Single- versus two-stage AI citation screening for clinical systematic reviews: a prevalence-aware, access-limited cost analysis

Sergii Ivakhno[1]
[1] Axelium, London, United Kingdom

Corresponding author: Sergii Ivakhno, sergii.ivakhno@axelium.ai
Word count: ~3,700 (main text, excluding references, tables, and figure captions)
Keywords: systematic review, citation screening, large language models, two-stage screening, cost-effectiveness, open access
Running title: Cost of single- vs two-stage AI citation screening

Abstract

Background. Large language models (LLMs) are increasingly used to screen citations for systematic reviews. A core design choice is *single-stage* screening (read every candidate's full text) versus *two-stage* (an AI title/abstract gate, then full text only for survivors). Prior work has derived the cost trade-off analytically or observed it on a single corpus; none has measured it across reviews of differing eligibility prevalence or treated full-text *obtainability* as an input.

Methods. We reconstructed the screening pools of four published COVID-era pairwise-RCT meta-analyses (ivermectin–Cochrane, colchicine–PLOS, vitamin-D–Frontiers, ivermectin–BMC) through one production LLM screening pipeline. Reproduced include-prevalence spanned 1.7–9.2%. Using each review's published included trials as ground truth (matched by PMID/DOI/registry ID), we measured the abstract gate's recall, specificity and pass-fraction (the share of records advanced to full text); per-record abstract and full-text read costs; and the open-access retrievability of the included trials. We expressed the single-versus-two-stage choice as a cost break-even — the full-text:abstract cost ratio at which two-stage starts to pay — and placed each corpus relative to it. Recall comparisons assume a perfect full-text screen ($r_{FT} = 1$; no full-text screening step was run — see §3.2).

Results. Against human-confirmed includes, the abstract gate's pooled recall was 43/46 = 93.5% (95% CI 82–98; a random-effects logit estimate gives 89%, 76–95), with specificity 71–85% and pass-fraction 17–31%. Two-stage screening was 29–43% cheaper than single-stage in all four corpora. This cost ranking was insensitive to the full-text:abstract cost ratio over its realistic range (two-stage cheaper whenever the ratio exceeds the per-corpus break-even of 1.20–1.45), because the gate's selectivity — not read-cost — drives the saving. The binding constraint on achievable recall was full-text access: only 35/46 = 76% (62–86) of included trials were openly retrievable, and the unretrievable ones were disproportionately high-impact paywalled trials. Under $r_{FT} = 1$, neither strategy reaches 95% recall by AI-reading alone. In supporting analyses, the gate's continuous confidence score added no relevance signal beyond its binary decision (score AUC 0.51 vs decision AUC 0.90); and in the few observed misses (n = 3, one corpus) the gate failed by confident single-criterion exclusion rather than flagged uncertainty.

Conclusions. Across these four COVID-era low-prevalence corpora, an AI abstract gate was the cost-efficient choice in every case and the full-text:abstract cost ratio did not change that; the practical frontier was full-text access rather than compute. We position this as building on recent agentic-SLR evaluation work, and as a single-pipeline proof-of-concept whose generalisation requires multi-model, multi-domain replication.

1. Introduction

Citation screening is among the most labour-intensive steps in a systematic review. A comprehensive search routinely returns thousands to tens of thousands of records, the large majority irrelevant, and each must be triaged against the review's eligibility criteria — conventionally in two passes, a rapid title-and-abstract screen followed by full-text assessment of the survivors. Performed by human reviewers this work is slow and a recognised bottleneck. An earlier generation of automation tackled it with machine-learning prioritization and active learning, reordering the screening queue so reviewers reach the eligible records sooner (Wallace et al. 2012; van de Schoot et al. 2021). The maturation of large language models (LLMs) has since shifted the emphasis from prioritization to direct judgement: recent systems apply general-purpose LLMs to decide eligibility from a record's text. Early evaluations report recall often competitive with human screeners — though with variable and frequently lower precision — at a fraction of the time and cost (Tran et al. 2024; Dennstädt et al. 2024; Khraisha et al. 2024). As these tools move into production review pipelines, the operative question shifts from whether an LLM can screen to how it should be deployed.

This paper studies the central design choice in that deployment. An automated screener can be run *single-stage* — reading every candidate's full text in one pass — or *two-stage*, interposing an AI title/abstract gate that advances only apparently-relevant records to a more expensive full-text read. The choice is the principal cost lever, because full-text reading dominates the per-record budget (more tokens than an abstract, plus retrieval and conversion). Single-stage spends that cost on every record; two-stage spends it only on the gate's survivors, but risks discarding eligible records the gate misjudges. The relevant question is therefore not which is cheaper in general but which is cheaper *at a fixed recall* — the fraction of truly eligible records retained. We treat recall as the held-fixed quantity and ask, across realistic reviews, when the gate earns its place.

Several literatures bear on this without settling it. The underlying economics are classical: whether to test items individually or pre-screen them in a cheaper batched stage is the group-testing problem of Dorfman (1943). The two-stage break-even algebra — the cost ratio at which gating begins to pay, as a function of gate selectivity — is inherited directly from that line, including its modern statistical treatment (Longford, *Classification in two-stage screening*, 2015). We claim no novelty for this algebra. Closer to the application, Shemilt et al. (2016, *Systematic Reviews*) costed alternative screening strategies for a single review using an empirical full-text-to-abstract cost ratio, but at one review and one prevalence. Most directly, Padarha et al. (the "AgentSLR" harness; arXiv:2603.22327) *observed* the abstract-versus-direct-full-text trade-off on one epidemiology corpus. They reported that a fully-automated abstract-gated funnel recovered fewer eligible records than reading full text directly (funnel recall ≈0.81 vs direct-full-text ≈0.89; their strongest configuration, human-abstract→AI-full-text, reached ≈0.92). This came at a stated cost of 2.3× screening runtime and a rise in OCR expenditure from US$36.6 to US$303.2 (an ≈8× increase, derived

from those figures). That work establishes the phenomenon but does not resolve *when* each strategy is preferable, nor calibrate the choice across reviews of differing prevalence.

A machine-learning literature on cost-sensitive cascades and classification with costly features (Trapeznikov & Saligrama; Janisch et al., 2019) develops the per-record routing machinery an adaptive screener might use; we draw on its framing but make no contribution to it. Full-text obtainability itself — open-access coverage, paywalls, and "could not retrieve" rates in review flow diagrams — is also well studied (Piwowar et al. 2018; Boudry et al. 2019; Uribe-Cavero et al. 2026; Jarvis et al. 2021). Our contribution is not that paywalls exist but the coupling of obtainability to achievable AI recall.

These strands share two gaps relevant here. First, none calibrates the single- versus two-stage choice across multiple real clinical reviews spanning eligibility prevalence — the parameter that, through pass-fraction, most plausibly governs which strategy wins. Second, none treats full-text obtainability as an explicit input to the cost/recall of the AI-screening decision. We address both with a prevalence-aware, access-aware analysis across four reconstructed clinical-RCT screening pools.

Our contributions are fourfold. First, we report measured gate operating characteristics and pass-fractions on four real clinical corpora, located on the inherited break-even. Second, we give the empirical observation that, over the realistic cost-ratio range, the cost-efficient choice does not change — a consequence of the small measured break-even, which we state and verify rather than claim as new theory. Third, to our knowledge, this is the first treatment of included-trial obtainability as an explicit input to the single- versus two-stage AI-screening cost/recall decision, with evidence that obtainability, not compute, is the binding recall constraint here. Fourth, we offer two supporting findings on the gate's confidence calibration and failure mode. As a four-corpus, single-model proof-of-concept we are explicit throughout about what does and does not generalise.

2. Methods

2.1 Corpora and reconstruction

We selected four published systematic reviews of COVID-19 randomised controlled trials, each a simple pairwise (single intervention versus control or standard care) meta-analysis: ivermectin (Cochrane, version pub3), colchicine (*PLOS ONE*), vitamin D (*Frontiers in Immunology*), and a second ivermectin review (*BMC Infectious Diseases*) — referred to as corpora A–D, respectively. We reconstructed each review's screening pool end-to-end through one production large-language-model systematic-review pipeline. PICO criteria were configured; a recall-oriented multi-source search (PubMed, ClinicalTrials.gov, Cochrane CENTRAL, Europe PMC, OpenAlex) assembled the candidate pool; every record received an LLM title/abstract screening decision; and full text was retrieved and parsed wherever openly available. The verbatim per-source search strategies are in Supplement S2. The reconstructed pools contained 311 (ivermectin–Cochrane), 161 (colchicine), 416 (vitamin D) and 259 (ivermectin–BMC) records, with reproduced include-prevalence 1.7–9.2%. Because reconstruction relied chiefly on freely-queryable sources, pools are smaller than the source reviews' multi-database screens and prevalence is correspondingly higher. We treat each reconstructed pool as the screening corpus and report drift against the published flow (Supplement S5).

2.2 Ground truth, matching, and its biases

For each review we took its **published list of finally-included trials** as ground truth, resolved to PubMed ID/DOI/registry identifier and matched to the pool by exact identifier (per-trial coverage and fate in Supplement S1). Coverage was 11/11, 5/5, 7/8 and 23/25 of the included trials surfaced in the pool; the three unsurfaced trials were not PubMed-indexed. Matched included trials were labelled positive (the *gold* set) and all other records negative; throughout, "gold" denotes a published included trial. Two biases follow and we flag them throughout. First, these positives already passed the original human two-stage screen *including a full-text read*, so measured gate recall is recall against human-confirmed includes and is plausibly **upward-biased** relative to eligibility-from-scratch. Second, eligible-but-out-of-window trials the review did not include are labelled negative. This can depress measured precision, and means some "false positives" may not be errors. We examine sensitivity to the second issue using one review's published excluded-with-reasons table and flag that a full multi-corpus sensitivity analysis is outstanding.

2.3 Abstract-gate operating characteristics

The screener emits a three-way decision (include / unsure / exclude). We evaluated a **strict** gate (advance `include`) and a **lenient** gate (advance `include + unsure`), the lenient variant trading a larger pass-fraction for a wider recall safety margin. On surfaced records, we computed, against ground truth, recall (sensitivity to the included trials), specificity, precision, and pass-fraction φ (the share advanced to full text). Per-corpus proportions are reported with Wilson 95% CIs. Because positive counts are small (5–23 per corpus) and three of four corpora sit at 100% recall, we report the **aggregate pooled proportion** (43/46) as the primary summary and a DerSimonian–Laird random-effects logit estimate alongside it. With k = 4 and several recall estimates pinned at the 100% boundary, between-corpus heterogeneity is not interpretable and we do not assert homogeneity.

2.4 Cost model and break-even

We measured the per-record abstract-screening cost, c_a, from pipeline telemetry (mean ≈0.44 milli-USD, ≈2,460 input tokens). We characterised the full-text-to-abstract cost ratio, $c = c_f/c_a$, from input-token volumes for a full-text read (≈5,400–8,000 tokens) relative to an abstract — ≈2–3× single-pass (up to ≈4–6× for multi-chunk papers) for the open-access HTML/XML of these corpora. We then swept c over 2–50 for generality.

In units of c_a, single-stage reads every record's full text at cost c. Two-stage instead pays 1 for the abstract screen on every record, plus c on the fraction φ it advances — a total of $1 + \varphi c$. Two-stage is therefore cheaper exactly when $c > c^* = 1/(1-\varphi)$. The ranking reverses (single-stage cheaper) only if $\varphi > 1 - 1/c$, i.e. a much less selective gate or near-free full text. This cost ratio is **input-token-only**: output tokens and retrieval/conversion cost are excluded (the latter was not instrumented), so measured c is a lower bound; for non-OA records requiring OCR it would be substantially higher.

On recall, final inclusion requires reading the full text, so achievable recall is bounded by obtainability. Let β_{gold} be the fraction of included trials retrievably parsed. Assuming a perfect full-text screen ($r_{FT} = 1$), single-stage recall $\leq \beta_{gold}$ and two-stage recall $\leq r_{gate} \cdot \beta_{gold}$, where r_{gate} is the gate's recall. No full-text screening step was run; the recall comparison is therefore a model, not a measurement, and we report it conditionally (§3.2).

2.5 Confidence and failure-mode analysis

The screener's structured output records a per-record confidence (0–1) and a per-PICO-question judgement (population, intervention, comparator, outcome); we release these per record (`confidence_per_record.csv`; the full released-data manifest is Supplement S7). To test whether the confidence *score* is a usable relevance signal, we computed the area under the ROC curve (Mann–Whitney) for predicting gold status from (i) the raw confidence, (ii) the decision-signed confidence, and (iii) the confidence within the advanced (included) set. To characterise the failure mode, we inspected every included trial the strict gate missed, recording its decision bin, confidence, and which PICO question(s) it judged unmet.

3. Results

3.1 The abstract gate is a high-recall, high-specificity filter

Across the four reconstructed pools the strict abstract gate behaved as a high-recall, high-specificity first stage. Against the human-confirmed included trials it recovered 43 of 46 surfaced golds, a pooled recall of 93.5% (Wilson 95% CI 82–98). A more conservative DerSimonian–Laird random-effects logit estimate, which down-weights the small perfect-recall corpora, gives 89% (76–95); we report both rather than choose between them. Three of the four corpora sat at 100% recall (ivermectin–Cochrane, colchicine and vitamin-D). The fourth, ivermectin–BMC, recovered 20 of 23 (87%). Which trials were surfaced, screened in, and full-text-retrievable is tabulated per trial in Supplement S1, and the per-corpus forest plot is Fig. 2.

Two features of the design temper how this recall should be read. First, the positive counts are small — 5 to 23 per corpus — so the per-corpus intervals are wide. Colchicine, for instance, runs from 0.57 to 1.00, and the three 100% values carry Wilson lower bounds of only 0.57–0.74. The aggregate is the more stable summary, which is why we lead with it. Second, these golds are not a from-scratch eligibility sample. Each had already passed the original reviewers' full two-stage screen, full-text read included, so the figure measures agreement with a human-validated decision and is plausibly optimistic relative to screening a fresh literature.

Specificity ranged 71–85% and the pass-fraction φ — the share of records advanced to full text — ranged 17–31%, the selective behaviour the cost argument below turns on. Precision was low (10–26% at the strict operating point), but this is the arithmetic of screening at 1.7–9.2% prevalence rather than a defect: when fewer than one record in ten is eligible, even an excellent filter advances many false positives. Adding the `unsure` tier (the lenient gate) bought no extra recall — after re-screening, no gold fell in the `unsure` bin — while inflating the pass-fraction, so we adopt the strict gate as the operating point throughout.

3.2 Two-stage screening is cheaper across the realistic cost-ratio range

Whether the abstract gate earns its place comes down to a single comparison. In units of one abstract read, single-stage screening pays the full-text cost c on every record, whereas two-stage pays 1 for the abstract screen on every record plus c only on the fraction φ it advances — a total of $1 + \varphi c$. The gate therefore pays for itself whenever c exceeds the break-even ratio $c^* = 1/(1-\varphi)$, a threshold fixed entirely by how selective the gate is. With the measured pass-fractions that break-even is low: $c^* = 1.20$–1.45 across the four corpora. A real full-text read costs far more than this: we measured ≈2–3× an abstract for the open-access HTML/XML of these trials, and OCR-heavy pipelines run higher still. Two-stage was therefore the cheaper strategy in every corpus, by 29–43% (Fig. 3). The decisive point is not the size of

the saving but its robustness: any realistic cost ratio (≥1.5) already clears a break-even of ≈1.45, so the verdict does not flip anywhere in the plausible range. The regime map (Fig. 1) places all four corpora well inside the two-stage-cheaper region, and the full prevalence × cost-ratio sweep behind it is given in Supplement S4.

The break-even algebra is established group-testing theory, not a contribution of this paper; what we add is the measured input that drives it — the empirical pass-fraction on real clinical pools — and the observation that, given that pass-fraction, the cost ranking is insensitive to read-cost. Two measurement choices make this a conservative reading. Our cost ratio counts input tokens only, excluding output tokens and the retrieval/conversion overhead we did not instrument, so the true ratio is higher than measured — and a higher ratio only widens the two-stage advantage. For external calibration, the OCR-heavy pipeline reported by AgentSLR sits at the expensive end of the range (direct full-text at ≈2.3× runtime, OCR cost rising US$36.6→US$303.2, an ≈8× increase derived from their figures). Yet it reaches the same qualitative verdict as our cheap open-access corpora.

The cost ranking is a statement about spend, not about whether either strategy meets a recall target; we keep the two separate (§3.3). Because no separate full-text-screening pass was run, the recall side of the comparison assumes that an advanced record's eligibility is settled correctly at full text ($r_{FT} = 1$); it is therefore a modelled comparison, and a measured full-text screen — ideally with an r_{FT} sweep over [0.7, 1.0] — is the natural next step. Under that assumption the two strategies differ only where the gate errs. In the three corpora where it missed no gold, two-stage is both cheaper and no worse on recall, while in ivermectin–BMC, where it missed 3 of 23, two-stage trades a 13-point recall loss for the saving. One scope limit remains: our reconstructed prevalence (1.7–9.2%) is compressed relative to the published 0.3–7.3% range, so the lowest-prevalence regime — where two-stage should dominate most — rests on the model rather than on data points of its own.

3.3 Full-text access, not compute, caps achievable recall

The more consequential constraint turned out to lie downstream of the gate altogether. Final inclusion requires reading the full text, so achievable recall is capped not by the gate's accuracy but by how many eligible trials can actually be obtained. Across the four reviews, only 35 of 46 included trials (76%, 95% CI 62–86) were openly retrievable. The shortfall was not random: the unretrievable trials were disproportionately the large, high-impact, paywalled ones — López-Medina in JAMA, I-TECH, ACTIV-6 — which Supplement S1 lists individually. Under $r_{FT} = 1$, single-stage recall is bounded by this retrievable fraction $\beta_{gold} \approx 0.76$ and two-stage by $r_{gate} \cdot \beta_{gold}$, so neither strategy can reach 0.95 recall by AI-reading alone. The ceiling is access, not compute or the gate (Fig. 4).

That ceiling is, if anything, optimistic. These corpora were chosen partly because pandemic open-access mandates made COVID-19 trials unusually retrievable; the broader literature is less accessible — roughly 43% of the RCTs in recent Cochrane reviews sit behind a paywall (Uribe-Cavero et al. 2026), and only a minority of the scholarly record is openly available (Piwowar et al. 2018; Boudry et al. 2019) — so a non-pandemic corpus would likely show a lower obtainability ceiling, not a higher one. The practical implication is direct: effort spent making more full text obtainable would raise achievable recall more than effort spent making the screener cheaper or marginally more accurate.

3.4 The confidence score reflects certainty, not relevance

The screener attaches a numeric confidence (0–1) to every decision, which invites an obvious refinement — escalate low-confidence records, or route on a confidence threshold rather than the binary call. The data do not support it. The raw confidence score did not separate included from excluded trials at all (AUC 0.51, indistinguishable from chance), whereas the binary include/exclude decision did (decision-signed AUC 0.90); even within the advanced set, confidence ranked the true golds only weakly above the false positives (AUC 0.66). The reliability table behind this is Supplement S3. Mean confidence sits near 0.98 while gold prevalence is ≈4% — a large expected calibration error (0.90) — with the gold rate essentially flat across confidence deciles. The score, in other words, reports the model's certainty in its own decision, not an estimate of relevance, so a confidence-threshold cascade has little to add beyond the decision itself (Fig. 5). This is a single-model observation, but it has a concrete consequence for the adaptive-routing idea: on this signal, per-record routing collapses to the strict-versus-lenient threshold choice already made in §3.1.

3.5 The recall failures were confident, not uncertain

A natural hope is that when the gate does miss an eligible trial it at least signals doubt — landing the record in the `unsure` bin, where a cautious reviewer would catch it. That is not what happened. All three missed golds (in ivermectin–BMC: Babalola, Biber, Niaee) were confident exclusions, assigned 0.74–0.99 confidence, and each failed on a single PICO element — the comparator — rather than on overall relevance; none was routed to `unsure`. These golds are a different trio from the unretrievable paywalled trials of §3.3 (López-Medina, I-TECH, ACTIV-6). The per-question judgements for these three records are in Supplement S6. The human heuristic of escalating uncertain records therefore does not help here, because the model expressed no uncertainty at the point of error. We report this as a single observed pattern, not a rate: n = 3, one corpus, one PICO element. It is plausibly a single eligibility-encoding quirk rather than a general failure mode, and separating quirk from mode would take more corpora and a second screener model.

4. Conclusion

Across four COVID-era, low-prevalence clinical-RCT corpora, an AI abstract gate was the cost-efficient screening choice in every case. The full-text-to-abstract cost ratio did not change that verdict anywhere in its realistic range. The binding limit on "let the AI read every full text" was not compute but full-text access. Roughly a quarter of the included trials were not openly retrievable, and the missing ones were disproportionately the high-impact paywalled trials that matter most. Two empirical contributions follow. The first is the measured pass-fractions that locate these reviews on the inherited cost break-even. The second is the coupling of full-text obtainability to achievable recall. Two supporting analyses round these out: the gate's confidence score is a certainty signal rather than a relevance one, and its rare misses are confident rather than hesitant. These results come from one production pipeline and one model, evaluated on four related corpora, so they are best read as a proof-of-concept rather than a general benchmark. Confirming them will take a measured full-text-screening step (with an r_{FT} sensitivity sweep), a second screener model, and evaluation on shared open screening benchmarks such as SYNERGY, CLEF-TAR or the AgentSLR dataset.

Figures

(PNG, 300 dpi, in `figures/`; `make_figures.py` computes values from the CSVs.)

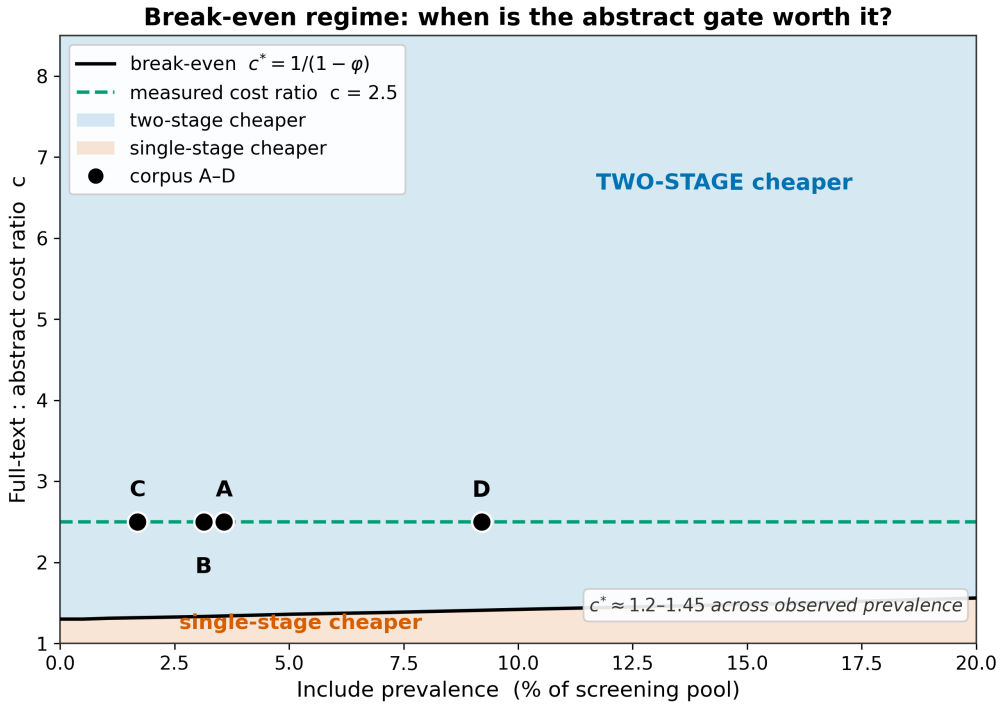

Fig. 1 — Break-even regime map. Cost-minimising strategy over include prevalence (x) and cost ratio c (y); boundary c* = 1/(1−φ); the four corpora at their prevalence and measured c ≈ 2.5, above c* ≈ 1.2–1.45 throughout.

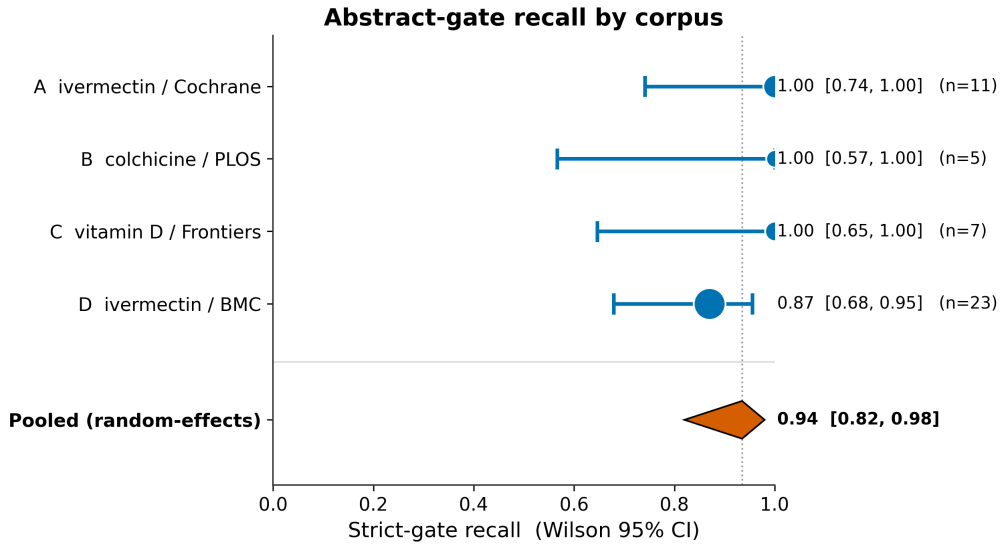

Fig. 2 — Abstract-gate recall. Per-corpus strict-gate recall with Wilson 95% CIs (A/B/C = 1.00, D = 0.87); aggregate pooled 43/46 = 93.5% [82–98] (random-effects 89% shown for comparison).

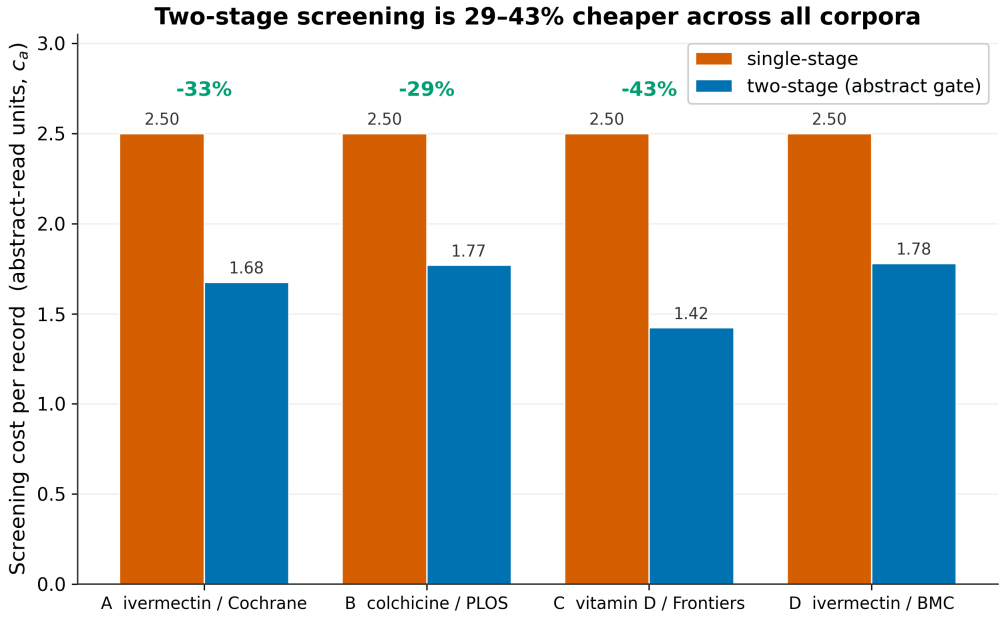

Fig. 3 — Single- vs two-stage cost. Per-record cost (abstract-read units): single-stage 2.50 vs two-stage; 29–43% cheaper.

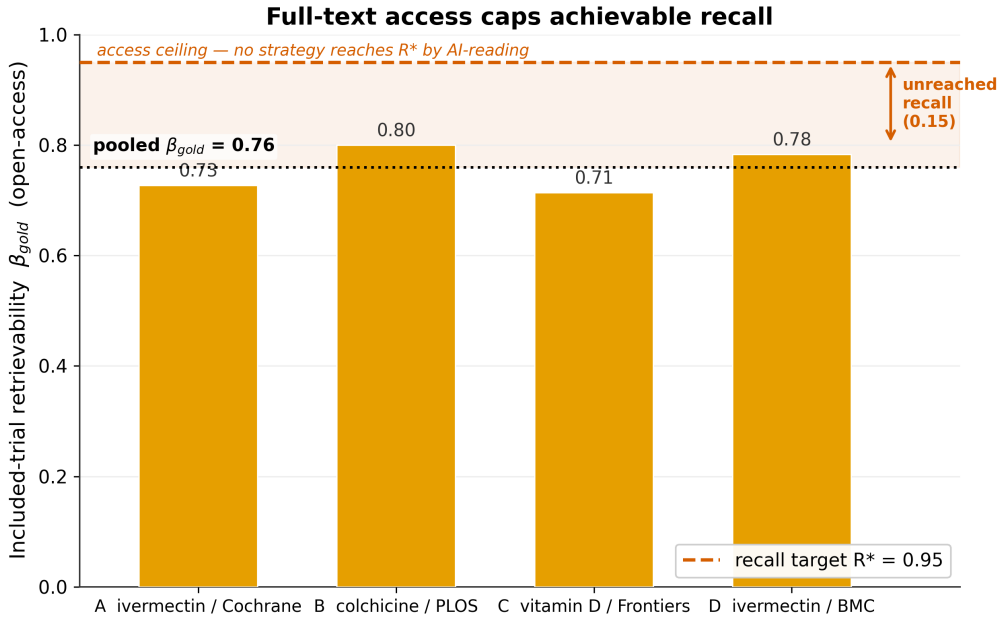

Fig. 4 — Full-text access ceiling. Per-corpus β_{gold} (8/11, 4/5, 5/7, 18/23) vs the 0.95 target; pooled β_{gold} = 35/46 = 0.76.

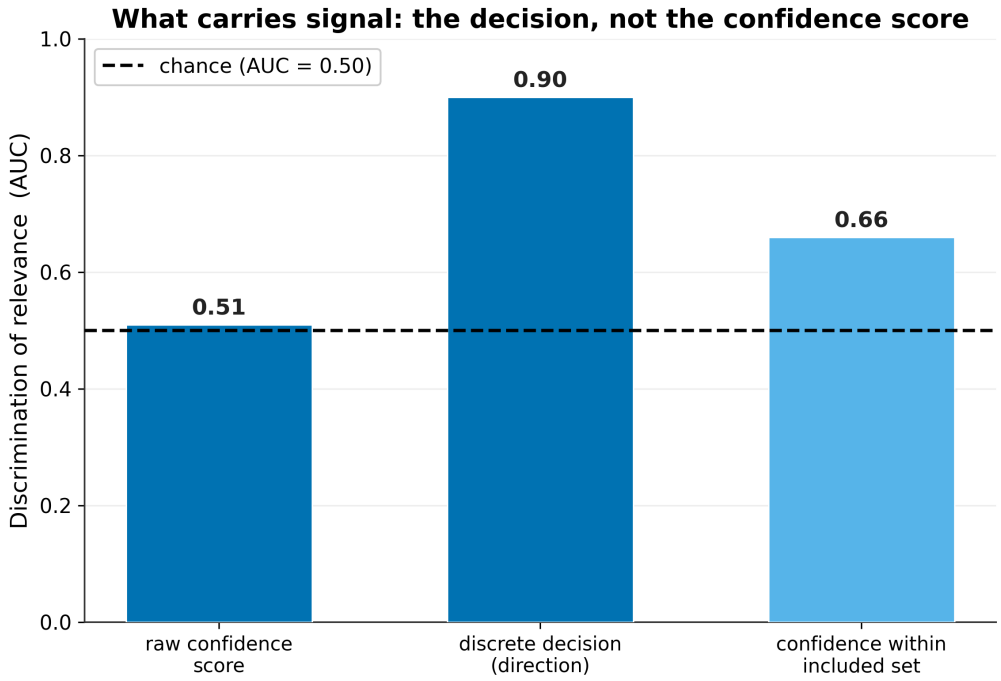

Fig. 5 — Confidence vs decision discrimination. Relevance-discrimination AUC: raw confidence 0.51, decision-signed 0.90, within-included 0.66; chance 0.50.

References

- Boudry C, Alvarez-Muñoz P, Arencibia-Jorge R, Ayena D, Brouwer NJ, Chaudhuri Z, et al. Worldwide inequality in access to full text scientific articles: the example of ophthalmology. *PeerJ* 2019;7:e7850. doi:10.7717/peerj.7850.
- Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Systematic Reviews* 2024;13:158. doi:10.1186/s13643-024-02575-4.
- Dorfman R. The detection of defective members of large populations. *Annals of Mathematical Statistics* 1943;14(4):436–440. doi:10.1214/aoms/1177731363.
- Janisch J, Pevný T, Lisý V. Classification with costly features using deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 2019;33(01):3959–3966. doi:10.1609/aaai.v33i01.33013959.
- Jarvis C, Gregory JM, Mortensen-Hayes A, McFarland M. Borrowing trouble? The impact of a systematic review service on interlibrary loan borrowing in an academic health sciences library. *Journal of the Medical Library Association* 2021;109(1):84–89. doi:10.5195/jmla.2021.1005.
- 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. *Research Synthesis Methods* 2024;15(4):616–626. doi:10.1002/jrsm.1715.
- Longford NT. Classification in two-stage screening. *Statistics in Medicine* 2015;34(25):3281–3297. doi:10.1002/sim.6554.

- Padarha S, Kearns RO, Naidoo T, et al. Evaluating AI-based Scientific Knowledge Synthesis with Epidemiological Systematic Reviews. arXiv:2603.22327 [cs.IR], v1 2026-03-20, v2 2026-06-04. (Harness: "AgentSLR".)
- Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, Farley A, West J, Haustein S. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* 2018;6:e4375. doi:10.7717/peerj.4375.
- Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews* 2016;5:140. doi:10.1186/s13643-016-0315-4.
- Trapeznikov K, Saligrama V. Supervised sequential classification under budget constraints. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 2013;31:581–589.
- Tran V-T, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, Sommer I, Alebouyeh F, Afach S, Meerpohl JJ, Ravaud P. Sensitivity and specificity of using GPT-3.5 Turbo models for title and abstract screening in systematic reviews and meta-analyses. *Annals of Internal Medicine* 2024;177(6):791–799. doi:10.7326/M23-3389.
- Uribe-Cavero LJ, Brañez-Condorena A, Parra-Huaroto AM, Vera-Maccha PJ, Taype-Rondan A. Clinical evidence behind a paywall: an analysis of randomised clinical trials included in Cochrane reviews. *Learned Publishing* 2026;39(2). doi:10.1002/leap.2058.
- van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence* 2021;3(2):125–133. doi:10.1038/s42256-020-00287-7.
- Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI '12)* 2012:819–824. doi:10.1145/2110363.2110464.

Data & code availability

We release the derived per-record dataset for all four pools — screening decisions, confidence scores, per-PICO-question judgements, token counts, and full-text retrievability flags, keyed by PMID/DOI (`confidence_per_record.csv` and the per-corpus metric/cost CSVs) — together with the analysis and figure code, which computes all reported statistics from those files. Raw third-party abstracts and full text are withheld. The screening model family/version and the eligibility/screening configuration are documented in the release to support independent reproduction.