unstable angina / coronary heart disease · Coronary revascularization (PCI, CABG) · Stroke (stroke, TIA) · Arrhythmias (atrial fibrillation, atrial flutter, ventricular fibrillation, ventricular tachycardia) · Heart failure / heart failure hospitalization · Diabetes · Chronic kidney disease

Question 2: *Does the study directly compare medication (e.g., mineralocorticoid receptor antagonist) vs. surgery for patients with PA, and additionally stratify outcomes according to the treatment received? [*Exclude if "No." Include for final analysis if "Yes"]*

a. Yes: it is possible to compare outcomes for PA patients who had surgery vs. medications

b. No: unable to compare outcomes for surgery vs. medications (e.g., because all patients with PA are grouped together)

Exposure of interest	We are interested in knowing how treatment outcomes differ between surgery and medications (not just the natural history of PA).

Question 3: *Does this study report a hard clinical outcome that can be extracted according to treatment received? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes

b. No

c. Unclear

Outcomes of interest	We are interested in hard clinical outcomes for primary aldosteronism (PA), such as mortality, MACE, ACS, stroke, arrhythmia, heart failure, chronic kidney disease, and/or incident diabetes.

Question 4: *Is this a research study reporting original data? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

a. Yes

b. No

c. Unclear

Design	We are only interested in original studies. We will exclude conference abstracts, reviews, editorials, and protocols.

PECOD Framework:

The *population* of interest are patients with PA.

The *exposure/intervention* of interest is surgical adrenalectomy and the *comparator* is medical treatment with a mineralocorticoid receptor antagonist (e.g., spironolactone, eplerenone, etc).

The *primary outcome* is all-cause mortality. *Secondary outcomes* include other commonly reported clinical events include incident major adverse cardiovascular events (and its individual components, such as myocardial infarction, stroke, revascularization, arrhythmia), congestive heart failure, atrial fibrillation, chronic kidney disease, and diabetes mellitus.

The study *designs* that will be considered include randomized controlled trials and observational studies.

SVCF

Simplified Objectives + Inclusion/Exclusion Criteria

Objectives: To evaluate the association of low SVC flow, diagnosed in the first 48 hours after birth echocardiography, with neurological morbidity and mortality, among very preterm neonates.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):

1. Preterm infants <32 weeks gestational age who had echocardiography done within the first 48 hours after birth with evaluation of SVC flow

A. Prognostic factor: Low SVC flow identified by Doppler assessment during echocardiography performed in the first 48 hours after birth.

B. Comparison: Normal SVC Flow

2. Reported outcomes including intraventricular hemorrhage (IVH), presence of periventricular leukomalacia (PVL), all-cause mortality before discharge from neonatal intensive care unit (NICU), neurodevelopmental impairment in early childhood or any diagnosed Cerebral Palsy, visual and/or hearing deficits, or necrotizing enterocolitis (NEC)

A. Any grade IVH diagnosed in the first 7 days after birth by cranial ultrasonography

B. Severe IVH (defined as stage 3 or higher according to Papile's classification 14) diagnosed in the first 7 days after birth

3. Randomized controlled trials, cohort or case-control studies

Exclusion criteria (if any met then exclude):

1. Non-human (e.g., in silico, animal, in vitro)
2. Cross-sectional studies, narrative reviews, case series or case reports

Original author protocol

Association of early-life low superior vena cava flow among preterm neonates and death or cerebral haemorrhage: a systematic review and meta-analysis: Protocol

The objective of this study is to systematically review and meta—analyse the association of low SVC flow during the transitional period among preterm neonates < 32 weeks GA with mortalityand adverse neurological morbidity..

PICO/PFO outline-

Population – Preterm infants <32 weeks gestational age

Prognostic factor- Low SVC flow identified by Doppler assessment during echocardiography performed in the first 48 hours after birth (measured as ml/kg/min)

Comparison- Patients with normal SVC flow

Outcomes-

Primary- **Any grade IVH** diagnosed in the first 7 days of life by cranial ultrasonography

Secondary-

- Severe IVH (defined as stage 3 or higher according to Papile's classification [14]) diagnosed in the first 7 days of life,
- Presence of PVL (diagnosed by 28 days of life)
- Mortality within the neonatal period (defined as the first 28 days of life)
- Neurodevelopmental impairment in early childhood (Defined as a composite outcome of any of the following: Cerebral palsy with Gross Motor Function Classification System score >=1 or Bayley-III motor composite <85; Bayley cognitive composite < 85; Bayley language composite <85; Any sensorineural/mixed hearing loss; Any unilateral or bilateral visual impairment)[15]
- Necrotizing Enterocolitis (NEC) compared between low and normal SVC flow groups

METHODOLOGY

This systematic review and meta-analysis will be conducted according to the PRISMA guidelines and Cochrane methodology.

Eligibility Criteria

Randomized controlled trials, cohort or case-control studies that evaluated the following population characteristics will be included for this study:

Preterm infants <32 weeks gestational age who had echocardiography done within the first 48 hours after birth with evaluation of SVC flow. Studies must have compared any of the above stated clinical outcomes among preterm neonates with low vs. normal SVC flow to be

considered for inclusion. Studies will be included only if they exclusively include human subjects and there will be no language limitations.

Exclusion criteria: Cross-sectional studies, narrative reviews, case series or case reports on this topic will be excluded. We will also exclude any animal studies. Any identified studies without any available full text from the included databases or the original authors will also be excluded.