

Stopping Is a Conditional Decision: Corpus Qualification, Bayesian Sequential Stopping, and the Limits of Early Termination in Scholarly Literature Screening

Mohammad Pasha✉

mohd.pasha@outlook.com

SR University

Mohammed Ali Shaik✉

SR University

Research Article

Keywords: scholarly literature screening, corpus qualification, Bayesian sequential stopping, information retrieval, bibliographic corpora, evidence synthesis

Posted Date: September 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10938971/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Stopping Is a Conditional Decision: Corpus Qualification, Bayesian Sequential Stopping, and the Limits of Early Termination in Scholarly Literature Screening

Mohammad Pasha^{a,*}; Mohammed Ali Shaik^b

^a Postdoctoral Research Fellow, CSE Department, SR University, Ananthasagar, Warangal 506371, Telangana, India ^b Associate Professor, School of Computer Science and Artificial Intelligence, SR University, Ananthasagar, Warangal 506371, Telangana, India

* Corresponding author: Mohammad Pasha, mohd.pasha@outlook.com; ORCID: 0000-0002-0596-1682
Mohammed Ali Shaik: niharali@gmail.com; ORCID: 0000-0002-5520-0830

Abstract

The construction of a scholarly literature corpus and the decision to stop screening it are distinct methodological problems. We formalize this distinction through a conditional architecture in which corpus qualification precedes within-corpus stopping. Adaptive Research Scoping (ARS) is a domain-general corpus-qualification layer that evaluates alignment, reference-set coverage, contamination, and contextual representation through a prospectively declared conjunctive gate. Only corpora that satisfy this gate enter Bayesian Sequential Stopping (BSC), a closed-corpus Beta-Binomial decision model with a finite-horizon loss recursion. Across controlled simulations, labelled review corpora, semantic-readiness tests, information-horizon analyses, and an exploratory retrieval-ordering study, posterior calibration, semantic redundancy, and improved ranking did not reliably imply safe early termination. In a final benchmark-qualified validation using ten labelled review corpora from the SYNERGY screening-data collection, with all benchmark-known relevant records present in each candidate corpus, the primary operating condition achieved 95% posterior-predictive interval coverage of 0.978 but a premature-stop rate of 0.915 and mean benchmark-relative recall of 0.165. A more conservative condition raised mean benchmark-relative recall to 0.979 but reduced mean work saving to 1.25%. These results show that posterior-predictive calibration, corpus qualification, and ranking benefit are not equivalent to stopping safety. The main contribution is an auditable separation of bibliographic corpus qualification, information ordering, and stopping admissibility for quantitative analysis of scholarly literature workflows.

Keywords: scholarly literature screening; corpus qualification; Bayesian sequential stopping; information retrieval; bibliographic corpora; evidence synthesis

Introduction

Scoping reviews and mapping reviews are designed to characterize the breadth, concepts, evidence clusters, and gaps in a literature rather than necessarily estimate a single intervention effect. PRISMA-ScR frames scoping reviews as systematic evidence-mapping exercises, while current JBI guidance emphasizes a priori scope, transparent eligibility criteria, systematic searching, and structured presentation of the evidence base. (Tricco et al., 2018; Peters et al., 2024)

Search quality is already a core concern of evidence-synthesis methodology. PRISMA-S formalizes transparent and reproducible reporting of database searching, supplementary searching, limits, deduplication, and search updating (Rethlefsen et al., 2021). Related scientometric work shows that bibliographic sources differ in reference coverage, metadata completeness, and downstream indicator behavior, so corpus construction is itself a methodological choice rather than a neutral preprocessing step (Culbert et al., 2025; Scheidsteger et al., 2025). The contribution here is therefore narrower: most stopping-rule evaluations condition on an already assembled candidate corpus, whereas ARS makes corpus adequacy an explicit admissibility interface before within-corpus stopping is interpreted.

Accordingly, the principal novelty claim is not that search quality has been overlooked, nor that the proposed architecture provides a universally successful stopping rule. The contribution is a boundary-condition argument: stopping-rule calibration inside a retrieved set should not be interpreted as external stopping safety unless the evidentiary status of that set is assessed separately. The corpus-qualification layer operationalizes that admissibility interface, and the experiments then test whether corpus adequacy, posterior calibration, semantic redundancy, additional information horizon, and improved ranking are individually sufficient to certify useful early termination. The study therefore treats negative results as informative constraints on when automation may support a stopping claim, rather than as evidence that a more aggressive threshold should simply be tuned post hoc.

Large language models are also emerging as screening assistants. Recent studies have evaluated LLMs for title/abstract relevance classification and report potentially substantial workload reductions, but also model-, prompt-, criterion-, and dataset-dependent sensitivity-specificity trade-offs. (Dennstädt et al., 2024; Delgado-Chaves et al., 2025; Oami et al., 2025) These systems address an important but distinct question: whether a model can classify or prioritize records efficiently. The present study asks whether the evidentiary conditions for a defensible stopping claim have been met. An LLM relevance score, explanation, or confidence signal is therefore not treated here as evidence of corpus completeness or as an external certificate that screening may safely terminate.

This breadth creates a workload problem. Search strategies that are intentionally sensitive may retrieve hundreds or thousands of records even when the final included evidence base is small. Historical methodological surveys have reported scoping reviews with as few as one included study, and practical guidance notes that very small evidence bases can limit the value of a scoping review while still requiring rigorous search and selection procedures. (Gottlieb et al., 2021)

The practical motivation arose from resource-constrained scholarly-literature reviews, including undergraduate computer-science and engineering research supervision, where literature work must fit bounded project schedules while still supporting defensible synthesis. AI-assisted search and summarization can make a shallow survey faster and more polished without making it more evidentially rigorous. This use case shaped the requirements for interpretable diagnostics, explicit failure states, reproducible decisions, and low computational burden. It is an intended application, not an evaluated education intervention: no student cohort was randomized, and no claim is made here about grades, learning outcomes, review time, publication rates, or classroom effectiveness.

Machine-learning-assisted priority screening addresses the workload by ranking likely relevant records earlier. ASReview and related systems have made this approach operationally practical, and recent evaluations show that realized workload reduction depends on the review corpus, prevalence, and screening model (van de Schoot et al., 2021; Spiero et al., 2025). Ranking alone, however, does not answer when screening may stop.

The stopping problem has become a distinct methodological literature. Statistical approaches seek confidence that a recall target has been reached; heuristic procedures combine minimum screening fractions, key-paper checks, and runs of irrelevant records; prevalence-estimation approaches have also been proposed for technology-assisted review; and recent empirical evaluations show strong dataset, prevalence, algorithm, and ranking dependence (Callaghan & Müller-Hansen, 2020; Boetje & van de Schoot, 2024; König, 2024; Bron et al., 2025).

Recent critiques and large benchmarks converge on a demanding requirement: a stopping criterion should communicate the risk of missing studies, should not require outcome-informed tuning, and should be evaluated across heterogeneous datasets and ranking conditions. In a 2026 evaluation of 15 stopping methods over 81 real-world datasets, most methods either missed relevant records or stopped too conservatively to realize the available work savings. (Callaghan et al., 2024; Repke et al., 2026)

We argue that this literature still leaves an upstream assumption insufficiently explicit. Stopping methods are normally applied to the retrieved and often ranked set of records. They therefore answer a within-corpus question: how much relevant evidence likely remains in the records currently available? They cannot by themselves determine whether the search retrieved the right literature in the first place. For scholarly-literature workflows that make claims about the structure or content of a field, that distinction matters because database and metadata coverage can alter the evidentiary universe available for downstream analysis (Culbert et al., 2025; Scheidsteger et al., 2025). A precisely calibrated stopping rule applied to an inadequately covered corpus can therefore produce a precise answer to the wrong evidentiary universe.

The study therefore evaluates a conditional decision architecture comprising Adaptive Research Scoping (ARS) for corpus qualification and Bayesian Sequential Stopping (BSC) for within-corpus termination. The central spine is:

corpus qualification -> within-corpus BSC boundary testing -> benchmark-qualified external falsification.

The later semantic-readiness, information-horizon, and retrieval-ordering analyses are not co-equal attempts to find a successful stopper; they test three alternative explanations for failure: perhaps paper counts are the wrong information unit, perhaps five relevant papers are too early to judge redundancy, or perhaps poor ordering delays useful evidence. The central question is therefore not whether one algorithm can always stop early, but which conditions must hold before stopping has evidential meaning. Fig. 1 summarizes this three-stage conditional architecture.

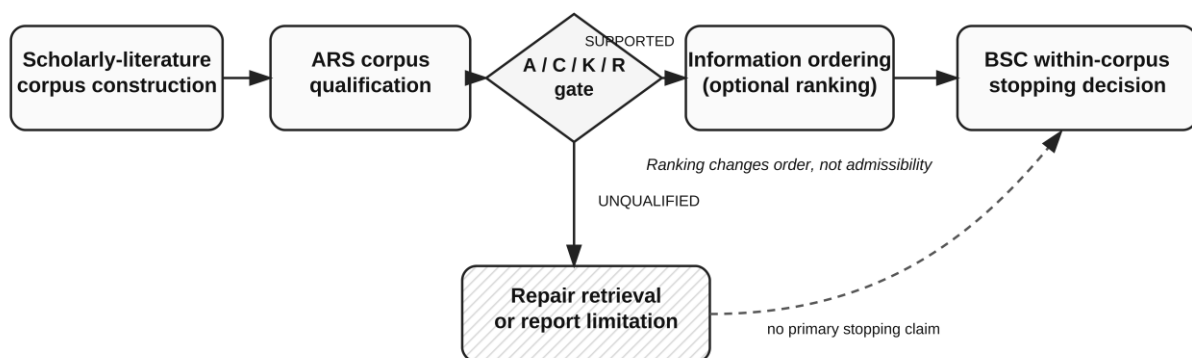


Fig. 1 ARS-BSC conditional architecture showing corpus construction, ARS qualification, optional ranking, and conditional BSC

Methods

Study design, preregistration, and evidentiary hierarchy

The study used a staged prospective design to distinguish confirmatory, prospectively frozen, exploratory, and post-outcome descriptive analyses. The domain-general ARS-BSC core protocol was prospectively registered before its confirmatory analyses. Subsequent semantic-readiness and information-horizon tests were prospectively frozen with SHA-256 manifests before execution on new external datasets. Retrieval-ordering analyses were conducted later as post-confirmatory exploratory stress tests. The final benchmark-qualified BSC validation was separately frozen before outcome inspection on its selected panel (2 September 2026, 12:01 Asia/Kolkata; freeze-package SHA-256: e91b95c8f8d12af6c0739583f8fc7c0e1d534eb526875a52e8b5853a17994198). Local cryptographic freezes establish tamper-evident chronology and file identity but are not equivalent to independent registry timestamps.

Full version chronology, hypothesis mapping, comparator accountability, freeze status, and protocol deviations are reported in Online Resource 1. Version labels are retained in the main text only where they determine evidentiary status.

Conditional architecture and estimands

The model sequence was selected to isolate a chain of increasingly specific failure mechanisms rather than to maximize algorithmic complexity. The intended application—resource-constrained scholarly-literature screening—favored components that are transparent enough to audit, computationally inexpensive to reproduce, and sufficiently interpretable that assumption failures remain identifiable rather than being absorbed into an end-to-end score.

Adaptive Research Scoping (ARS) is the corpus-qualification layer of the architecture and functions as an admissibility interface before within-corpus stopping. It is evaluated first because stopping is downstream of retrieval: if relevant evidence is absent from the candidate corpus, no stopping rule operating exclusively within that corpus can recover it. ARS therefore evaluates four non-substitutable properties—Alignment, Coverage, Contamination, and Context Representation—rather than collapsing corpus quality into a single relevance score. The conjunctive gate prevents a strong value on one dimension from compensating for a structurally different failure on another.

Bayesian Sequential Stopping (BSC) addresses the conditional problem that remains after retrieval has been frozen: estimating how much relevant evidence may remain within a corpus of known size. A Beta-Binomial model was selected because, under conditional exchangeability, it provides exact conjugate updating, direct posterior-predictive statements about remaining relevant records, and stable computation in sparse settings without requiring large training sets. The model is intentionally simpler than neural or rank-dependent alternatives because interpretability and exact updating are design requirements for an auditable stopping decision.

A Bellman decision rule converts posterior prediction into a finite-horizon action. Let $\rho=c_s/c_m$ denote the screening-cost to missed-relevant-record loss ratio. This quantity encodes an operating preference; it is not a recall target. Consequently, the loss-optimal stopping action under a given ρ need not satisfy an externally imposed recall criterion such as 0.80. Linear loss is used as a transparent operating model rather than as a claim that all missed studies have equal scientific value (Bellman, 1957).

Semantic analyses were introduced only after the paper-count model exposed a distinct question: relevant papers need not be equivalent units of information. Online semantic clustering therefore tested whether conceptual novelty declined before paper-level relevance was exhausted. Readiness and information-horizon analyses then tested whether an early semantic signal was sufficiently stable to support a stopping decision.

Retrieval ordering was examined separately as an explanatory stress test. If poor stopping performance resulted mainly from relevant evidence arriving too late, improved ranking should increase early discovery. BM25 and semantic reranking therefore tested information ordering without changing the stopping estimand. This separation preserves the distinction between ranking benefit and stopping admissibility (Robertson & Zaragoza, 2009). Table 1 summarizes the purpose and principal limitation of each component.

Table 1 Why each model/component was studied and the failure mode it was intended to isolate

Model / component	Question it was designed to answer	Why this model	Primary limitation tested
ARS conjunctive gate	Do we have the right corpus before asking when to stop?	Transparent non-substitutable diagnostics for alignment, coverage, contamination and context	Qualification validity depends on the quality of the declared reference/coverage evidence.
Beta-Binomial BSC	How many relevant records may remain inside a fixed corpus?	Exact conjugate small-N predictive count model; auditable and low-compute	Conditional exchangeability; fixed retrieved corpus; not a model of the whole field
Bellman loss rule	When is further screening worth its cost?	Turns posterior prediction into an explicit finite-horizon decision using $\rho=c_s/c_m$	Linear constant costs simplify scientific utility
Online semantic clustering	Are new relevant papers still adding new conceptual content?	Tests conceptual redundancy separately from relevance counts	Whole-document similarity is a proxy, not verified claim novelty
Readiness / information horizon	Is there enough observed relevant evidence to trust a semantic stop?	Tests whether early-prefix information predicts later safe savings	External transport was weak and the m=10 advantage did not replicate
BM25 / semantic reranking	Is stopping failure mainly caused by poor evidence ordering?	Strong, interpretable retrieval baseline plus semantic ordering stress test	Better early retrieval did not imply stopping safety

Let R^* denote the relevant reference set used for evaluation, D the frozen candidate corpus, and $S \subseteq D$ the records screened when stopping occurs. Reference-set-relative recall factorizes into corpus coverage and within-corpus screening recall:

$$Recall_{R^*} = C_{R^*} \times Recall_{within}, \quad C_{R^*} = \frac{|D \cap R^*|}{|R^*|}, \quad Recall_{within} = \frac{|S \cap R^*|}{|D \cap R^*|}$$

Therefore $Recall_{R^*} \leq C_{R^*}$. No stopping rule operating exclusively inside D can recover reference-set items absent from D . This identity is the formal reason ARS precedes BSC: BSC can be internally calibrated and still have poor reference-set-relative validity when corpus coverage is low. The term 'field recall' is reserved for simulated settings in which the complete generating truth is known; empirical constructed standards are reported as reference-set-relative recall, and the final external validation is reported as benchmark-relative recall.

Adaptive Research Scoping as a corpus-qualification layer

ARS is the corpus-qualification layer of the architecture, formalized as an admissibility operator applied before BSC. Let Q denote the frozen review formulation, D the retrieved candidate corpus, and $\Theta_Q=(\tau_A, \tau_C, \tau_K, \tau_R)$ the prospectively declared operating thresholds. The corpus state is $X(D, Q)=(A, C, K, R)$, where the four dimensions represent alignment, coverage, contamination, and context representation. Qualification is defined as:

$$Qualified(D, Q, \theta_Q) = \mathbf{1}\{A \geq \tau_A \wedge C \geq \tau_C \wedge K \leq \tau_K \wedge R \geq \tau_R\}$$

A denotes alignment with the research question; C denotes coverage relative to an independently defined reference standard or other prospectively declared coverage evidence; K denotes contamination/noise, for which lower values are preferable; and R denotes representation of a context-defining dimension. The conjunctive form is deliberate: a highly aligned corpus can still be unusable if it misses key evidence, and a clean corpus can still be unusable if a required context is absent. The dimensions are fixed by the architecture, whereas their operational measures and thresholds are context-parameterized and must be declared prospectively to prevent post-hoc qualification.

For the frozen YOLOv8/edge-device demonstration, the operating vector was $\tau_A=0.50$, $\tau_C=0.40$, $\tau_K=0.25$, and $\tau_R=0.40$. Alignment was operationalized by the frozen query-pool alignment diagnostic (QPFS), coverage by relative reference-set recall, contamination by the frozen contamination estimate, and Context Representation by EdgeCoverage.

For corpus D,

$$QPFS(D) = |D|^{-1} \sum_i I[YOLO_i \text{ AND } OD_i \text{ AND } (EDGE_i \text{ OR } LIGHT_i)],$$

whereas

$$EdgeCoverage(D) = |D|^{-1} \sum_i I[EDGE_i]$$

Thus, QPFS measured the fraction of records jointly matching the detector family, object-detection context, and at least one edge/lightweight term; EdgeCoverage measured the fraction matching at least one frozen edge-context term. Exact lexical dictionaries are reported in Online Resource 1. These are demonstration-specific operationalizations of the ARS dimensions, not universal corpus-quality measures. The constructed reference set R^* contained 30 independently curated items and was used only after retrieval decisions were complete. Because R^* was not a complete field census or a probability sample of all field-relevant studies, coverage is reported only as relative recall against this fixed constructed standard.

Corpus-repair policy and terminal states

At iteration t, ARS applies

$$EXPAND \rightarrow ALIGN \rightarrow ASSESS \text{ MARGINAL CONTRIBUTION} \rightarrow KEEP/QUARANTINE/DISCARD$$

while leaving the registered review formulation Q0 unchanged. ARS terminates after at most three iterations or when no registered action targets the remaining diagnostic defect. The terminal states are

$$SUPPORTED, NEAR, \text{ and } UNQUALIFIED. SUPPORTED$$

which permits primary BSC analysis; NEAR permits sensitivity analysis only with an explicit limitation; UNQUALIFIED blocks primary BSC.

Illustrative qualification case: YOLOv8/edge-device corpus

The YOLOv8/edge-device demonstration began with 663 OpenAlex records. A single confirmatory citation-network repair used ten operational seeds recovered from the existing corpus, retrieved backward references and forward citations, canonicalized candidates, and applied the frozen KEEP/QUARANTINE/DISCARD rules before R^* evaluation. Missing R^* titles were never used as hidden queries or seeds.

Within-corpus Bayesian Sequential Stopping

BSC addresses the conditional problem that remains after retrieval has been frozen: how much relevant evidence may remain inside a fixed candidate corpus D of known size $N=|D|$, N is the size of the retrieved corpus, not the size of the scientific literature. Let Y_i be the relevance indicator for screened record i and assume conditional exchangeability, $Y_i | \pi \sim \text{Bernoulli}(\pi)$, with prior $\pi \sim \text{Beta}(\alpha_d, \beta_d)$. After n screened records with k relevant observations:

$$\pi | \mathcal{D}_n \sim \text{Beta}(a_n, b_n), \quad a_n = \alpha_d + k, \quad b_n = \beta_d + n - k$$

For $M=N-n$ remaining records, the predictive count R_n of relevant records remaining follows a Beta-Binomial distribution. Its posterior-predictive mean and variance, together with the probability that no relevant records remain, are:

$$R_n | \mathcal{D}_n \sim \text{BetaBinomial}(M, a_n, b_n), \quad \mathbb{E}[R_n | \mathcal{D}_n] = M \frac{a_n}{a_n + b_n}$$

$$\text{Var}(R_n | \mathcal{D}_n) = \frac{M a_n b_n (a_n + b_n + M)}{(a_n + b_n)^2 (a_n + b_n + 1)}, \quad \Pr(R_n = 0 | \mathcal{D}_n) = \frac{B(a_n, b_n + M)}{B(a_n, b_n)}$$

The zero-remaining probability is reported as a secondary diagnostic rather than used as an additional mandatory stop criterion.

The decision rule adds explicit screening and omission costs. Let $c_s > 0$ be the cost of screening one additional record and $c_m > 0$ the loss assigned to leaving one relevant record unscreened; $\rho = c_s / c_m$ is the operating characteristic. The stopping loss and next-record predictive relevance probability are:

$$L_{\text{stop}}(n, k) = c_m \mathbb{E}[R_n | \mathcal{D}_n], \quad p_{nk} = \frac{\alpha_d + k}{\alpha_d + \beta_d + n}$$

For a finite horizon, the continuation loss and Bellman recursion are:

$$L_{\text{continue}}(n, k) = c_s + p_{nk} V(n+1, k+1) + (1 - p_{nk}) V(n+1, k) \\ V(N, k) = 0, \quad V(n, k) = \min\{L_{\text{stop}}, L_{\text{continue}}\}, \quad \text{STOP} \Leftrightarrow L_{\text{stop}} \leq L_{\text{continue}}$$

The loss ratio ρ is an operating preference, not a recall target. The Bellman policy minimizes expected screening-plus-omission loss under the stated model; it does not directly optimize or constrain the prespecified benchmark-relative recall criterion of 0.80. In a local one-step intuition, screening becomes less attractive as the posterior predictive probability of relevance approaches the screening-to-missing cost ratio, but the exact finite-horizon rule also includes the option value of future observations. The 80% recall threshold is therefore an external safety criterion used to evaluate the decision policy, not a quantity encoded directly in its loss function. No post-outcome recall-constrained rule was introduced to repair this mismatch.

Four assumptions delimit the interpretation. First, $\mathcal{D}$ is closed before BSC begins. Second, N is known because it is the size of $\mathcal{D}$. Third, relevance labels are conditionally exchangeable under the primary model; this assumption is satisfied by the random-permutation validation design but is generally not guaranteed under relevance-ranked or adaptively reordered screening streams. Fourth, residual scientific loss is represented linearly and screening cost is constant. These are modeling devices for the within-corpus decision problem, not assumptions that the scientific literature itself is fixed, exchangeable, or uniformly valuable. Ranked retrieval experiments are therefore treated as robustness analyses rather than evidence that the Bernoulli model is generatively correct under ranking. Fig. 2 summarizes the operational relationship between the ARS qualification gate and the BSC finite-horizon decision.

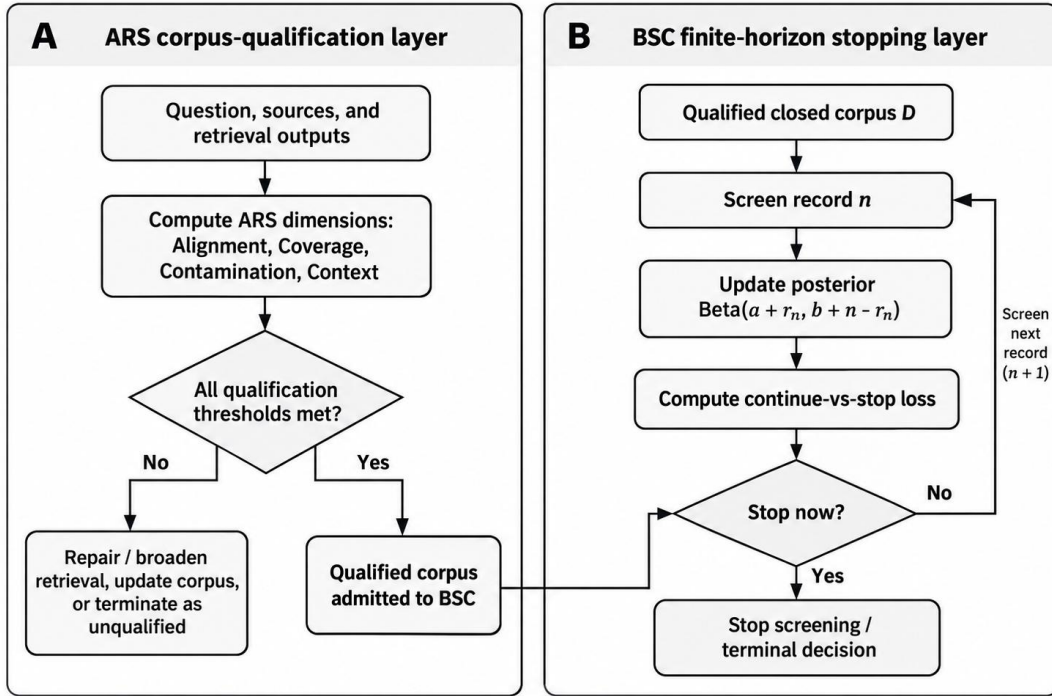


Fig. 2 Operational ARS-BSC methods schematic showing the ARS corpus-qualification gate (A) and BSC finite-horizon decision process (B)

Experimental evaluation framework

The preregistered core tests evaluated six questions: whether corpus coverage structurally bounds achievable recall when generating truth is known (H0); whether multidimensional qualification distinguishes induced corpus defects better than alignment alone (H1); whether controlled adaptive retrieval improves matched-budget performance (H2); whether contamination degrades stopping safety (H3); whether field-informed priors improve sparse-setting performance (H4); and

whether BSC exhibits a prespecified valid and useful operating region after qualification without post-hoc retuning (H5). Full hypothesis definitions are reported in Online Resource 1.

The corrected v1.5 qualification-boundary extension sampled 24 Latin-hypercube points over A,C,K,R, four corpus sizes (50, 100, 250, 500), four relevance prevalences (0.02, 0.05, 0.10, 0.25), two priors, and 500 Monte Carlo replicates per cell, yielding 768 primary cells and 384,000 runs. Cov95 and Cov80 denote empirical coverage rates of the corresponding posterior-predictive intervals for the number of relevant records remaining. A cell was BSC-valid only if $\text{Cov95} \geq 0.925$, $\text{Cov80} \geq 0.760$, premature stopping ≤ 0.05 , and mean truth-relative recall at stop ≥ 0.80 . A valid cell was additionally useful only if mean work saved ≥ 0.10 . Table 2 maps these evaluation questions to the corresponding analyses.

Table 2 Evaluation questions and analyses used to test alternative explanations

Evaluation question	Analysis used to answer it
Can multidimensional qualification distinguish structurally different corpus failures?	Controlled degradation and diagnostic-identifiability experiments.
Does retrieval repair improve a corpus enough to make stopping admissible?	Matched-budget retrieval and the frozen YOLOv8 citation-repair demonstration.
Does the closed-corpus BSC remain safe and useful after qualification?	Qualification-boundary simulations and labelled-review replay.
Are failures explained by the information representation rather than the decision rule?	Semantic-saturation and semantic-readiness analyses.
Are failures caused by judging readiness too early?	Prospectively frozen information-horizon analysis.
Are failures caused mainly by poor record ordering?	Post-confirmatory BM25 and latent-semantic retrieval-ordering stress test.
Does BSC work when benchmark-relative corpus loss is removed?	Prospectively frozen validation on ten SYNERGY/SYNERGY+ review corpora.

Benchmark-qualified external validation on SYNERGY review corpora

The final external validation was designed to isolate BSC from the strongest alternative explanation for earlier failure: inadequate corpus coverage. Ten previously unconsumed labelled review corpora from the SYNERGY/SYNERGY+ screening-data collection were selected using a deterministic rule frozen before BSC outcome inspection. Eligibility required an exact CSV, binary inclusion labels, 5–30 benchmark-known included records, $N \leq 1000$, and no prior use in model development. Exact dataset identities, source-file SHA-256 hashes, priors, ρ grid, random seed, 500-permutation replay plan, estimands, validity thresholds, usefulness threshold, and negative-result interpretation were prospectively frozen (2 September 2026, 12:01 Asia/Kolkata; freeze-package SHA-256: e91b95c8f8d12af6c0739583f8fc7c0e1d534eb526875a52e8b5853a17994198). Each complete labelled candidate set was used as `D_j`, and all records with `label_included=1` defined `R*_j`. Therefore `C_benchmark=1` by construction. This establishes completeness relative to the supplied benchmark labels only; it does not establish complete coverage of the scientific field.

Each review was replayed under 500 uniform random permutations with seed 42. The random-permutation design was chosen to align the empirical replay with the conditional-exchangeability assumption of the primary Beta-Binomial model, not to mimic a priority-ranked screening workflow. The primary prior was $\text{Beta}(1,1)$, with $\text{Beta}(1,19)$ as a prespecified sparse-prior sensitivity analysis. The frozen ρ grid was $\{0.01, 0.02, 0.05, 0.10, 0.20, 0.50, 1.00\}$, with $\rho=0.10$ primary. The validity rule was unchanged from v1.5: $\text{Cov95} \geq 0.925$, $\text{Cov80} \geq 0.760$, premature-stop rate ≤ 0.05 , and mean benchmark-relative recall at stop ≥ 0.80 ; usefulness additionally required mean work saving ≥ 0.10 .

Secondary validation and exploratory stress tests

Semantic representation, readiness, and information horizon

Semantic development began after the original core freeze and introduced a different estimand: conceptual expansion among relevant evidence rather than the paper-level remaining-relevant-record count of BSC. Development began on Wolters_2018 (5,019 candidates; 19 included papers), followed by external labelled reviews. Whole-document title/abstract/keyword representations were used. The online stopper used a stateless HashingVectorizer (2^{14} features, English stop words, unigram+bigrams, L2 normalization), online centroid formation, and a $\text{Beta}(1,1)$ posterior over new-cluster events. Setting B declared a new online cluster when maximum centroid cosine similarity was < 0.30 and stopped when posterior mean $P(\text{new cluster}) \leq 0.20$ for two consecutive relevant observations after a minimum prefix. Subsequent v1.6 semantic-readiness and v1.7

information-horizon tests were prospectively frozen before external evaluation; their small number of independent reviews limits generalization but does not make those tests retrospective.

The semantic representation was intentionally non-generative. Stateless hashing and TF-IDF were selected because they are deterministic, inexpensive, auditable at replay time, and introduce no prompt template, proprietary model version, decoding stochasticity, or hidden training-data dependence into the stopping experiment. This is a choice for experimental identifiability, not a claim that sparse lexical representations are semantically superior to modern LLM embeddings. An LLM-based screener or representation would add another adaptive layer whose model, version, prompt, and decoding choices could materially alter observed stopping behavior; such comparisons therefore require a separately frozen evaluation.

Final TF-IDF/agglomerative clusters ($k=3-8$, restricted to $k \leq N_{\text{relevant}}$) were constructed only for retrospective evaluation. They were invisible to the stopping algorithm. We therefore describe the outcome as coverage of evaluation-derived semantic clusters, not coverage of all true scientific concepts.

v1.6 prospectively froze Setting B and evaluated early semantic readiness after exactly five relevant papers. The primary early semantic redundancy index was $ESRI_5 = (5 - K_5)/4$, where K_5 is the number of online clusters among the first five relevant papers. The primary joint safe-savings outcome required paper recall ≥ 0.80 , mean evaluation-cluster coverage ≥ 0.90 , and work saved ≥ 0.10 . Four untouched datasets were selected and hashed before execution: Webers_2022 (8,078/84 candidate/included), Zakeri-Nasrabadi_2023b (1,720/46), Zanframundo_2022 (1,334/40), and Xu_2022 (960/21).

For v1.6 and v1.7, the external review-not the replay-is the independent generalization unit. Repeated random replays quantify sensitivity to screening order within each fixed review; they do not create hundreds of independent reviews. Accordingly, review-level results and leave-one-review-out discrimination are given priority, while pooled replay-level AUCs are treated as descriptive measures of ordering variability rather than population-level precision.

v1.7 froze an information-horizon hypothesis after development analyses suggested that five relevant papers might be too early. The same three-variable standardized logistic readiness model - $ESRI_m$, mean semantic similarity over relevant papers $2..m$, and early relevance discovery rate - was compared at $m=5$ and $m=10$. The primary estimand was $\Delta AUC = AUC_{10} - AUC_5$. Four new untouched datasets were selected and SHA-256 hashed before execution: Attai_2022 (1,726/135), Harms_2024 (840/48), Moran_2020 (1,992/39), and Tumkaya_2018 (481/39).

Retrieval ordering

A final post-confirmatory analysis asked whether poor information ordering explained conservative semantic stopping. Review titles, with generic review terms removed, were used as queries. The same four v1.7 datasets were ordered by random permutation, Okapi BM25 ($k1=1.5$, $b=0.75$), a TF-IDF/Truncated-SVD latent semantic ranker, and a BM25-first-stage/LSA reranker. The LSA reranker is an offline semantic proxy and must not be interpreted as a transformer cross-encoder. Retrieval was evaluated using average precision, recall after screening 10% and 20% of the corpus, and nDCG at 10%. The downstream Setting B stopping logic remained unchanged.

Comparator and protocol-accountability summary

The preregistered protocol specified consecutive-irrelevance rules ($k=10,20,50$), SAFE/Callaghan-type procedures where ranker outputs were available, Chao-type estimators when at least 10 relevant records had been observed, an evaluation-only 80%-recall oracle, and broader knee/fixed-budget comparator classes together with applicability rules. The primary YOLOv8 path terminated UNQUALIFIED, so BSC and its screening-stop comparators were protocol-defined not applicable on that path.

A complete comparator benchmark was not archived with the preregistered core analyses. After primary outcomes were known, only the unambiguously specified consecutive-irrelevance $k=\{10,20,50\}$ rules and the evaluation-only 80%-recall oracle were completed on the frozen v1.8 replay orders, explicitly as descriptive/non-confirmatory analyses. Comparator variants requiring new rank inputs, tuning, or implementation choices were not reconstructed post hoc. Accordingly, no empirical superiority claim is made for BSC over SAFE, Callaghan-type, Chao-type, knee, or fixed-budget procedures. Full comparator accountability is provided in Online Resource 1.

Simulation-based calibration (SBC) was preregistered, but no frozen SBC implementation/output was produced. This is treated as a protocol deviation; no SBC validation claim is made. Empirical 80%/95% posterior-predictive interval coverage is reported separately and is not presented as a substitute for SBC.

Full comparator applicability, protocol-accountability details, and the evidentiary chronology are reported in Online Resource 1.

Results

Corpus qualification and admissibility

Under fixed screening recall of 0.95, truth-relative recall equaled $C \times 0.95$, giving Pearson $r=1.000$ across $C=0.05-0.95$. This analytically supports H0 in the truth-known simulation: screening cannot recover relevant evidence that was never retrieved.

Across 9,000 synthetically perturbed corpus states, the four-dimensional gate achieved 0.993 supported/unqualified classification accuracy compared with 0.375 for alignment alone, while exact defect-set identification accuracy was 0.914. At the highest perturbation level ($\sigma=0.10$), combined classification accuracy remained 0.979 and macro defect F1 was 0.881. These results establish internal identifiability of the registered A/C/K/R defect structure under controlled perturbations, not external construct validity of ARS as a universal corpus-quality measure. Table 3 summarizes the core preregistered and prospectively frozen results.

Table 3 Core preregistered/frozen ARS-BSC results

Experiment	Primary result	Interpretation
H0 input dominance	truth-relative recall = $C \times 0.95$; $r(C, \text{truth-relative recall}) = 1.000$	Search coverage structurally bounds performance when the complete generating truth is known.
H1 diagnostic identifiability	Combined gate accuracy=0.993; alignment-only=0.375; exact defects=0.914	Internal implementation validation: the multidimensional rule detects defects induced along its registered dimensions; external construct validity is not established.
H2 matched-budget adaptive retrieval	At budget 300, screening recall +0.384 vs D0, but reference-set-relative recall -0.049; C fell 0.133 to 0.033	Within-corpus efficiency can improve while reference-set coverage worsens.
YOLOv8 D2 repair	n 663->742; A 0.638->0.677; K 0.205->0.183; C remained 0.133	Alignment/cleanliness improved without repairing coverage; terminal UNQUALIFIED.
H3 contamination	Unmanaged premature stopping 0.711->0.761->0.808 as K rose 0.1->0.3->0.5	Apparent work savings can be produced by unsafe contamination-driven early stops.
H4 field density/prior	Field-informed prior mean RMSE 4.39 vs weak 4.05 overall; effects varied by density	Field-informed priors were not uniformly safer; H4 was mixed.

The adaptive KEEP policy improved alignment from 0.523 to 0.717, reduced contamination from 0.205 to 0.088, and increased contextual representation from 0.474 to 0.549 relative to the one-shot corpus. Coverage nevertheless fell from 0.133 to 0.033. At a matched budget of 300 records, within-corpus screening recall increased by 0.384 while reference-set-relative recall decreased by 0.049. H2 was therefore not supported on its primary reference-set-relative interpretation.

In the confirmatory YOLOv8/edge-device repair, citation-network expansion added 79 KEEP records, increasing corpus size from 663 to 742. Alignment improved from 0.638 to 0.677 and contamination fell from 0.205 to 0.183, but contextual representation changed only from 0.481 to 0.482 and coverage remained 0.133. Because $\Delta C=0$, below the frozen minimum gain of $1/30=0.0333$, the terminal ARS state was UNQUALIFIED and primary BSC was not admissible.

Posterior-predictive calibration and stopping safety

H3 was supported directionally. Relative to matched clean corpora, unmanaged contamination increased premature stopping by 0.025, 0.075, and 0.122 at contamination targets 0.10, 0.30, and 0.50, respectively. Mean recall simultaneously decreased by 0.020, 0.071, and 0.124. Work saving increased, but the apparent efficiency reflected earlier unsafe stopping rather than improved evidence recovery. Perfect removal restored the clean-corpus behavior.

H4 was mixed. Field-informed priors improved error in some prevalence strata but did not reliably improve stopping safety. Overall mean RMSE was 4.39 for the field-informed prior versus 4.05 for the weak prior, and both produced only four valid cells in the summarized primary conditions.

The corrected v1.5 boundary experiment produced the strongest negative result. Only 2 of 768 primary cells (0.26%) satisfied the complete BSC-valid definition, and no cell was BSC-useful. Mean Cov95 was 0.945, yet mean truth-relative recall was only 0.226 and the mean premature-stop rate was 0.918. Mean work saved was 0.621, showing that substantial work saving can coexist with poor truth-relative safety. Fig. 3 summarizes the safety-efficiency boundary by contrasting the aggregate operating mean with the two valid cells, both of which lie at essentially exhaustive screening.

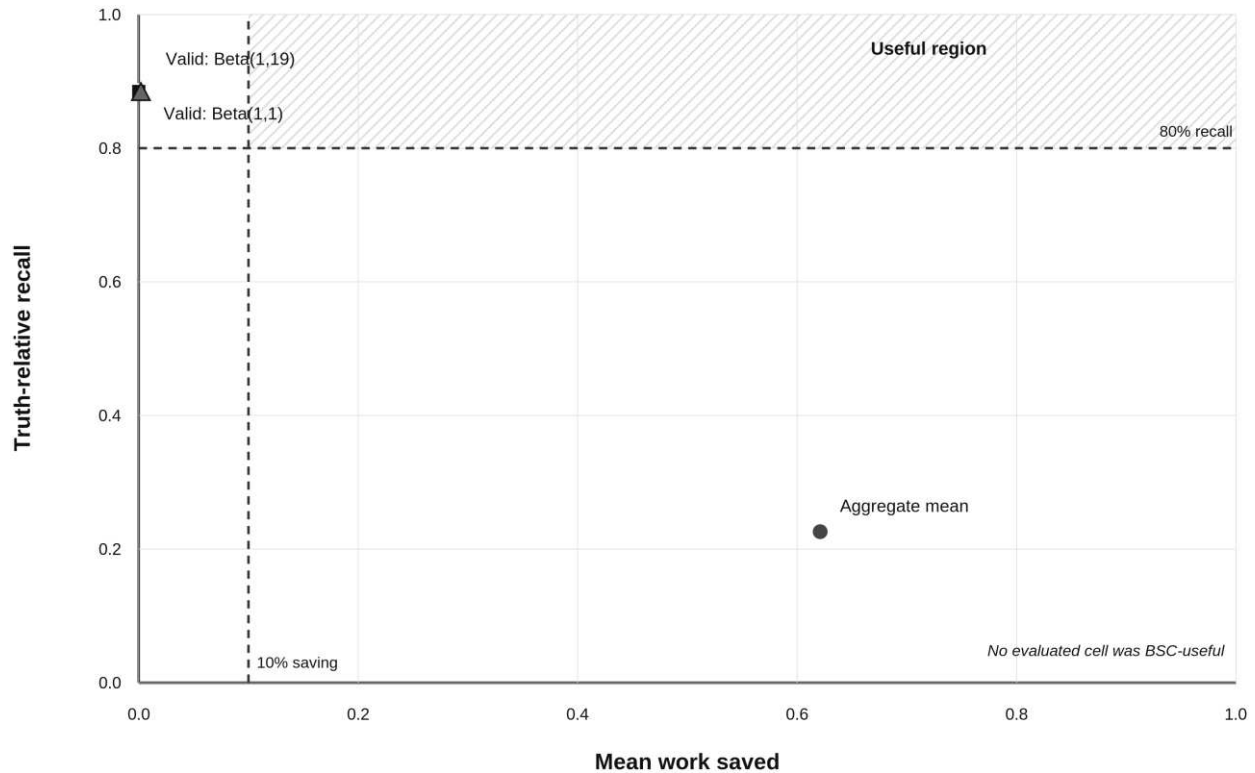


Fig. 3 Safety-efficiency boundary in the prospectively frozen v1.5 qualification-boundary experiment

The two valid cells occurred at $N=500$, prevalence=0.25, and realized coverage approximately 0.885. They achieved validity by screening essentially the full corpus (mean work saved 0-0.2%). No prespecified rho benchmark produced a non-zero aggregate useful rate. These results separate posterior-predictive interval calibration from truth-relative stopping safety.

Alternative explanations: semantics, horizon, and ordering

Semantic redundancy and readiness

One alternative explanation was that paper-level relevance counts were the wrong information unit. Early semantic development analyses did not identify useful intermediate stops using ordinary relevance-count BSC, binary whole-paper novelty, or continuous document-level semantic gain on Wolters_2018. An online new-cluster process was more promising in development and was therefore carried forward to separately frozen external testing.

Full-stream replay corrected that optimism. Across four external reviews, Setting B achieved mean paper recall 93.9%, mean evaluation-cluster coverage 97.9%, and mean work saving 11.6%, but the mean replicate-level joint safe-savings probability was only 24.45%. Setting A illustrated false efficiency most clearly in Leenaars_2020: 83.7% mean work saving but only 16.3% paper recall and 61.8% concept coverage. Fig. 4 summarizes the semantic stopping trade-off using the four-review Setting B mean and the Leenaars_2020 Setting A false-efficiency exemplar; Table 4 reports the broader progression from development to frozen external tests.

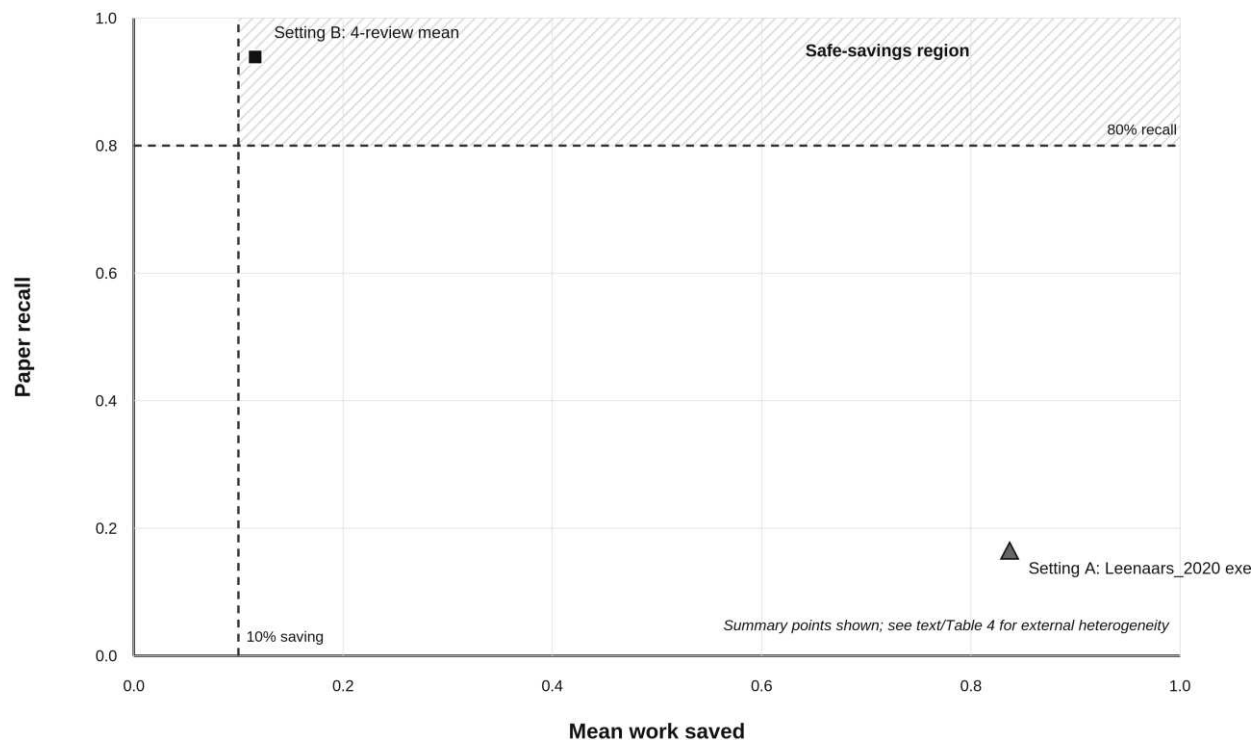


Fig. 4 External semantic stopping trade-off in the full-stream replay

Table 4 Progression from semantic development to frozen external tests

Stage	External evidence	Safety/efficiency result	Verdict
Full-stream Setting B	4 external reviews	Recall 0.939; concept coverage 0.979; work saved 0.116; joint useful probability 0.245	Promising mean operating region, inconsistent replicate-level reliability
Pre-stop ESRI_5 predictor	4 external reviews	AUC 0.624; early similarity AUC 0.650	Risk indicator only; not a deployment gate
v1.6 frozen readiness test	4 untouched reviews	Unconditional: recall 0.885, coverage 0.958, saving 0.135, safe-savings 0.062	Weak external transport
v1.6 readiness-gated	Same 4 reviews	Recall 0.898, coverage 0.962, saving 0.112, safe-savings 0.044	Safety improved modestly; utility worsened
v1.7 frozen m=5 vs m=10	4 new untouched reviews	AUC5=0.279; AUC10=0.239; Delta=-0.040; safe-savings prevalence 0.0055	Information-horizon effect not supported

In the frozen v1.6 panel of four untouched external reviews, replay-level discrimination was weak: ESRI_5 pooled ROC-AUC was 0.569 and mean early semantic similarity pooled ROC-AUC was 0.594. These pooled values summarize replay/order variability and must not be interpreted as if thousands of independent reviews had been observed. At the review level, the frozen gate improved mean joint-valid probability from 0.751 to 0.786 and mean paper recall from 0.885 to 0.898, but reduced work saving from 0.135 to 0.112 and reduced joint safe-savings probability from 0.062 to 0.044. With only four independent external reviews, the frozen v1.6 semantic-readiness hypotheses were not supported strongly; low readiness was more informative for ruling out stopping than high readiness was for supporting it.

Information horizon

Development analyses suggested improved cross-review discrimination around 8-10 relevant papers. This motivated a prospectively frozen m=5 versus m=10 test on four new datasets. The prospectively frozen result was opposite to the hypothesized direction: pooled ROC-AUC was 0.279 at m=5 and 0.239 at m=10 (delta AUC=-0.040). PR-AUC also decreased, while Brier score changed negligibly. Both horizons achieved mean paper recall and semantic-cluster coverage of 1.00 because

the semantic stopper almost always continued to the end; mean work saving was only 1.74%, and safe-savings prevalence was 0.55%. The information-horizon hypothesis was therefore not supported externally.

Retrieval ordering

The post-confirmatory retrieval-ordering analysis tested whether poor evidence ordering explained conservative stopping. After screening 10% of the corpus, random ordering retrieved 9.9% of relevant papers, compared with 18.5% for BM25, 19.8% for TF-IDF/Truncated-SVD latent-semantic ranking, and 18.4% for BM25-to-LSA reranking. Mean average precision increased from 0.065 under random ordering to 0.094 with BM25 and approximately 0.102 with the latent-semantic rankers. Thus, ranking improved early relevant-paper discovery; this analysis is descriptive and not a preregistered comparator test. Fig. 5 shows the distribution of 10%-screening recall across the four ordering strategies. The contrast between improved early relevant-paper discovery in Fig. 5 and the near-exhaustive downstream stopping behavior reported below is the empirical basis for treating ranking benefit and stopping benefit as distinct estimands.

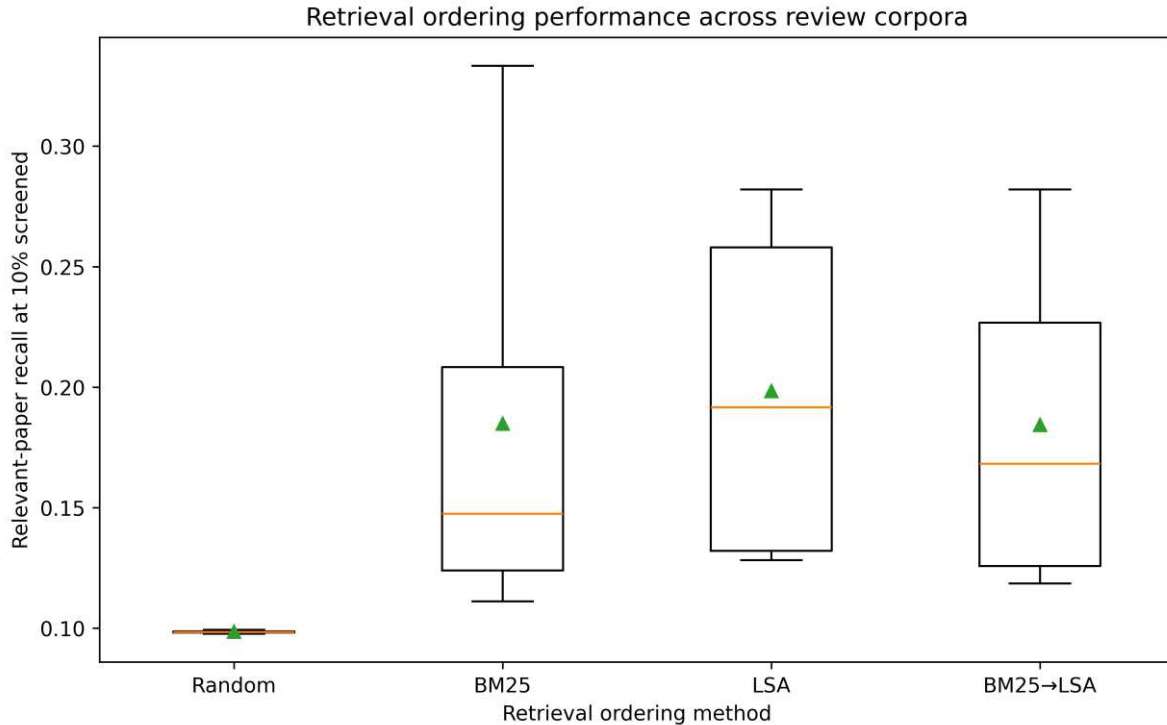


Fig. 5 Distribution of relevant-paper recall at 10% screened across random, BM25, LSA, and BM25-to-LSA retrieval ordering

The downstream semantic stopper nevertheless remained conservative. BM25 and BM25-to-LSA reranking achieved 100% mean paper recall and evaluation-cluster coverage but only 3.48% mean work saving, with no deterministic safe-savings cases under the 10% criterion; LSA produced 2.08% mean work saving. Improved ordering therefore did not by itself produce a useful stopping rule.

Benchmark-qualified external validation

The prospectively frozen experiment directly tested BSC on ten labelled review corpora distributed through the SYNERGY/SYNERGY+ screening-data collection under a benchmark-relative qualified condition in which every benchmark-known relevant record was present in D. At the frozen primary setting (Beta(1,1), $\rho=0.10$), posterior-predictive interval coverage exceeded the prespecified coverage thresholds: cross-review mean Cov95 was 0.9776 and mean Cov80 was 0.9046. The stopping policy nevertheless remained unsafe: mean premature-stop rate was 0.9154, mean benchmark-relative recall at stop was 0.1647, and mean work saved was 0.8286. None of the 10 reviews met the complete validity rule, and none were useful.

The sparse-prior sensitivity condition Beta(1,19) was more aggressive: at $p=0.10$, mean Cov95 was 0.9956 and mean Cov80 was 0.9780, but the premature-stop rate was 1.000, mean recall was 0.0270, and mean work saved was 0.9726. Again, 0/10 reviews were valid and 0/10 were useful.

The prespecified rho sensitivity analysis exposed a clear safety-efficiency frontier. With the primary Beta(1,1) prior, $\rho=0.01$ produced mean benchmark-relative recall of 0.9790 and only 0.0125 mean work saving; 7/10 reviews were valid but none reached the 10% usefulness threshold. Increasing rho raised work saving but rapidly reduced recall: at $\rho=0.02$, mean recall was 0.7787 with 0.1832 work saving and 4/10 valid reviews; at $\rho=0.05$, mean recall fell to 0.3546 with 0.6266 work saving and only 1/10 valid review; at $\rho \geq 0.10$, no review was valid. Across the frozen rho grid and both priors, no tested condition was simultaneously valid and useful.

Within the benchmark-defined candidate sets, this result rules out benchmark-relative corpus incompleteness as the sole explanation for BSC failure. Even with $C_benchmark=1$, posterior-predictive interval calibration did not translate into a safe stopping policy.

Because the 500 permutations within a review are dependent replications of one review context, uncertainty about cross-review heterogeneity was summarized descriptively by resampling the 10 reviews, not the 5,000 replay runs. Under the primary Beta(1,1) prior, the review-bootstrap 95% interval for mean recall at $\rho=0.01$ was 0.960-0.996 and for mean work saving was 0.002-0.024; at $\rho=0.10$ the corresponding intervals were 0.075-0.323 for recall and 0.662-0.923 for work saving. These intervals are descriptive robustness summaries over the 10 frozen reviews, not claims about a super-population of all systematic reviews. Table 5 reports the review-level safety-efficiency results directly so that between-review heterogeneity is visible.

Table 5 Review-level v1.8 safety-efficiency results under the primary Beta(1,1) prior. Each cell reports benchmark-relative recall / work saved. The 500 replays estimate ordering variability within each review; the table displays the independent review contexts directly

Review	$\rho=0.01$ Recall / WorkSaved	$\rho=0.02$ Recall / WorkSaved	$\rho=0.05$ Recall / WorkSaved	$\rho=0.1$ Recall / WorkSaved	$\rho=0.2$ Recall / WorkSaved	$\rho=0.5$ Recall / WorkSaved	$\rho=1$ Recall / WorkSaved
Abgaz	1.000 / 0.000	0.981 / 0.012	0.391 / 0.551	0.134 / 0.859	0.053 / 0.943	0.014 / 0.985	0.000 / 1.000
Ali	1.000 / 0.000	1.000 / 0.000	0.831 / 0.143	0.145 / 0.853	0.044 / 0.956	0.012 / 0.989	0.000 / 1.000
Anmarkrud	1.000 / 0.000	0.900 / 0.076	0.187 / 0.803	0.070 / 0.930	0.029 / 0.969	0.009 / 0.991	0.000 / 1.000
Aouad	1.000 / 0.000	0.994 / 0.005	0.318 / 0.648	0.077 / 0.920	0.030 / 0.971	0.009 / 0.992	0.000 / 1.000
Appenzeller-Herzog	0.931 / 0.043	0.397 / 0.547	0.122 / 0.873	0.058 / 0.942	0.028 / 0.972	0.009 / 0.991	0.000 / 1.000
Bech	0.996 / 0.003	0.894 / 0.079	0.192 / 0.796	0.073 / 0.926	0.028 / 0.968	0.009 / 0.990	0.000 / 1.000
Theobald	1.000 / 0.000	1.000 / 0.000	1.000 / 0.000	0.844 / 0.108	0.290 / 0.679	0.054 / 0.943	0.000 / 1.000
Wijnen	0.940 / 0.033	0.494 / 0.411	0.164 / 0.828	0.078 / 0.921	0.039 / 0.961	0.014 / 0.989	0.000 / 1.000
Williamson	0.988 / 0.005	0.736 / 0.153	0.227 / 0.744	0.118 / 0.884	0.061 / 0.946	0.015 / 0.984	0.000 / 1.000
Zakeri-Nasrabadi	0.935 / 0.041	0.390 / 0.549	0.114 / 0.879	0.050 / 0.945	0.027 / 0.973	0.011 / 0.992	0.000 / 1.000

Distributional synthesis and descriptive comparators

Post-outcome descriptive completion was limited to the preregistered comparator classes that required no new tuning: consecutive-irrelevance rules $k=\{10,20,50\}$ and the evaluation-only 80%-recall oracle. SAFE, Callaghan-type, Chao-type, and knee/fixed-budget variants remained unexecuted because their implementation was not uniquely fixed for this setting.

On the same frozen v1.8 panel and the same 500 replay orders, consecutive-10 achieved mean recall 0.052 with 94.6% work saved and a 0.996 premature-stop rate; consecutive-20 achieved recall 0.134 with 86.0% work saved and a 0.931 premature-stop rate; and consecutive-50 achieved recall 0.324 with 66.2% work saved and a 0.831 premature-stop rate. The non-deployable 80%-recall oracle achieved mean recall 0.832 with 23.9% work saved and zero premature stops by construction. These analyses are descriptive and do not support a confirmatory superiority claim for BSC over the preregistered comparator suite.

To make between-dataset and between-review heterogeneity explicit, Figs. 5–8 present distributional summaries. Fig. 5 shows the distribution of Recall@10 across the four retrieval-ordering methods. Figs. 6 and 7 show the review-level distributions of benchmark-relative recall and work saved across the frozen ρ grid under the primary prior. Fig. 8 plots the response of mean benchmark-relative recall to ρ for the primary and sensitivity prior configurations. The sensitivity prior showed the same qualitative trade-off and was generally more aggressive, with no useful condition anywhere on the frozen grid.

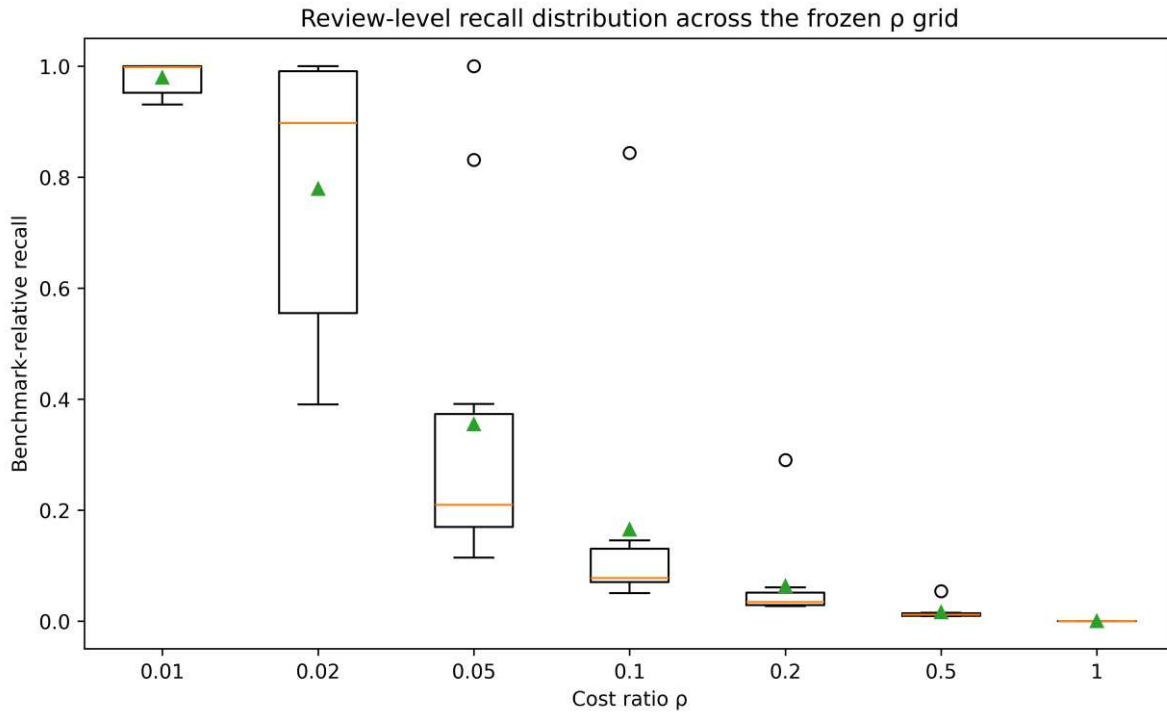


Fig. 6 Distribution of review-level benchmark-relative recall across the frozen p grid under the primary Beta(1,1) configuration (across ten review corpora)

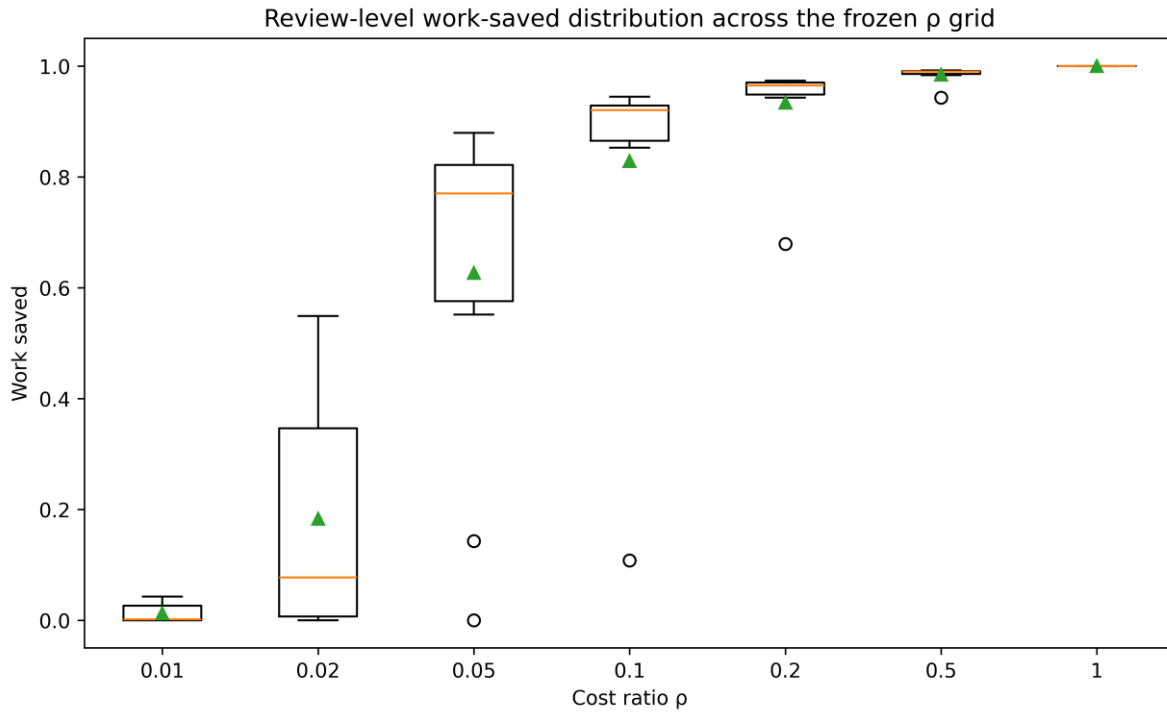


Fig. 7 Distribution of review-level work saved across the frozen p grid under the primary Beta(1,1) configuration (across ten review corpora)

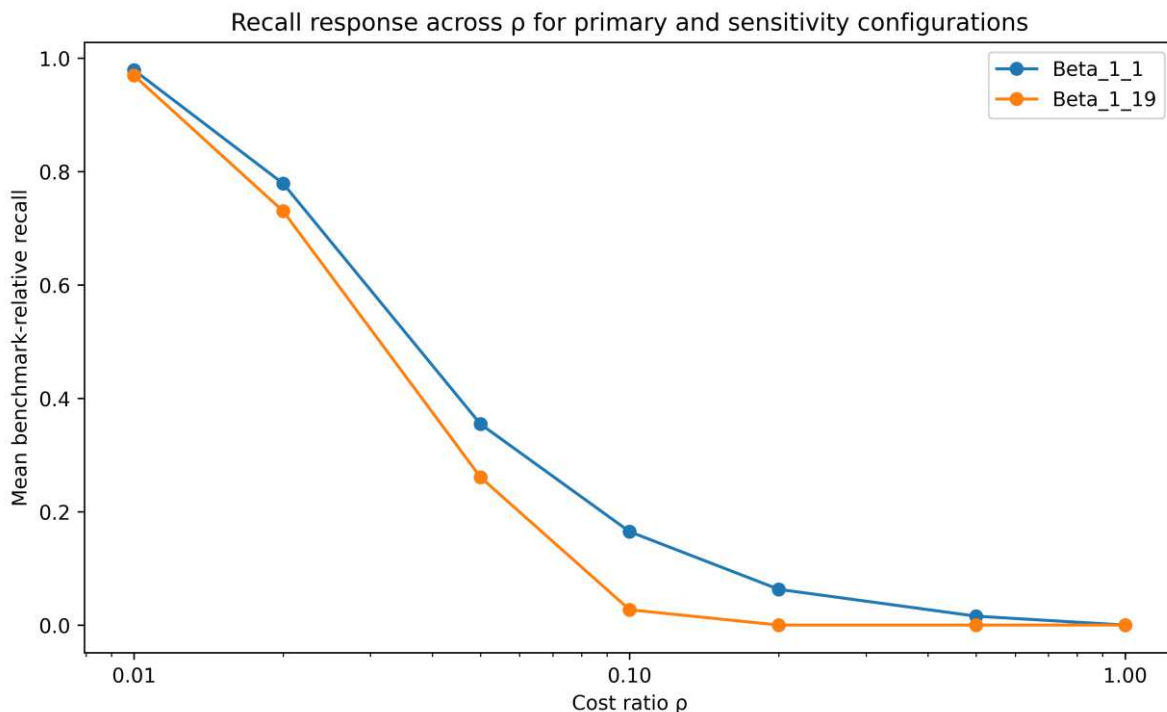


Fig. 8 Mean benchmark-relative recall across p for the primary Beta(1,1) and sensitivity Beta(1,19) configurations

Discussion

Principal findings

The principal finding is conditional rather than algorithmic: ARS-BSC does not establish a generally effective early-stopping rule. Its contribution is to identify and test boundary conditions under which a stopping decision is scientifically interpretable. The architecture separates corpus qualification from within-corpus stopping, and the empirical program shows that corpus qualification, posterior-predictive calibration, semantic redundancy, additional information horizon, and improved ranking are individually insufficient to certify useful early termination under the tested conditions. The methodological object is therefore not search quality or stopping in isolation, but the interface between the evidentiary status of a scholarly corpus and the interpretation of a stopping decision.

Across the staged evaluations, the recurring pattern was the absence of a robust safety-efficiency region. In v1.5, nearly all apparently efficient conditions stopped prematurely. In v1.6, readiness gating traded some efficiency for modest safety without improving joint safe savings. In v1.7, the hypothesized information-horizon gain disappeared on untouched data because screening became effectively exhaustive. In v1.8, a benchmark-qualified corpus still failed to make the primary BSC policy both safe and useful. Together, these results indicate that uncertainty about the unseen tail was not reducible to a reliable early saturation signal under the tested conditions.

Relationship to scientometrics and screening literature

The findings connect two literatures that are often treated separately: bibliographic corpus construction in scientometrics and stopping-rule evaluation in technology-assisted screening. Recent comparisons of OpenAlex, Web of Science, Scopus, and related sources show that bibliographic coverage, reference availability, and database-specific differences can materially affect downstream quantitative analyses and field-normalized indicators (Culbert et al., 2025; Scheidsteger et al., 2025). These studies reinforce a general point relevant to the present architecture: the properties of a scholarly corpus are not neutral with respect to the conclusions drawn from it. ARS formalizes this dependence as an explicit qualification problem before stopping is considered.

In parallel, the screening literature shows that work saving and recall depend strongly on prevalence, dataset characteristics, ranking algorithms, and stopping criteria. PRISMA-S emphasizes transparent and reproducible searching, while ASReview and related systems operationalize priority screening after a candidate set has been assembled (Rethlefsen et al., 2021; van de Schoot et al., 2021). Statistical stopping, SAFE, and related criteria address the downstream screening problem from different

directions (Callaghan & Müller-Hansen, 2020; Boetje & van de Schoot, 2024; König, 2024; Callaghan et al., 2024). The contribution here is complementary: ARS asks whether the corpus is qualified for a stopping claim before a within-corpus stopping rule is interpreted.

Recent evaluations sharpen this distinction. Kempny et al. (2026) found no universal ASReview stop criterion across five SYNERGY datasets, and Repke et al. (2026), evaluating 15 methods across 81 real-world datasets, concluded that most methods either miss relevant records or stop too conservatively. Chao-based prevalence estimation has also been proposed as a stopping criterion for technology-assisted review, illustrating the continuing search for residual-relevance estimators that can support termination decisions (Bron et al., 2025). These findings are consistent with the present safety-efficiency boundary.

The retrieval-ordering analysis adds a direct connection to information-retrieval research. BM25 and latent-semantic ranking improved the order in which relevant evidence was encountered, yet the downstream semantic stopper remained nearly exhaustive. Ranking benefit and stopping benefit are therefore distinct estimands: better information ordering can improve screening efficiency without providing evidence that residual unseen relevance risk is sufficiently small to justify termination.

The same distinction applies to rapidly developing LLM-assisted screening. Recent studies report substantial but model- and task-dependent gains in title/abstract classification and workload reduction, alongside sensitivity to model choice, prompting, eligibility formulation, cost, and operational setting (Dennstädt et al., 2024; Delgado-Chaves et al., 2025; Oami et al., 2025). Improved language understanding may enhance screening or ordering, but model-level screening performance does not by itself establish corpus qualification or stopping admissibility. Table 6 positions the present findings relative to recent stopping-rule evidence.

Table 6 Positioning relative to recent stopping-rule evidence

Recent literature	Main finding	Relationship to this study
Callaghan & Müller-Hansen 2020 (Callaghan & Müller-Hansen, 2020)	Statistical confidence-based stopping can reduce work while targeting recall.	Provides the statistical stopping tradition; ARS adds an upstream admissibility check.
Boetje & van de Schoot 2024 SAFE (Boetje & van de Schoot, 2024)	Conservative multi-condition heuristic; context and key papers matter.	Supports using multiple safeguards rather than one saturation signal.
König 2024 (König, 2024)	Stopping performance depends on prevalence and model/rule combination.	Matches our H4 and cross-review heterogeneity findings.
Callaghan et al. 2024 (Callaghan et al., 2024)	Stopping criteria should be calibrated and robust across datasets/models.	Directly motivates our no-post-hoc-tuning and external-freeze design.
Kempny et al. 2026 (Kempny et al., 2026)	No universal criterion across five SYNERGY datasets.	Converges with v1.6/v1.7 failure to transport a simple readiness rule.
Repke et al. 2026 (Repke et al., 2026)	Most methods unsafe or conservative; performance depends on good rankings.	Our BM25 analysis confirms better ordering but shows it need not create useful stopping.

Qualification, calibration, and stopping safety are distinct

ARS, BSC, semantic analysis, and retrieval ranking address different methodological objects. ARS is not a relevance predictor; it is a corpus-qualification and admissibility layer. BSC is not a model of the open scientific literature; it is a predictive count model conditional on a frozen corpus. Semantic clustering does not replace relevance screening; it tests whether conceptual redundancy contains information that paper counts omit. BM25 and semantic reranking are not stopping rules; they test whether evidence ordering explains stopping failure. Keeping these estimands separate prevents success in one layer from being misinterpreted as evidence of success in another.

The Beta-Binomial model was retained because exact conjugate updating makes the residual-count estimand transparent in sparse closed corpora. More flexible alternatives—hierarchical, rank-dependent, finite-population, capture-recapture, or neural policies—may be appropriate for different estimands or screening orders, but introducing them after observing the present failures would add post-outcome modeling degrees of freedom. The value of the present model is therefore not optimality but falsifiability.

The same principle applies to richer semantic and LLM-based representations. Transformer embeddings, cross-encoder rerankers, generative screening, and claim extraction may provide stronger representations, but their outputs depend on model family/version, prompting, decoding, training data, and domain (Reimers & Gurevych, 2019; Dennstädt et al., 2024; Delgado-

Chaves et al., 2025; Oami et al., 2025). Retrofitting those choices after observing negative outcomes would confound representation quality with post-outcome model selection; they require separately frozen comparisons.

The benchmark-qualified v1.8 experiment was designed to remove one major alternative explanation for earlier BSC failure. Every benchmark-known relevant record was present in the candidate set, so benchmark-relative retrieval loss could not by itself explain the observed stopping behavior. Nevertheless, at the primary $\rho=0.10$ condition, mean benchmark-relative recall was only 0.165. This removes benchmark-relative corpus incompleteness as the sole explanation, not field-level incompleteness.

The v1.8 result also separates posterior-predictive calibration from decision safety. At $\rho=0.10$, the Beta(1,1) model achieved $\text{Cov95}=0.978$, yet the premature-stop rate was 0.915 and mean benchmark-relative recall was 0.165. These results do not imply that the predictive intervals were intrinsically incorrect. Interval coverage concerns uncertainty about the number of relevant records remaining inside D; the stopping action additionally depends on the loss function that trades screening cost against missed-evidence loss. A calibrated predictive distribution can therefore induce an unsafe policy when the decision objective is misaligned with an external recall requirement.

The prespecified ρ grid makes the mismatch explicit. At $\rho=0.01$, mean benchmark-relative recall approached 0.98, but mean work saving fell to 1.25%. Increasing ρ increased work saving while rapidly reducing recall, and no tested condition was both valid and useful. Because $\text{recall} \geq 0.80$ was an external safety criterion rather than an objective directly encoded in the linear loss, a post-outcome recall-constrained rule was not introduced. Such a rule would constitute a different optimization problem and requires prospective evaluation.

Accordingly, ARS-BSC should not be interpreted as a two-stage mechanism that guarantees early stopping. ARS remains a corpus-qualification layer that determines whether a stopping analysis is admissible. BSC remains a transparent conditional model for residual relevance and decision trade-offs. The empirical evidence supports the architecture as a diagnostic and falsification framework, but not the tested BSC policy as a generally deployment-ready early-termination rule.

The semantic experiments reinforce this separation of estimands. Relevant-paper saturation asks how many included records remain, whereas conceptual saturation asks whether new relevant records add new conceptual information. These quantities can diverge: many relevant papers may be conceptually redundant, while a semantically similar paper may still contain a distinct result or claim. High early redundancy therefore cannot establish that few relevant papers remain downstream.

The current semantic layer measures document-level redundancy rather than verified scientific claims. Whole-document hashing and retrospective TF-IDF clusters were deliberately lightweight and auditable, but they are not a validated ontology of scientific concepts. Richer semantic units may improve representation fidelity, yet they should be evaluated prospectively rather than retrofitted to the present negative results.

Implications for automated scholarly-literature workflows

For automated scholarly-literature workflows, the practical implication is procedural rather than algorithmic: define the analytical question and inclusion criteria, construct and diagnose the corpus, distinguish ranking from stopping, and require stopping evidence that is appropriate to the corpus, screening order, and recall objective. A paper count, apparent semantic saturation, relevance-ranking score, or AI-generated summary should not be treated as evidence that screening is safe to terminate.

ARS can therefore remain useful even when BSC does not support early termination. An UNQUALIFIED state is operationally actionable: expand retrieval, repair a missing context, reduce contamination, or explicitly limit the strength of downstream claims. A SUPPORTED state means that the corpus has passed the declared qualification gate; it does not imply that screening may stop. Resource-constrained and student-led reviews were part of the motivating use case, but educational effectiveness is not an outcome of this study. Table 7 summarizes the practical interpretation for scholarly-literature screening workflows.

Table 7 Practical interpretation for scholarly-literature screening workflows

Decision point	Recommended practice for scholarly-literature screening	Reason
Before screening	Declare question, inclusion criteria, sources, coverage evidence, and corpus-quality metrics prospectively.	Stopping cannot repair evidence never retrieved.
After retrieval	Assess alignment, coverage, contamination, and context representation separately.	A high relevance score can coexist with poor reference-set coverage.

During prioritization	Use ranking to improve screening order, but report ranking and stopping as separate components.	Better BM25/semantic ranking did not guarantee safe early termination.
Before stopping	Require an externally validated, uncertainty-aware stopping method appropriate to the dataset and target recall.	Internal posterior confidence is insufficient if model assumptions or corpus adequacy fail.
If safe-saving evidence is weak	Continue to exhaustive screening and report the attempted stopping analysis as non-admissible or non-useful.	If safety evidence is insufficient, continued screening is the more defensible endpoint.

Limitations and future research

The study has five methodological strengths. First, it separates corpus qualification from downstream stopping rather than assuming the candidate set is adequate. Second, the core protocol and later external tests use explicit prospective freezes and failure states, while post-confirmatory retrieval ordering remains labelled as exploratory. Third, promising semantic findings were tested on new untouched external panels rather than repeatedly tuned on the same reviews. Fourth, negative frozen findings were retained without lowering safety or work-saving thresholds. Fifth, benchmark-qualified v1.8 was designed specifically to test whether earlier stopping failure could be explained solely by inadequate benchmark-relative corpus coverage.

- First, corpus qualification is benchmark- and operationalization-dependent. The YOLOv8 coverage measure is relative recall against a fixed 30-item constructed reference set, not an estimate of complete field coverage, and the H1 perturbation study establishes internal diagnostic identifiability rather than external construct validity of A/C/K/R. Likewise, C_benchmark=1 in v1.8 means that every benchmark-known included record was available to BSC; it does not establish that the original review corpus contained every relevant study in the broader literature. External validation of ARS will require larger independently assembled standards and prospectively justified domain-specific qualification measures.
- Second, BSC is a closed-corpus Beta-Binomial model. Its conditional-exchangeability assumption is aligned with the random-permutation validation design but is not guaranteed under relevance-ranked or adaptive screening streams. A hypergeometric or other finite-population model is a natural prospectively specified sensitivity analysis for fixed labelled corpora, while living or dynamically updated literature-search processes require different models.
- Third, external generalization is limited. Most labelled datasets originated from systematic-review screening corpora, and the v1.6 and v1.7 semantic panels contained only four independent reviews each. Replay replication estimates within-review order variability rather than increasing the number of independent review contexts. The semantic proxy operates at whole-document level, the LSA reranker is not a modern cross-encoder, and no LLM-based screening or stopping model was tested. Richer representations may improve ranking or semantic characterization, but they require separately frozen evaluation.
- Fourth, several operating and protocol constraints limit interpretation. The 80% paper-recall, 90% semantic-coverage, and 10% work-saving thresholds are experimental operating criteria rather than guideline recommendations. The full preregistered comparator suite was not completed across later extensions, and simulation-based calibration was preregistered but not implemented in the frozen analysis chain. These omissions limit comparative and implementation-validation claims; accordingly, the paper makes neither a superiority claim over the unexecuted stopping methods nor an SBC validation claim.
- Random replay simulations and fixed review labels also cannot reproduce all human screening effects, adjudication errors, query reformulation, or real-time database growth. These factors should be treated as implementation-context limitations rather than absorbed into the present closed-corpus estimand.
- Future research should test whether the identified boundary persists under alternative prospectively specified decision models and corpus representations. Priorities include finite-population residual-count models; rank-aware stopping policies that explicitly condition on screening order; independent validation of corpus-qualification dimensions using larger reference standards; and representation ablations comparing sparse lexical, sentence-transformer, cross-encoder, and claim-level semantic units while holding the decision criterion fixed. A separate prospective implementation program is required to evaluate whether ARS-style qualification improves real review workflows, including resource-constrained or student-led reviews.

Conclusion

This study asked when a stopping decision in scholarly literature screening is scientifically interpretable rather than merely computationally convenient. The resulting conditional architecture separates corpus qualification from within-corpus stopping: ARS evaluates whether the retrieved corpus is admissible for downstream stopping analysis, whereas BSC estimates residual relevance and applies an explicit finite-horizon decision rule inside that frozen corpus.

Across simulations, corpus-repair tests, semantic-readiness and information-horizon analyses, retrieval-ordering experiments, and benchmark-qualified external validation, no tested operating condition provided a robust combination of stopping safety and practically useful work saving. Benchmark-qualified v1.8 further showed that strong posterior-predictive interval coverage does not by itself imply a safe stopping policy, while improved ranking and semantic redundancy did not provide stopping certificates. The boundary condition is therefore explicit: corpus qualification, predictive calibration, and information ordering are necessary methodological considerations but are not interchangeable with stopping safety.

For quantitative scholarly-literature workflows, stopping should therefore be treated as a conditional decision whose interpretation depends on both the evidentiary status of the candidate corpus and the validity of the within-corpus decision rule. When residual-evidence risk cannot be bounded adequately under a prospectively specified policy, continued or exhaustive screening remains a scientifically defensible endpoint rather than a methodological failure.

Statements and Declarations

Protocol and registration

The domain-general v1.4 core protocol was prospectively registered on OSF before the core confirmatory analyses. v1.6 and v1.7 were prospectively SHA-256 frozen locally before new external dataset selection and execution. Retrieval ordering was conducted afterward as a post-confirmatory exploratory analysis and was neither preregistered nor represented as prospectively frozen. The v1.8 benchmark-qualified BSC validation was separately frozen locally before any BSC outcome inspection on the selected panel: freeze timestamp 2 September 2026, 12:01 Asia/Kolkata; freeze-package SHA-256 e91b95c8f8d12af6c0739583f8fc7c0e1d534eb526875a52e8b5853a17994198. These local cryptographic freezes establish tamper-evident chronology and file identity but are not represented as independent OSF or trusted-timestamp-authority registrations. The unchanged freeze packages, manifests, checksums, and analysis outputs will be archived with the reproducibility release.

Data and code availability

The final benchmark-qualified validation used labelled review corpora distributed through the SYNERGY/SYNERGY+ screening-data collection (De Bruin et al., 2023; <https://doi.org/10.34894/HE6NAQ>). The ARS-BSC reproducibility release will contain registration materials, frozen scripts and results, manifests, checksum files, dataset identifiers, source-file hashes, fixed seeds, software-environment information, and reconstruction instructions. Where bibliographic abstracts cannot be redistributed, the release will provide identifiers, exact hashes, and reconstruction instructions rather than copyrighted text. Permanent repository URL: [https://osf.io/zj7qy/overview?view_only=e3d78e4360734dcaaebbf55fb43501c0].

Ethics

Ethics approval and informed consent: Not applicable. This study did not recruit human participants, use human biological material, involve animals, or conduct an intervention. The analyses used bibliographic records, public or research-use labelled review datasets, and simulation outputs.

Funding

No funds, grants, or other support were received.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

AI-assisted technologies declaration

During manuscript preparation, ChatGPT (OpenAI) was used to assist with language refinement, consistency checking and structural organization. No AI-assisted technologies were used to generate scientific content, analyze data, interpret results, or draw scientific conclusions. The authors carefully reviewed and edited the manuscript and take full responsibility for its content.

Author contributions

CRedit: Mohammad Pasha: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft, Writing - review & editing, Project administration.

Mohammed Ali Shaik: Validation, Supervision, Writing - review & editing. Both authors reviewed and approved the final manuscript and agree to be accountable for the work.

References

- Bellman, R. (1957). *Dynamic programming*. Princeton University Press.
- Boetje, J., & van de Schoot, R. (2024). The SAFE procedure: A practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic Reviews*, 13, 81. <https://doi.org/10.1186/s13643-024-02502-7>
- Bron, M., van der Heijden, P. G. M., Feelders, A. J., et al. (2025). Using Chao's estimator as a stopping criterion for technology-assisted review. *ACM Transactions on Information Systems*. <https://doi.org/10.1145/3724116>
- Callaghan, M., Müller-Hansen, F., Bond, M., et al. (2024). Computer-assisted screening in systematic evidence synthesis requires robust and well-evaluated stopping criteria. *Systematic Reviews*, 13(1), 284. <https://doi.org/10.1186/s13643-024-02699-7>
- Callaghan, M. W., & Müller-Hansen, F. (2020). Statistical stopping criteria for automated screening in systematic reviews. *Systematic Reviews*, 9(1), 273. <https://doi.org/10.1186/s13643-020-01521-4>
- Campos, D. G., Futterer, T., Gfrorer, T., et al. (2024). Screening smarter, not harder: A comparative analysis of machine learning screening algorithms and heuristic stopping criteria for systematic reviews in educational research. *Educational Psychology Review*, 36(1), 19. <https://doi.org/10.1007/s10648-024-09862-5>
- Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., et al. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*. <https://doi.org/10.1007/s11192-025-05293-3>
- De Bruin, J., Ma, Y., Ferdinands, G., Teijema, J., & van de Schoot, R. (2023). SYNERGY - Open machine learning dataset on study selection in systematic reviews (Version 1) [Data set]. *DataverseNL*. <https://doi.org/10.34894/HE6NAQ>
- Delgado-Chaves, F. M., Jennings, M. J., Atalaia, A., et al. (2025). Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences of the United States of America*, 122(2), e2411962122. <https://doi.org/10.1073/pnas.2411962122>
- Dennstädt, F., Zink, J., Putora, P. M., Hastings, J., & Cihoric, N. (2024). Title and abstract screening for literature reviews using large language models: An exploratory study in the biomedical domain. *Systematic Reviews*, 13, 158. <https://doi.org/10.1186/s13643-024-02575-4>
- Gottlieb, M., Haas, M. R. C., Daniel, M., & Chan, T. M. (2021). The scoping review: A flexible, inclusive, and iterative approach to knowledge synthesis. *AEM Education and Training*, 5(3), e10609. <https://doi.org/10.1002/aet2.10609>
- Kempny, C., Annac, K., Wahidie, D., Yilmaz-Aslan, Y., & Brzoska, P. (2026). When to stop reviewing: Validation of stop criteria in ASReview. *BMC Medical Research Methodology*, 26(1), 109. <https://doi.org/10.1186/s12874-026-02866-5>
- König, L. (2024). An evaluation of the performance of stopping rules in AI-aided screening for psychological meta-analytical research. *Research Synthesis Methods*, 15(6), 1120-1146. <https://doi.org/10.1002/jrsm.1762>
- Oami, T., Okada, Y., & Nakada, T. (2025). Optimal large language models to screen citations for systematic reviews. *Research Synthesis Methods*, 16(6), 859-875. <https://doi.org/10.1017/rsm.2025.10014>
- Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2024). Scoping reviews. In E. Aromataris, C. Lockwood, K. Porritt, B. Pilla, & Z. Jordan (Eds.), *JBIM manual for evidence synthesis*. JBI. <https://doi.org/10.46658/JBIMES-24-09>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3980-3990. <https://doi.org/10.18653/v1/D19-1410>
- Repke, T., Tinsdeall, F., Danilenko, D., et al. (2026). Don't Stop Me Now, 'Cause I'm Having a Good Time Screening: Evaluation of stopping methods for safe use of priority screening in systematic reviews. *Cochrane Evidence Synthesis and Methods*, 4(1), e70068. <https://doi.org/10.1002/cesm.70068>
- Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., et al. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, 10, 39. <https://doi.org/10.1186/s13643-020-01542-z>
- Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333-389. <https://doi.org/10.1561/15000000019>

- Scheidsteger, T., Haunschild, R., & Bornmann, L. (2025). How similar are field-normalized citation impact scores obtained from OpenAlex and three popular commercial databases? An empirical comparison based on large-scale data. *Scientometrics*. <https://doi.org/10.1007/s11192-025-05338-7>
- Shemilt, I., Simon, A., Hollands, G. J., et al. (2014). Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1), 31-49. <https://doi.org/10.1002/jrsm.1093>
- Spiero, I., Leeuwenberg, A. M., Moons, K. G. M., et al. (2025). Evaluation of semi-automated record screening methods for systematic reviews of prognosis studies and intervention studies. *Research Synthesis Methods*. <https://www.cambridge.org/core/journals/research-synthesis-methods/article/evaluation-of-semiautomated-record-screening-methods-for-systematic-reviews-of-prognosis-studies-and-intervention-studies/2FB928A59A8B753BFD1C1EC2F20225E0>
- Tricco, A. C., Lillie, E., Zarin, W., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467-473. <https://doi.org/10.7326/M18-0850>
- van de Schoot, R., de Bruin, J., Schram, R., et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2), 125-133. <https://doi.org/10.1038/s42256-020-00287-7>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1.pdf](#)