Article

Not peer-reviewed version

Comparing Single-Agent and Multi-Agent Strategies in LLM-Based Title-Abstract Screening

Irina Radeva *, Teodora Noncheva, Lyubka Doukovska, Ivan Popchev

Posted Date: 26 March 2026

doi: 10.20944/preprints202603.2107.v1

Keywords: Large Language Models (LLMs); screening tasks; LLM coordination strategies; model selection; few-shot prompting; reproducibility; blockchain-verified audit trail

 doi:10.20944/preprints202603.2107.v1

Article

Comparing Single-Agent and Multi-Agent Strategies in LLM-Based Title-Abstract Screening

Irina Radeva [1,2,*], Teodora Noncheva [1,2], Lyubka Doukovska [1,2] and Ivan Popchev [3]

[1] Intelligent Systems Department, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria

[2] Trakia University, Faculty of Digital and Green Technologies, 6000 Stara Zagora, Bulgaria

[3] Bulgarian Academy of Sciences, 1040 Sofia, Bulgaria

* Correspondence: irina.radeva@iict.bas.bg

Abstract

Title-and-abstract screening remains labour-intensive, especially in interdisciplinary domains where shared terminology increases misclassification risk. This study compared five LLM coordination strategies — single-agent baseline, majority voting, recall-focused ensemble, confidence-weighted aggregation, and two-stage debate — using four 4-bit quantised open-source models (Mistral 7B, LLaMA 3.1 8B, Granite 3.3 8B, Qwen 2.5 7B) in zero-shot and few-shot configurations. The evaluation was conducted on a Gold Standard of 200 papers from a corpus of 2,036 records on blockchain-based e-voting. The best-performing configuration — a single-agent strategy with Qwen 2.5 7B in few-shot mode — achieved recall of 100%, precision of 70.4%, F1 of 82.6%, and a 43.4% reduction in manual screening effort, outperforming all multi-agent alternatives. Confidence-weighted aggregation produced results identical to majority voting, indicating that self-reported confidence from 7–8B parameter models did not add discriminative value. All decisions were recorded on a private Antelope blockchain with OpenTimestamps anchoring and Zenodo archival. These results suggest that, for domain-specific screening tasks, careful model selection outweighs multi-agent coordination overhead, and that few-shot prompting with a well-matched model can achieve human-level recall with substantially reduced manual effort.

Keywords: Large Language Models (LLMs); screening tasks; LLM coordination strategies; model selection; few-shot prompting; reproducibility; blockchain-verified audit trail

1. Introduction

1.1. Problem

Structured literature screening processes are the foundation of evidence-based research. They follow structured protocols to identify, screen, and synthesise all relevant studies on a given topic [1]. The screening stage is widely recognised as the most resource-intensive phase of this process. Thousands of candidate papers must be evaluated against predefined criteria, typically by two independent reviewers [2].

The challenge is not only one of scale. In interdisciplinary research areas, the vocabulary of the target field often overlaps with that of adjacent fields. The same terms carry different meanings across disciplines, and papers from neighbouring domains may appear relevant based on their titles and abstracts but fall outside the scope of the review. This terminological overlap increases both the volume of candidates and the difficulty of distinguishing target from non-target papers. Even experienced reviewers may disagree on borderline cases, making screening in such domains particularly slow and inconsistent.

 doi:10.20944/preprints202603.2107.v1

1.2. Opportunity

Large language models can process title-abstract pairs and produce structured inclusion/exclusion decisions without task-specific training [3]. Several studies have shown that LLM-assisted screening can reduce manual workload while maintaining high recall [4]. However, most existing work has relied on proprietary models accessed through commercial APIs. Few studies have explored whether locally deployed open-source models can perform this task effectively. Fewer still have examined how multiple LLM agents can be coordinated to improve screening decisions, and whether different coordination strategies suit different types of domains.

This combination of a practical problem and an underexplored technical direction defines the focus of this study.

1.3. Research Questions

Three research questions define the scope of this study.

RQ1. How does the type of coordination strategy — from simple voting to structured debate — affect screening performance when using heterogeneous open-source LLM ensembles?

RQ2. Which strategy and model combination achieves the best balance between recall and screening effort reduction?

RQ3. What systematic error patterns emerge when LLM models screen papers in a terminologically overloaded domain?

1.4. Goal and Tasks

To address these questions, the goal of this study is to compare single-agent and multi-agent LLM coordination strategies for automated title-abstract screening. Four locally deployed, 4-bit quantised open-source models are tested in a terminologically overloaded domain – blockchain-based e-voting, i.e., systems that employ distributed ledger technology to record, verify, or audit votes in public or institutional elections. In this domain, terms such as “voting”, “election”, and “consensus” carry purely technical meaning in blockchain literature, making automated screening particularly challenging.

The study was structured into five sequential tasks::

1. To implement LLM-based screening strategies and human expert screening within a single platform with blockchain audit logging.
2. To construct the target corpus through systematic search, deduplication, and keyword-based filtering across five open-access academic databases.
3. To define domain-specific inclusion and exclusion criteria for title-abstract screening.
4. To construct a gold standard through dual independent human screening with disagreement resolution, and to select few-shot examples for LLM calibration.
5. To evaluate all configurations on the gold standard, apply the top-ranked ones to the full corpus, and analyse systematic error patterns.

This study makes three contributions. First, it provides a controlled comparison of single-agent and multi-agent LLM coordination strategies for structured title-and-abstract screening tasks under local deployment constraints. Second, it shows that, at the 7–8B parameter scale, model capability was a stronger determinant of performance than coordination strategy. Third, it identifies a persistent precision ceiling driven by ambiguity in title-and-abstract screening, supported by systematic error analysis.

1.5. Paper Organisation

The remainder of this paper is organised as follows. Section 2 reviews related work on literature screening and LLM-assisted screening approaches. Section 3 describes the proposed framework, including corpus construction, gold standard creation, strategy design, and evaluation protocol.

 doi:10.20944/preprints202603.2107.v1

Section 4 presents the experimental results. Section 5 discusses findings, practical recommendations, and limitations. Section 6 concludes the paper.

2. Related Work

This section reviews research in three areas relevant to the present study: (a) LLM-assisted screening in systematic reviews, (b) multi-agent and ensemble LLM strategies, and (c) inter-rater reliability in screening. The section concludes with a summary of the identified research gap.

2.1. LLM-Assisted Screening in Systematic Reviews

The PRISMA 2020 guidelines require transparent documentation of study selection procedures, including any automation tools [1]. Title-and-abstract screening remains one of the most labour-intensive phases. An analysis of human reviewers across multiple systematic reviews reported a mean error rate of 10.76% during abstract screening [5], providing an empirical baseline for evaluating automated approaches.

Early automation relied on traditional machine learning. A voting perceptron classifier was applied to 15 drug class reviews, and the Work Saved over Sampling at 95% recall (WSS@95) metric was introduced as a standard measure of screening efficiency [2]. Active learning tools such as ASReview further improved efficiency by prioritising records for human review using multiple classifiers and query strategies [6]. However, these tools still require iterative human labelling.

Large language models introduced a different approach. An evaluation of ChatGPT (GPT-3.5 Turbo) for systematic review screening found performance comparable to traditional classifiers such as support vector machines, without task-specific training [7]. A pre-registered study of GPT-4 tested title/abstract screening, full-text review, and data extraction across peer-reviewed, grey, and non-English literature. GPT-4 achieved high specificity but variable sensitivity depending on dataset balance. It was concluded that GPT-4 may function as a secondary reviewer but should not replace human judgement entirely [8]. A hybrid workflow combining LLM analysis with human verification was also proposed, where the LLM identified misclassified articles that were missed during human-only screening [9]. A methodological review emphasised that recall must be prioritised in early screening phases and that iterative prompt refinement is essential [10]. A three-layer strategy using GPT-3.5 and GPT-4 was proposed, where each layer evaluated a different inclusion criterion: research design, target population, and intervention [11]. This layered approach is conceptually similar to the two-stage filtering strategy (S5) in the present study. The insufficiency of existing reporting standards for AI-aided screening has been explicitly documented. The RDAL checklist demonstrated that PRISMA guidelines do not require detailed recording of screening decisions, model settings, or training data in active learning-aided reviews [12]. A broader extension, PRISMA-trAIce, was subsequently developed to cover transparent AI reporting across all phases of a systematic review [13]. Both initiatives confirm that reproducibility in AI-assisted screening remains an open methodological challenge.

2.2. Multi-Agent and Ensemble LLM Strategies

A survey of LLM-based multi-agent systems identified three communication paradigms: cooperative, debate, and competitive [14]. Cooperative communication underlies majority voting (S2) and recall-focused aggregation (S3) in the present study, while the debate paradigm corresponds to strategy S5. A self-consistency decoding method samples multiple reasoning paths from a language model and selects the most frequent answer through majority voting [15]. This approach improved accuracy on arithmetic and commonsense benchmarks and provides the theoretical basis for strategy S2. In a complementary direction, a multi-agent debate approach refines responses through structured argumentation rounds among multiple model instances [16]. The debate mechanism improved both factual accuracy and reasoning by exposing errors through peer critique, informing the debate component of strategy S5. Adversarial debate combined with voting mechanisms was also

 doi:10.20944/preprints202603.2107.v1

investigated to reduce LLM hallucinations, using dynamic weighting to prioritise high-performing models [17].

A framework for reliable decision-making in multi-agent LLM systems compared aggregation strategies including majority voting, decentralised communication, and spoke-and-wheel architectures. Majority voting and decentralised approaches consistently formed the Pareto front of reliability across tasks [18]. These aggregation patterns are consistent with the majority voting (S2) and confidence-weighted (S4) strategies adopted in the present study. Scaling the number of agents in a majority-voting framework also yielded consistent accuracy improvements [19], supporting the use of three-agent ensembles in the present study.

Ensemble techniques have been applied to scientific literature classification by combining outputs from multiple LLMs using a confidence calibration framework. The ensemble achieved higher accuracy than any individual model [20]. A survey of ensemble approaches distinguished between model-level, parameter-level, and task-specific ensembles. Ensemble methods improved robustness but introduced additional computational costs [21]. This trade-off is a central consideration in the present study, where all models were executed locally on consumer hardware. A comparative evaluation of four coordination strategies (collaborative, sequential, competitive, and hierarchical) against calibrated single-agent RAG baselines reported statistically significant performance degradation across all 28 tested configurations, with coordination overhead identified as the primary contributing factor [22]. A recent evaluation applied three multi-agent collaboration strategies — majority voting, multiagent debate, and LLM-based adjudication — directly to abstract screening across 28 biomedical systematic reviews. Majority voting with three API-based models consistently outperformed individual models, while adjudicator-as-a-ranker achieved the best results among adjudication variants [23].

2.3. Inter-Rater Reliability in Systematic Reviews

Inter-rater reliability (IRR) is a fundamental concern in systematic reviews. An analysis of screening practices found that IRR is widely under-reported and that coding behaviour varies both between and within individuals over time [24]. A study of inter-reviewer reliability across clinical systematic reviews reported a mean Cohen's kappa of 0.82 for abstract screening [25]. However, agreement decreased for interdisciplinary or emerging research areas where terminology was not yet standardised. Inter-rater agreement was also assessed using the PROBAST tool for prediction model studies, revealing kappa values between 0.04 and 0.26 at the domain level [26].

These findings suggest that moderate agreement levels are expected in interdisciplinary domains where terminology spans multiple fields, particularly when screening criteria require distinguishing application context rather than topic [24].

2.4. Research Gaps

The reviewed literature reveals several gaps. First, most evaluations of LLM screening have relied on proprietary models such as GPT-3.5 and GPT-4 [7,8,10,11]. Recent work has begun testing open-source models deployed locally via frameworks such as Ollama [27], but these evaluations assessed models individually rather than in coordinated multi-agent configurations. Although multi-agent strategies have recently been applied to biomedical screening with API-based models [23], no study has applied multi-agent coordination strategies using quantised open-source models deployed locally on consumer hardware, nor evaluated such strategies in terminologically overloaded interdisciplinary domains. Second, multi-agent strategies have been explored for general reasoning tasks [14–19] but not applied to systematic review screening. The present study compares five strategies (S1–S5) implemented from established multi-agent decision-making patterns [14–21]. A recent evaluation of multi-agent coordination for RAG-based question answering confirmed consistent performance degradation when applied to 7–8B parameter models [22]. Whether similar patterns emerge in the structurally different task of binary screening classification has not been investigated. Third, no prior work has combined LLM-based screening with blockchain-based audit

 doi:10.20944/preprints202603.2107.v1

trails to ensure decision provenance and reproducibility. The reproducibility gap in AI-aided screening has been explicitly recognised — the RDAL checklist [12] and PRISMA-trAIce [13] address reporting standards, but infrastructure-level decision provenance remains unaddressed. A framework for recording AI decisions using blockchain ledgers has been proposed for IoT environments [28]. A systematic analysis of blockchain integration frameworks, covering oracles, distributed file systems, cross-chain protocols, and middleware, demonstrated that integration choices encode governance assumptions and require evaluation beyond technical performance alone [29]. However, the application of blockchain-based audit trails to systematic review screening has not been investigated. Fourth, prior evaluations used primarily biomedical datasets [2,8,9,11]. The present study applies multi-agent screening to an interdisciplinary domain where terminology spans multiple fields.

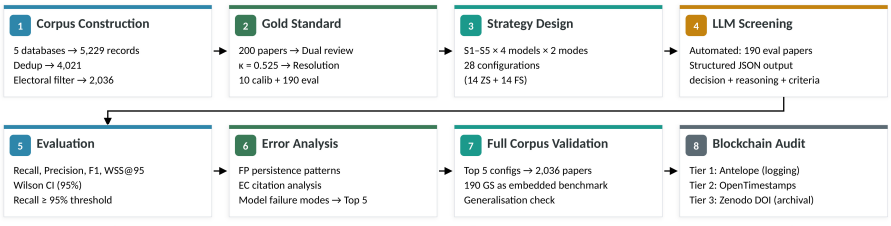
3. Methods

3.1. Framework Overview

This study presents a task-driven framework for designing and evaluating multi-agent LLM coordination strategies in structured title-and-abstract screening tasks. The framework is domain-independent and can be applied to any systematic review corpus. It consists of five sequential phases:

1. Corpus construction. A domain-specific corpus is assembled from multiple open-access databases using a Boolean search strategy. The collected records are deduplicated and filtered to retain only papers relevant to the target domain.
2. Gold Standard creation. A subset of the corpus is sampled and screened independently by two human reviewers. Inter-rater agreement is measured using Cohen's Kappa and PABAK. Disagreements are resolved by a third reviewer. The resulting consensus labels serve as ground truth for LLM evaluation.
3. Strategy design. Five LLM coordination strategies of increasing complexity are defined: single-agent screening (S1), majority voting (S2), recall-focused ensemble (S3), confidence-weighted aggregation (S4), and two-stage screening with debate (S5). Each strategy is tested in zero-shot and few-shot modes.
4. Evaluation. LLM screening decisions are compared against the Gold Standard using Recall, Precision, F1 Score, and Work Saved over Sampling at 95% recall (WSS@95). A minimum Recall threshold of 95% is applied, reflecting the requirement that systematic reviews must not miss relevant studies.
5. Blockchain audit. All screening decisions — both human and LLM — are logged to a private Antelope blockchain. Periodic Merkle root anchoring to a public repository (Zenodo) and timestamping (OpenTimestamps) provide external verifiability without exposing individual records.

The full experimental workflow, including the evaluation and validation stages that follow screening, is illustrated in Figure 1. Steps 1–4 (top row) cover corpus construction through automated LLM screening; steps 5–8 (bottom row) cover evaluation, error analysis, full corpus validation, and blockchain-based audit logging. The framework was integrated into PaSSER-SR, an open-source platform described in the PaSSER-SR Platform section (Section 3.8). It was validated on a case study domain described in Section 3.2.



 doi:10.20944/preprints202603.2107.v1

Figure 1. Experimental workflow of the screening evaluation framework.

3.2. Case Study Domain

The framework was validated on a corpus of blockchain-based e-voting systems within the broader electoral process context. This domain was selected because it presents a particularly challenging case for automated screening due to high terminological overlap with adjacent fields.

Blockchain-based e-voting systems constitute a narrow research area. Key terms such as "voting", "consensus", "verification", and "governance" carry fundamentally different meanings depending on context. In blockchain consensus mechanisms, "voting" refers to validator agreement on block validity. In decentralised autonomous organisations (DAOs), "voting" denotes token-based governance decisions. In corporate governance, "voting" describes shareholder participation. Only in the e-voting context does "voting" refer to citizen participation in public elections.

This semantic ambiguity creates a difficult screening task. The models must distinguish the application context of shared terms rather than rely on keyword matching alone. A paper titled "A Secure Voting Protocol for Blockchain Governance" may contain all expected keywords yet fall entirely outside the scope of e-voting systems research.

The domain therefore serves as a rigorous test case. If the proposed coordination strategies achieve acceptable screening performance in a terminologically overloaded domain, they can be expected to perform at least as well in domains with cleaner terminological boundaries.

3.3. Dataset Construction

3.3.1. Search Strategy

A Boolean search strategy was designed to capture blockchain applications in electoral processes. Search terms were organised into two groups: Group A contained blockchain-related terms (*blockchain*, *distributed ledger*, *DLT*, *smart contract*, *decentralised*), and Group B contained electoral terms (*voting*, *election*, *e-voting*, *electoral*, *ballot*, *referendum*, *voter registration*, *vote counting*). Each query required at least one term from each group. Terms related to non-electoral blockchain mechanisms were excluded during post-processing (*DAO voting*, *governance voting*, *governance token*, *token voting*). The search covered publications from 2015 to 2025, spanning the period from blockchain's emergence as a research topic to the present.

3.3.2. Data Sources

Papers were collected from five open-access databases: OpenAlex, Semantic Scholar, CORE, arXiv, and MDPI. This selection ensured broad coverage without reliance on subscription-based services. The Boolean query was adapted for each database API. Table 1 summarises the retrieval results.

Table 1. Data sources and retrieval statistics.

Database	Access Method	Records Retrieved
OpenAlex	REST API	1,644
Semantic Scholar	API with rate limiting	2,392
CORE	API with registration	848
arXiv	OAI-PMH API	259
MDPI	BibTeX export	86
Total		5,229

All collection scripts are publicly available in the project repository https://github.com/scpdxtest/PaSSER-SR.

 doi:10.20944/preprints202603.2107.v1

3.3.3. Deduplication

A two-stage deduplication process was applied to the combined corpus. First, exact DOI matching identified duplicates across databases. Second, for records without DOIs, normalised title similarity was computed using the Ratcliff/Obershelp algorithm (Python SequenceMatcher), with a threshold of 0.85. When duplicates were found, metadata was merged to retain the most complete record. This process removed 1,208 duplicates, yielding 4,021 unique papers. A total of 857 papers appeared in more than one database.

3.3.4. E-Voting Context Filtering

The unified corpus contained papers matching the broad Boolean query, including works on blockchain consensus mechanisms, decentralised governance, and smart city platforms that were not related to electoral processes. A keyword-based filter was applied to retain only papers with explicit electoral context. The filter matched 48 domain-specific terms (e.g., election, ballot, voter, referendum, polling station, voter registration, parliamentary, presidential) against each paper's title and abstract. Papers without any matching term were excluded. This step removed 1,985 papers, producing a final corpus of 2,036 papers for screening. Figure 2 presents the PRISMA 2020 flow diagram of the selection process.

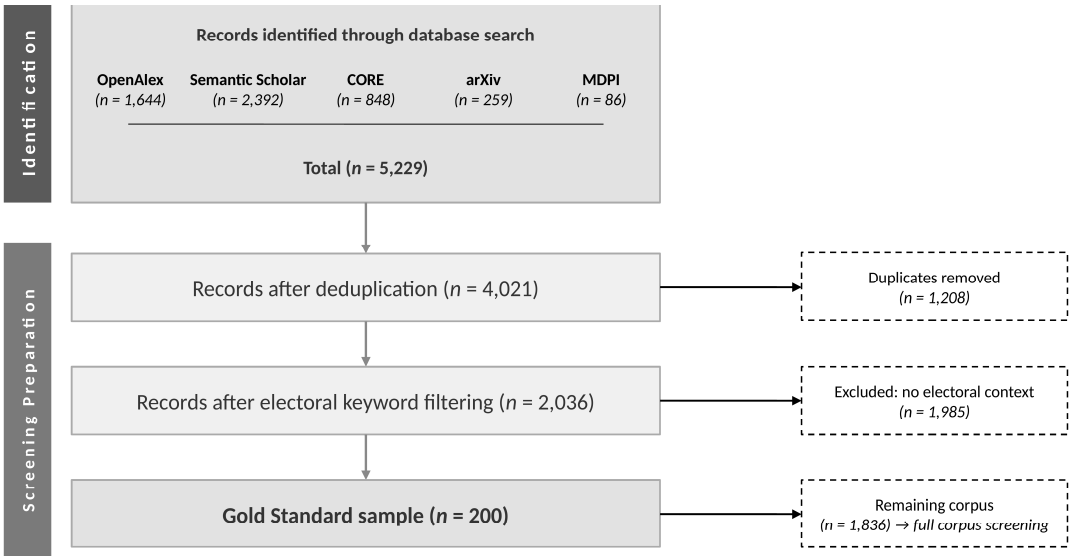


Figure 2. PRISMA 2020 flow diagram of the selection process.

3.4. Gold Standard Protocol

The evaluation of automated screening tools requires a set of expert-labelled decisions serving as ground truth — commonly referred to as a gold standard in the systematic review literature [2,6]. The following protocol was applied to construct such a set for the present study.

3.4.1. Sampling

The Gold Standard was constructed from the filtered corpus of 2,036 papers. Papers were partitioned into two pools based on the presence of electoral keywords in the title or abstract: Pool A (1,954 papers with keywords) and Pool B (82 papers without keywords). A random sample of 200 papers was drawn from Pool A using a fixed seed (seed = 42) to ensure reproducibility. Pool B was not sampled due to its small size and the expectation that few papers would meet the inclusion criteria. All 200 papers contained at least one electoral keyword, providing a sample enriched for borderline cases where screening decisions are most difficult. The sampling strategy intentionally

 doi:10.20944/preprints202603.2107.v1

focuses on keyword-rich records, increasing the proportion of difficult cases. The resulting Gold Standard should therefore be interpreted as a stress-test set rather than a statistically representative sample of the full corpus.

3.4.2. Screening Criteria

Five inclusion criteria (IC1–IC5) and six exclusion criteria (EC1–EC6) were defined by the authors specifically for the present study, based on the scope and boundaries of blockchain-based electoral systems research. A paper was included if it satisfied IC1 and IC2, and at least one of IC3–IC5, and did not meet any exclusion criterion. Table 2 presents the full criteria definitions.

Table 2. Inclusion and exclusion criteria for title-abstract screening.

Code	Type	Criterion
IC1	Incl.	Proposes, describes, or evaluates a blockchain-based model, framework, or system
IC2	Incl.	Addresses electoral process (voter authentication, registration, petition signing, voting, counting, auditing, dispute resolution) for public or institutional elections (national, regional, local, university, organisation)
IC3	Incl.	Includes empirical evaluation or experimental results
IC4	Incl.	Contains security/privacy analysis
IC5	Incl.	Describes implementation or prototype
EC1	Excl.	No blockchain technology discussed, or mentions blockchain without specific implementation
EC2	Excl.	Focuses on non-electoral domain (e.g., finance, supply chain, healthcare, IoT, energy) or discusses decentralisation/blockchain in general without electoral application
EC3	Excl.	Opinion pieces, position papers, tutorials, or general overviews/surveys without systematic method or original contribution
EC4	Excl.	DAO governance, corporate voting, or technical voting/election mechanisms (consensus protocols, node/notary/leader election, Byzantine voting)
EC5	Excl.	Abstract missing, insufficient, unclear scope, or not in English
EC6	Excl.	Only theoretical discussion, or general blockchain/smart contract concepts without concrete electoral application

3.4.3. Screening Procedure

Prior to independent screening, the screening criteria were discussed between the two reviewers to align interpretation. Each reviewer then screened all 200 papers independently in blind mode using the PaSSER-SR Human Screening Module. For each paper, reviewers recorded a decision (INCLUDE, EXCLUDE, or UNCERTAIN), a confidence level (HIGH, MEDIUM, or LOW), the specific criteria met or violated, and free-text reasoning.

3.4.4. Inter-Rater Agreement and Disagreement Resolution

Inter-rater reliability was measured using Cohen's Kappa (κ) and Prevalence-Adjusted Bias-Adjusted Kappa (PABAK). Disagreements were resolved by a third reviewer who examined the original paper, both reviewers' decisions and reasoning, and made a final determination. The agreement statistics and disagreement patterns are reported in the Gold Standard Results section.

3.5. LLM Coordination Strategies

The five coordination strategies were implemented for the present study, each based on an established multi-agent decision-making paradigm identified in the Section 2.2 [15,16,18,20] and adapted for the title-abstract screening task. Each strategy receives a paper's title and abstract as input

 doi:10.20944/preprints202603.2107.v1

and produces a binary decision (INCLUDE or EXCLUDE), a confidence level (HIGH, MEDIUM, or LOW), the criteria applied, and free-text reasoning.

S1: Single Agent (Baseline). A single LLM model screens each paper independently. This strategy serves as the baseline against which multi-agent approaches are compared.

S2: Majority Voting. Three LLM models screen each paper independently. The final decision is determined by simple majority [15,19,20]: if two or more models agree on INCLUDE or EXCLUDE, that decision is adopted. If no majority is reached, the paper is marked UNCERTAIN. The aggregated confidence is computed as the mean of individual confidence scores, mapped to HIGH (≥ 0.85), MEDIUM (≥ 0.65), or LOW.

S3: Recall-Focused Ensemble. Three LLM models screen each paper independently. If any model votes INCLUDE, the final decision is INCLUDE (OR logic). This strategy prioritises recall at the expense of precision, reflecting the principle that missing a relevant study is more costly than including an irrelevant one. If no model votes INCLUDE but at least one votes UNCERTAIN, the paper is marked UNCERTAIN; otherwise, it is marked EXCLUDE.

S4: Confidence-Weighted Aggregation. Three LLM models screen each paper independently. Each vote is weighted by the model's self-reported confidence level, mapped to numerical weights: HIGH = 0.9, MEDIUM = 0.7, LOW = 0.5. INCLUDE votes contribute positive weight, EXCLUDE votes contribute negative weight, and UNCERTAIN votes contribute zero to the numerator but their confidence weight is included in the normalisation denominator. The weighted scores are normalised, and the final decision is determined by threshold: a normalised score above +0.2 results in INCLUDE, below −0.2 in EXCLUDE, and between these values in UNCERTAIN.

S5: Two-Stage Screening with Debate. This strategy separates screening into two stages, combining fast filtering with multi-agent debate [16]. In Stage 1, a designated fast-filter model screens each paper. Papers receiving a HIGH-confidence EXCLUDE decision are immediately excluded. All remaining papers proceed to Stage 2, where two additional models independently screen the paper. If all three models (including the Stage 1 response) reach consensus, that decision is final. If disagreement persists, the majority decision is adopted; in the event of a tie, INCLUDE is selected to preserve recall. The fast-filter model and debate models are configurable, allowing role assignment based on individual model strengths.

3.5.1. Models

All strategies were tested with four open-source LLM models: Mistral 7B Instruct v0.3 (Mistral AI), Meta LLaMA 3.1 8B Instruct, Qwen 2.5 7B Instruct (Alibaba), and IBM Granite 3.3 8B Instruct. Model selection was governed by two requirements. First, a core design principle of PaSSER-SR is privacy-preserving, cloud-independent operation. All inference is performed locally on Apple Silicon hardware using the Apple MLX framework (mlx-lm) with 4-bit quantisation. This restricts the candidate pool to models that have MLX-compatible quantised variants and fit within the unified memory of consumer-grade hardware. No cloud-based or commercial APIs were used. Second, to maximise architectural diversity within this constraint, four models were chosen from independent development pipelines with different pre-training corpora, fine-tuning procedures, and alignment strategies.

Four models represent the maximum number of independently developed 7–8B instruction-tuned models with MLX-compatible 4-bit variants available at the time of the study; this count also satisfies the minimum ensemble size of three required for majority voting (S2), with one additional model enabling varied three-model combinations for S5. The 7–8 billion parameter range represents the practical upper bound for 4-bit local inference on devices with 16–32 GB of unified memory. The PaSSER-SR platform extends the original PaSSER framework [30], which employed Mistral 7B, Llama2 7B, and Orca2 7B for RAG evaluation in the smart agriculture domain. The present study uses newer model versions (LLaMA 3.1, Qwen 2.5) and adds IBM Granite. Three of the four model families (Mistral, LLaMA, and Granite) were previously evaluated in a multi-agent RAG context

 doi:10.20944/preprints202603.2107.v1

using the PaSSER platform [22], enabling cross-study comparison of model behaviour across different tasks.

Each strategy–model combination was tested in two prompt modes: zero-shot (screening criteria only) and few-shot (criteria plus labelled examples), as described in the Few-Shot Example Selection Protocol section.

3.6. Few-Shot Example Selection Protocol

Few-shot examples were selected through error-driven analysis of zero-shot results rather than random sampling. The selection protocol consisted of three steps.

Step 1: Cross-strategy error aggregation. Error analysis reports from all zero-shot runs were aggregated across strategy–model configurations. For each paper appearing as a false positive (FP) or false negative (FN), the number of configurations in which the error occurred was counted. Papers producing errors in a larger number of configurations were considered more representative of systematic model weaknesses.

Step 2: Error pattern categorisation. FP errors were categorised by confusion type: (a) domain boundary confusion, where blockchain consensus terminology was mistaken for electoral voting (related to EC2 and EC4); (b) missing implementation, where papers discussed elections without proposing a concrete blockchain system (EC1); (c) opinion or review papers without an original contribution (EC3); and (d) governance confusion, where DAO or corporate voting was misidentified as public electoral voting (EC4). FN errors were categorised by: (a) overly aggressive application of exclusion criteria (EC2 and EC3); and (b) terminology mismatch, where synonyms such as "distributed ledger" were not recognised as blockchain-related.

Step 3: Representative selection. Five EXCLUDE and five INCLUDE examples were selected to maximise pattern coverage across the identified error categories. EXCLUDE examples prioritised papers that appeared as FP in the largest number of configurations. INCLUDE examples targeted specific model weaknesses, such as the tendency of certain models to over-apply EC2 and EC3 when screening review papers.

The 10 selected papers were marked as calibration examples (is_calibration = true) in the Gold Standard database and excluded from all evaluation runs to prevent data leakage. Each few-shot example included the paper title, abstract, ground-truth decision, criteria applied, and a brief reasoning statement. Few-shot runs were therefore evaluated on the remaining approximately 190 papers from the Gold Standard.

3.7. Evaluation Protocol

Each strategy–model–prompt mode configuration was evaluated against the human ground truth established during gold standard screening (Section 3.3). The ground truth decision for each paper was determined as follows: where both screeners agreed, the consensus decision was used; where they disagreed, the resolution decision provided by the third reviewer was adopted.

Screening was framed as a binary classification task. INCLUDE was treated as the positive class and EXCLUDE as the negative class. Four standard confusion matrix counts were computed: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these, four primary metrics were derived.

Recall (sensitivity) measured the proportion of relevant papers correctly identified by the LLM strategy:

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

Precision (positive predictive value) measured the proportion of papers classified as INCLUDE that were genuinely relevant:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

 doi:10.20944/preprints202603.2107.v1

F1 Score provided the harmonic mean of Recall and Precision:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

Work Saved over Sampling at 95% recall (WSS@95) quantified the reduction in screening workload compared to random sampling [2]:

$$WSS95 = \frac{TN + FN}{N} - 0.05 \quad (4)$$

This metric is meaningful only when Recall ≥ 0.95 and represents the proportion of papers that need not be manually screened beyond the 5% baseline cost of random sampling [31].

Recall ≥ 0.95 was adopted as the primary acceptance threshold. In systematic review screening, missing relevant studies (false negatives) poses a greater risk to review validity than including irrelevant ones (false positives), which are eliminated during full-text review [2]. Strategies meeting this threshold were ranked by WSS@95 in descending order. Strategies failing to meet the threshold were ranked separately by Recall.

Confidence intervals. For Recall and Precision, 95% confidence intervals were computed using the Wilson score method [32]. The Wilson interval was selected over the Wald (normal approximation) interval because it provides more accurate coverage for small sample sizes and for proportions near 0 or 1 [33], both of which apply to the present evaluation set.

UNCERTAIN treatment. Human screening decisions included an UNCERTAIN category for ambiguous cases. Two treatment modes were evaluated: (a) recall-focused, where UNCERTAIN decisions in both ground truth and predictions were mapped to INCLUDE, representing a conservative approach that avoids missing potentially relevant papers; and (b) precision-focused, where UNCERTAIN decisions were mapped to EXCLUDE. All results reported in Section 4 use the recall-focused treatment unless otherwise stated. Both treatments are reported in full in the supplementary materials.

Calibration exclusion. The 10 papers marked as few-shot calibration examples (Section 3.6) were excluded from all evaluation runs. Each configuration was therefore evaluated on approximately 190 papers from the Gold Standard.

Full corpus validation. To assess whether screening performance generalises beyond the Gold Standard, the best-performing configurations were applied to the full corpus of 2,036 papers. Evaluation metrics were computed on the 190 GS evaluation papers embedded within the corpus, using the same ground truth and the same formulas described above. Consistency between GS-only and full-corpus metrics served as evidence that LLM performance does not degrade at scale.

3.8. PaSSER-SR Platform

All experimental procedures were conducted within PaSSER-SR, a web-based platform developed as an extension of the PaSSER framework [30]. PaSSER-SR was designed to support the full systematic review screening workflow, from corpus management through human and LLM-based screening to evaluation and audit. The system architecture is illustrated in Figure 3.

Presentation layer. The user interface was implemented in React with the PrimeReact component library. Four functional modules were provided: (a) Corpus Collection, for browsing and filtering the imported paper corpus; (b) Human Screening, for manual title-abstract screening with structured criteria selection, confidence tracking, and disagreement resolution; (c) LLM Screening, for configuring and executing automated screening runs with real-time progress monitoring via WebSocket; and (d) Evaluation and Audit, for computing performance metrics, comparing strategies, and managing the blockchain audit trail.

Application layer. The backend consisted of two Python FastAPI services and two supporting modules (Figure 3). The first service (screening_api.py) handled project management, corpus and Gold Standard operations, human screening decisions, disagreement resolution, and role-based access control. Three user roles were defined: screener (submits decisions), resolver (handles

 doi:10.20944/preprints202603.2107.v1

disagreements), and admin (manages projects, exports, and audit). The second service (llm_screening_api.py) managed LLM model loading and inference via the Apple MLX framework, strategy execution, and few-shot example retrieval. A dedicated evaluation module (evaluate.py) computed Recall, Precision, F1, and WSS@95 with Wilson confidence intervals across all strategy–model–prompt mode configurations. A blockchain logging module (blockchain_logger) recorded all screening decisions to the Antelope blockchain, computed Merkle roots via the logexport action, and coordinated OpenTimestamps anchoring and Zenodo DOI publication (Section 3.9).

Data and infrastructure layer. Three components supported data storage and processing. MongoDB served as the primary database, storing corpus papers, Gold Standard records, human screening decisions, LLM decisions, disagreement resolutions, and evaluation results. The Apple MLX framework (mlx-lm) provided local LLM inference on Apple Silicon hardware with 4-bit quantised models, eliminating the need for cloud-based API services. A private Antelope blockchain recorded all screening decisions as immutable audit entries; the audit trail is described in Section 3.9.

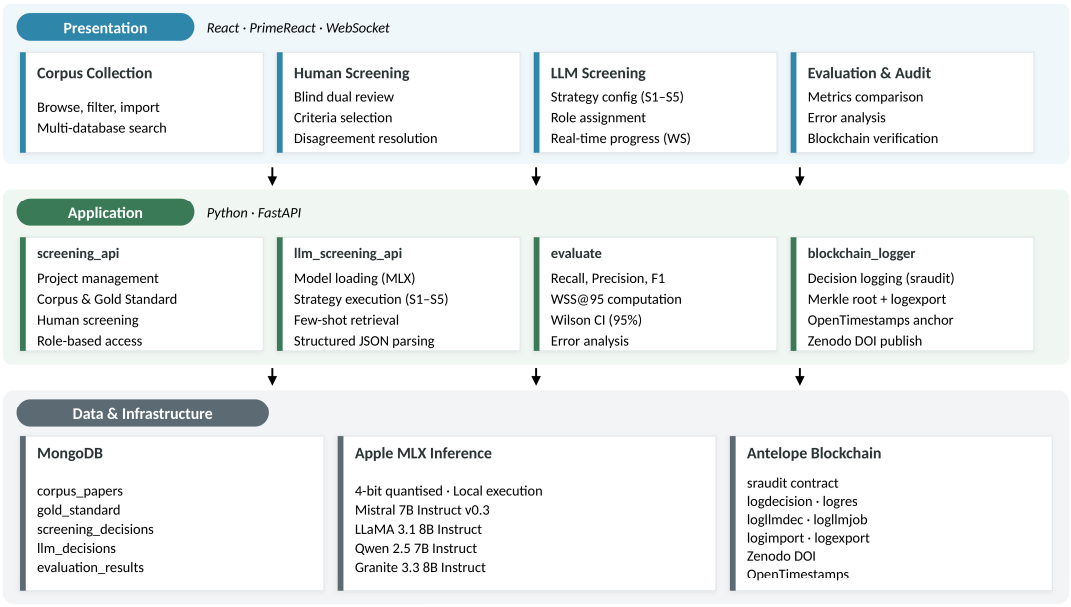


Figure 3. PaSSER-SR system architecture.

The platform enforced a strict separation between calibration and evaluation data. Papers marked as few-shot calibration examples (is_calibration = true) were automatically excluded from LLM screening evaluation runs. Human screening followed a dual independent review protocol: two screeners assessed each paper independently, and disagreements were resolved by a third reviewer with the resolver role. For full corpus screening jobs, the evaluation_only parameter ensured that the 10 calibration papers were excluded from metric computation, preventing data leakage from few-shot examples into the evaluation set. All decisions, including timestamps, criteria selections, and confidence levels, were logged to both MongoDB and the Antelope blockchain.

All experimental configurations, prompts, and evaluation procedures are available in the project repository, enabling full reproducibility of the reported results.

3.9. Blockchain Audit Trail

A three-tier verification model was implemented to ensure the reproducibility and integrity of all screening decisions.

Tier 1: Operational logging (Antelope blockchain). Every screening event was recorded as an immutable transaction on a private Antelope blockchain instance via the sraudit smart contract. Six

 doi:10.20944/preprints202603.2107.v1

action types were defined: logdecision (human screening decision), logres (disagreement resolution), logimport (corpus import event), logexport (audit export event), logllmdec (individual LLM screening decision), and logllmjob (LLM screening job metadata). For human decisions, the logged fields included the screener account, project identifier, paper identifier, decision (INCLUDE, EXCLUDE, or UNCERTAIN), confidence level, and a SHA-256 hash of the full decision payload. For LLM decisions, the model's name, strategy, and job identifier were additionally recorded. Each transaction received a unique blockchain transaction ID, linking the database record to its on-chain counterpart.

Tier 2: Temporal anchoring (OpenTimestamps). At defined project milestones, all accumulated decisions and resolutions were exported as a single JSON file. A Merkle tree was constructed from the individual decision records, and the resulting Merkle root was computed. The export file, its SHA-256 hash, and the Merkle root were recorded on the Antelope blockchain via the logexport action. The file was then submitted to OpenTimestamps, which created a timestamp proof anchored to the Bitcoin blockchain. This established an independent, publicly verifiable proof that the screening data existed at a specific point in time.

Tier 3: Archival publication (Zenodo). The complete audit export file, together with the OpenTimestamps proof (.ots file), was deposited on Zenodo [35], receiving a persistent Digital Object Identifier (DOI). This provided long-term public access to the full screening log, enabling external parties to independently verify the Merkle root, reconstruct the decision history, and confirm data integrity against the blockchain records. Detailed verification instructions are provided in Appendix A.

The audit trail design builds on a broader framework for internal audit and control procedures in blockchain-based systems, where verification, segregation of duties, and risk assessment are identified as essential components of blockchain adoption [34]. The combination of private operational logging, Bitcoin-anchored temporal proofs, and DOI-based archival publication addressed a common limitation of private blockchains: the absence of external verifiability. The Antelope blockchain recorded each screening decision as a tamper-evident transaction; the OpenTimestamps anchor and the published Zenodo archive allowed any third party to verify that the data had not been altered after the reported timestamp. The complete audit log, including human and LLM screening decisions with their transaction identifiers, the Gold Standard and Full Corpus evaluation results, and the final inclusion list for full-text review, has been deposited on Zenodo [35].

4. Results

This section presents the empirical results obtained from the corpus construction pipeline, the Gold Standard screening, and the evaluation of five LLM screening strategies across four open-source models. A total of 25 configurations (12 zero-shot and 13 few-shot) were evaluated against the Gold Standard of 200 papers. The primary acceptance criterion was recall ≥ 0.95, ensuring that at least 95% of relevant studies were retained.

4.1. Corpus Statistics

Papers were collected from five open access databases using the search query described in Section 3.2. The raw record counts per database (Table 3, column "Raw Records") correspond to the retrieval statistics reported in Table 1. Table 3 extends this with deduplication results and cross-database overlap. "Unique to DB" indicates records found exclusively in one database. "Shared" indicates records found in the given database and at least one other. "% of Corpus" is calculated as After Dedup / 4,021. Percentages exceed 100% because shared papers are counted under each contributing database.

 doi:10.20944/preprints202603.2107.v1

Table 3. Distribution of records across databases before and after deduplication.

Database	Raw Records	After Dedup	Unique to DB	Shared	% of Corpus
OpenAlex	1,644	1,554	811	743	38.6
Semantic Scholar	2,392	2,289	1,511	778	56.9
CORE	848	809	689	120	20.1
arXiv	259	259	130	129	6.4
MDPI	86	86	23	63	2.1
Total	5,229	4,021	3,164	857	—

A total of 5,229 raw records were retrieved. After deduplication by DOI matching and title similarity (≥ 0.85 threshold), the unified corpus contained 4,021 unique papers. Of these, 857 (21.3%) appeared in two or more databases, confirming the value of multi-source search for comprehensive coverage.

Semantic Scholar contributed the largest share of the corpus (56.9%), followed by OpenAlex (38.6%) and CORE (20.1%). The arXiv and MDPI databases contributed smaller but complementary sets of 259 and 86 records, respectively. A total of 1,511 papers (37.6%) were found exclusively in Semantic Scholar, indicating that reliance on a single database would have resulted in substantial omissions.

Electoral keyword filtering (Section 3.3) reduced the corpus from 4,021 to 2,036 papers. The excluded 1,985 papers lacked e-voting context in their title, abstract, or keywords. The Gold Standard sample of 200 papers was drawn from the filtered corpus using simple random sampling with a fixed seed (Section 3.4).

4.2. Gold Standard and Inter-Rater Agreement

Two independent reviewers screened all 200 Gold Standard papers independently against the inclusion and exclusion criteria defined in Table 2. Disagreements were resolved through discussion and, where necessary, adjudication by a third reviewer. Table 4 presents the inter-rater reliability metrics. Cohen's κ was interpreted according to the Landis and Koch scale [36].

Table 4. Inter-rater reliability metrics for the Gold Standard screening.

Metric	Value
Total papers screened (dual review)	200
Observed agreement (Po)	0.750 (75.0%)
Expected agreement (Pe)	0.485 (48.5%)
Cohen's κ	0.515
PABAK	0.500
Interpretation (Landis and Koch)	Moderate
Agreed INCLUDE	58
Agreed EXCLUDE	92
Agreed UNCERTAIN	0
Disagreements	50
Resolved by third reviewer	50

Cohen's κ of 0.515 indicates moderate inter-rater agreement, with an observed agreement of 75.0% and a PABAK of 0.500. This level of agreement is consistent with the interdisciplinary nature of the domain, where terminology such as „voting", „consensus", and „governance" carries different meanings across blockchain, e-voting, and governance contexts.

Table 5 presents the agreement matrix between the two reviewers. The largest source of disagreement was between INCLUDE and EXCLUDE decisions: 14 papers were classified as INCLUDE by Reviewer 1 but EXCLUDE by Reviewer 2, while 23 papers showed the reverse pattern.

 doi:10.20944/preprints202603.2107.v1

A further 13 disagreements involved the UNCERTAIN category. All 50 disagreements were resolved by a third reviewer.

Table 5. Agreement matrix between Reviewer 1 and Reviewer 2 (n = 200).

	R2: INCLUDE	R2: EXCLUDE	R2: UNCERTAIN	Total
R1: INCLUDE	58	14	3	75
R1: EXCLUDE	23	92	8	123
R1: UNCERTAIN	0	2	0	2
Total	81	108	11	200

After disagreement resolution, the final Gold Standard comprised 200 papers: 67 classified as INCLUDE (33.5%), 126 as EXCLUDE (63.0%), and 7 as UNCERTAIN (3.5%). Table 6 presents the final distribution and the mapping to LLM evaluation ground truth.

Table 6. Final Gold Standard decision distribution after disagreement resolution.

Decision	Count	Percentage	LLM Ground Truth
INCLUDE	67	33.5%	INCLUDE (67)
EXCLUDE	126	63.0%	EXCLUDE (126)
UNCERTAIN	7	3.5%	INCLUDE (7)
Total	200	100%	74 INC / 126 EXC

The UNCERTAIN papers were treated as INCLUDE for LLM evaluation purposes, following a conservative approach that maximises the recall requirement. This yielded an effective evaluation set of 74 positive (INCLUDE + UNCERTAIN) and 126 negative (EXCLUDE) cases for zero-shot evaluation (N = 200). For few-shot evaluation, 10 calibration papers were excluded, resulting in 69 positive and 121 negative cases (N = 190)

4.2.1. Zero-Shot Screening Results

A total of 28 strategy–model–prompt configurations were tested (14 zero-shot, 14 few-shot). Three S5 configurations were excluded from the main results: two (L→Q+M in both prompt modes) did not complete screening for all papers due to inference timeouts, and one (Q→G, zero-shot) used only two models instead of three, representing a degenerate case. The remaining 25 configurations are reported below. Table 7 presents the performance of 12 zero-shot configurations across five strategies and four models. Each configuration was evaluated on the full Gold Standard of 200 papers (N = 200). All metrics reported as percentages except TP, FP, FN, TN (counts). Status: ✓ = recall ≥ 95% (qualified); ✗ = recall < 95% (unqualified). WSS@95 values are reported for all configurations but are meaningful only when recall ≥ 0.95; values for unqualified configurations (✗) should not be compared against qualified ones. Ensemble abbreviations: M = Mistral 7B, L = LLaMA 3.1 8B, Q = Qwen 2.5 7B, G = Granite 3.3 8B. For S5, the notation X → Y+Z indicates Gate model → Debate panel.

Table 7. Zero-shot screening performance on the Gold Standard (N = 200).

Str.	Model	Rec. %	Prec. %	F1%	WSS@95%	TP	FP	FN	TN	Stat.
S1	Mistral 7B	98.7	52.1	68.2	25.0	73	67	1	59	✓
	LLaMA 3.1 8B	100.0	59.7	74.8	33.0	74	50	0	76	✓
	Qwen 2.5 7B	94.6	72.9	82.3	47.0	70	26	4	100	✗
	Granite 3.3 8B	98.7	36.7	53.5	−4.5	73	126	1	0	✓
S2	MLG	100.0	51.7	68.2	23.5	74	69	0	57	✓
	MLQ	100.0	60.2	75.1	33.5	74	49	0	77	✓

 doi:10.20944/preprints202603.2107.v1

S3	MLQ	100.0	51.7	68.2	23.5	74	69	0	57	✓
S4	MLG	100.0	52.1	68.5	24.0	74	68	0	58	✓
	MLQ	100.0	60.2	75.1	33.5	74	49	0	77	✓
S5	G → M+L	100.0	52.1	68.5	24.0	74	68	0	58	✓
	L → M+G	100.0	59.7	74.8	33.0	74	50	0	76	✓
	M → L+Q	98.7	59.8	74.5	34.0	73	49	1	77	✓

Of the 12 zero-shot configurations, 11 met the recall threshold of 95%. The single unqualified configuration was S1 with Qwen 2.5 7B, which achieved a recall of 94.6% (FN = 4). Despite failing the recall criterion, this configuration achieved the highest precision (72.9%) and F1 score (82.3%) among all zero-shot runs, demonstrating a clear recall–precision trade-off.

Granite 3.3 8B exhibited a distinct failure pattern. Under S1, it classified 199 out of 200 papers as INCLUDE (TN = 0, FP = 126), yielding a negative WSS@95 of −4.5%. This indicates that Granite showed no discriminative power in this setting and performed no better than including all papers without screening. This behaviour propagated to all ensemble strategies containing Granite (S2 MLG, S4 MLG, S5 G→M+L), which consistently underperformed their Qwen-based counterparts.

The best zero-shot performance by WSS@95 was S5 (M → L+Q) at 34.0%, closely followed by S2 MLQ and S4 MLQ at 33.5%. An identical result pattern was observed between S2 MLQ and S4 MLQ across all metrics, suggesting that the confidence-weighted aggregation in S4 produced no benefit over simple majority voting in S2 for this dataset.

The following section examines whether few-shot prompting addresses these zero-shot limitations, particularly Qwen's recall failure and Granite's lack of discriminative power.

4.2.2. Few-Shot Screening Results

Table 8 presents the results of 13 few-shot configurations. Each was evaluated on N = 190 papers, as 10 papers used for calibration examples (Section 3.6) were excluded from the evaluation set to prevent data leakage. Notation follows Table 7. All 13 configurations met the recall ≥ 95% threshold.

Table 8. Few-shot screening performance (N = 190).

Str.	Model	Rec. %	Prec. %	F1%	WSS@95%	TP	FP	FN	TN	Stat.
S1	Mistral 7B	100.0	48.6	65.4	20.3	69	73	0	48	✓
	LLaMA 3.1 8B	98.6	56.2	71.6	31.3	68	53	1	68	✓
	Qwen 2.5 7B	100.0	70.4	82.6	43.4	69	29	0	92	✓
	Granite 3.3 8B	100.0	36.3	53.3	−5.0	69	121	0	0	✓
S2	MLG	100.0	47.6	64.5	18.7	69	76	0	45	✓
	MLQ	100.0	58.0	73.4	32.4	69	50	0	71	✓
S3	MLG	100.0	36.3	53.3	−5.0	69	121	0	0	✓
	MLQ	100.0	47.3	64.2	18.2	69	77	0	44	✓
S4	MLG	100.0	47.6	64.5	18.7	69	76	0	45	✓
	MLQ	100.0	58.0	73.4	32.4	69	50	0	71	✓
S5	Q → M+L	100.0	61.6	76.2	36.0	69	43	0	78	✓
	M → L+Q	100.0	58.0	73.4	32.4	69	50	0	71	✓
	L → M+Q	98.6	57.6	72.7	32.9	68	50	1	71	✓

 doi:10.20944/preprints202603.2107.v1

All 13 few-shot configurations met the recall threshold. The best overall configuration was S1 with Qwen 2.5 7B (few-shot), which achieved perfect recall (100.0%), precision of 70.4%, F1 of 82.6%, and WSS@95 of 43.4%. This was the highest-ranked configuration across all 25 tested combinations.

The most notable few-shot effect was observed for Qwen 2.5 7B under S1. In zero-shot mode, Qwen was the only model that failed the recall threshold (94.6%, FN = 4). With few-shot prompting, recall increased to 100.0% while precision remained high (70.4% vs. 72.9% in zero-shot). The few-shot examples effectively corrected the overly conservative exclusion pattern that caused the zero-shot failure.

Granite 3.3 8B showed no improvement with few-shot prompting. Under S1, it again classified nearly all papers as INCLUDE (FP = 121, TN = 0), and the same pattern appeared under S3 MLG (FP = 121, TN = 0). Granite's inability to discriminate between relevant and irrelevant papers persisted regardless of prompt mode.

Across all matched strategy–model pairs, few-shot prompting produced mixed effects on precision. For Mistral 7B, precision decreased in most configurations (e.g., S1: 52.1% → 48.6%), suggesting that the few-shot examples introduced a bias towards inclusion. For LLaMA 3.1 8B, precision remained comparable (e.g., S1: 59.7% → 56.2%). Only Qwen 2.5 7B showed a consistent beneficial pattern where the recall improvement outweighed the modest precision decrease.

4.2.3. Strategy Comparison and Ranking

Table 9 presents the top five configurations ranked by a composite criterion: recall ≥ 95% as a hard threshold, then F1 score as the primary ranking metric. Abbreviations: ZS = zero-shot, FS = few-shot. Wilson 95% confidence intervals are shown for Recall and Precision.

Table 9. Top five configurations by F1 score.

Rank	Str.	Model(s)	Mode	Rec. %	Prec. % [95% CI]	F1%	WSS@95%
1	S1	Qwen 2.5 7B	FS	100.0	70.4 [60.7–78.5]	82.6	43.4
2	S5	Q → M+L	FS	100.0	61.6 [52.4–70.1]	76.2	36.0
3	S5	M → L+Q	ZS	98.7	59.8 [51.0–68.1]	74.5	34.0
4	S2	MLQ	ZS	100.0	60.2 [51.3–68.4]	75.1	33.5
5	S4	MLQ	ZS	100.0	60.2 [51.3–68.4]	75.1	33.5

The single-agent baseline (S1) achieved the highest rank, followed by two S5 (two-stage) configurations and S2/S4 with identical metrics.

Figure 4 illustrates the performance trends across all five strategies. In panels (a) and (b), S1 and S5 consistently outperform S2, S3, and S4 in both F1 and WSS@95. Few-shot prompting improved S1 and S5 but provided marginal or negative gains for S2–S4. Panel (c) confirms that all strategies maintained recall above 98.5% in few-shot mode, indicating that the primary differentiator was precision, not recall. Each bar represents the best-qualified configuration per prompt mode. Best zero-shot / few-shot models per strategy: S1 — LLaMA / Qwen; S2, S3, S4 — M+L+Q / M+L+Q; S5 — L+M+G / Q+M+L. The dashed red line in (c) marks the 95% recall threshold.

Three observations emerge from the 25 tested configurations. First, configurations including Qwen consistently outperformed those using Granite, regardless of strategy. Second, the equivalence of S2 and S4 observed in Section 4.2.1 persisted across all tested ensembles. Third, multi-agent strategies did not consistently outperform the single-agent baseline — the best S1 configuration surpassed all multi-agent alternatives. The 95% Wilson confidence intervals for precision overlapped across the top five configurations (Table 9), suggesting that the observed ranking differences may not be statistically robust at the current sample size.

 doi:10.20944/preprints202603.2107.v1

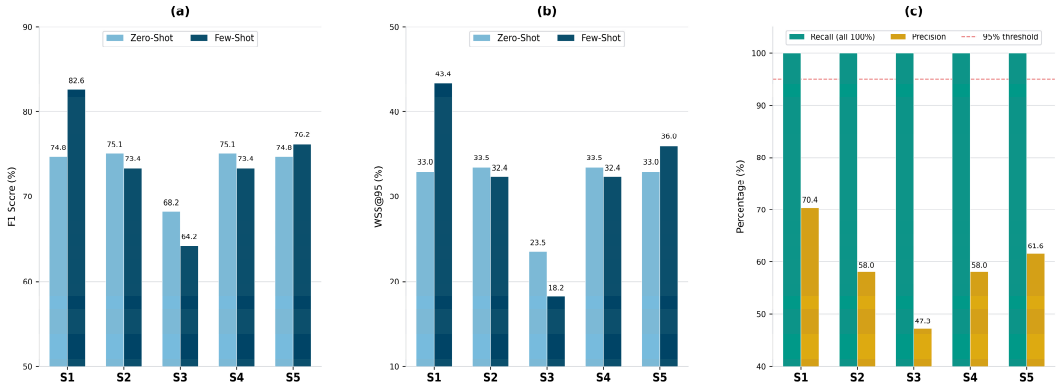


Figure 4. Performance metrics across LLM coordination strategies.

4.3. Full Corpus Screening

To assess whether screening performance generalises beyond the Gold Standard, the top five configurations from Table 9 were applied to the full corpus of 2,036 papers. Evaluation metrics were computed on the 190 Gold Standard evaluation papers embedded within the corpus (Section 3.7). Table 10 presents the screening results. GS-embedded metrics computed on N = 190 (calibration excluded). Included = papers selected for full-text review from the full corpus of 2,036 (INCLUDE + UNCERTAIN). FP and FN = GS-embedded error counts. For Ranks 1 and 2 (few-shot, N = 190), all confusion matrix counts are identical to Table 8. For Ranks 3–5 (zero-shot), the metrics differ from Table 7 due to the exclusion of calibration papers (N = 190 vs N = 200).

Table 10. Full corpus screening results.

Rank	Str.	Model(s)	Mode	Rec. %	Prec. %	F1%	WSS@95%	Included	FP	FN
1	S1	Qwen 2.5 7B	FS	100.0	70.4	82.6	43.4	950	29	0
2	S5	Q → M+L	FS	100.0	61.6	76.2	36.0	1,131	43	0
3	S5	M → L+Q	ZS	100.0	61.1	75.8	35.5	1,121	44	0
4	S2	MLQ	ZS	100.0	61.1	75.8	35.5	1,125	44	0
5	S4	MLQ	ZS	100.0	61.1	75.8	35.5	1,124	44	0

All five configurations maintained 100% recall on the embedded Gold Standard papers, confirming that no relevant studies were lost during full corpus screening. The single false negative observed for S5 M → L+Q ZS in the GS-only evaluation (Table 7, N = 200) was a calibration paper excluded from the N = 190 set.

Cross-strategy agreement on the candidate set was high: 926 of the 950 papers selected by S1 (97.5%) were also selected by all four multi-agent configurations. Only 2 papers were unique to S1, while 140 papers were selected by all multi-agent strategies but excluded by S1. Whether these 140 papers represent genuine inclusions or additional false positives cannot be determined without full-text review. The three zero-shot multi-agent configurations (Ranks 3–5) produced nearly identical INCLUDE sets, differing on fewer than 5 papers.

The 950 papers selected by the top-ranked configuration (of which one was classified as uncertain and excluded from the final review list) constitute the candidate set for full-text review.

 doi:10.20944/preprints202603.2107.v1

4.4. Error Analysis

Error analysis was performed on all 25 Gold Standard configurations using a consistent evaluation set of N = 190 papers (calibration examples excluded). The analysis was then extended to the five full corpus configurations to assess whether error patterns persisted at scale.

False positive persistence. Across 25 GS configurations, 52 unique papers were flagged as false positives at least once. The distribution was highly concentrated: a persistent pool of 20 papers appeared in the FP set of all five strategies, while 13 papers appeared exclusively in S1 configurations. Multi-agent screening did not introduce new error types – no paper appeared as a false positive only in a multi-agent strategy. Figure 5 illustrates the FP overlap structure for both GS and full corpus evaluations. Panels (a) and (b) show the GS results aggregated across all 25 configurations (N = 190): the persistent pool of 20 papers common to all five strategies is clearly separated from strategy-specific false positives. Panels (c) and (d) show the full corpus evaluation (five top-ranked configurations, N = 190 GS-embedded papers), where this concentration intensified – 28 of 48 unique false positives were shared across all five configurations. Dark shading indicates papers appearing as FP in all strategies; red indicates papers unique to S1.

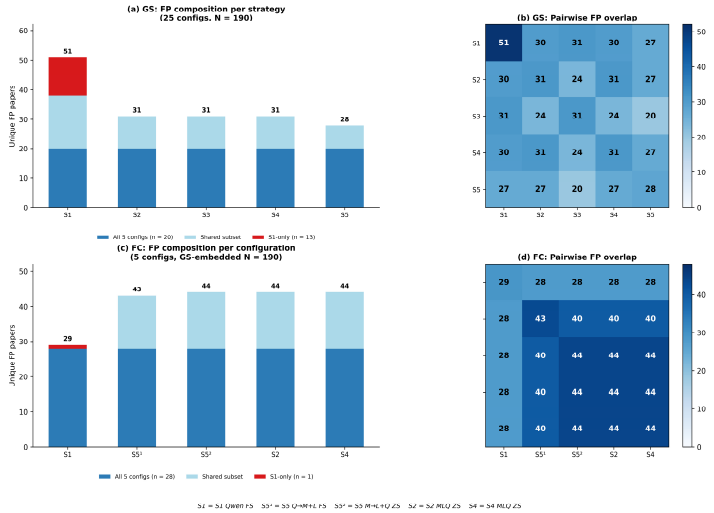


Figure 5. False positive overlap across LLM coordination strategies.

Exclusion criteria application. An inverse relationship was observed between EC citation frequency and false positive count (Table 11). Granite cited no exclusion criteria in either mode, consistent with keyword matching rather than criteria-based reasoning. Qwen cited EC most frequently and produced the fewest false positives.

Table 11. False positive counts and exclusion criteria citations per model (S1, N = 190).

Model	FP (ZS)	FP (FS)	EC cited (ZS)	EC cited (FS)
Granite 3.3 8B	121	121	0	0
Mistral 7B	62	73	30	84
LLaMA 3.1 8B	45	53	163	270
Qwen 2.5 7B	25	29	248	365

Few-shot prompting effect. Few-shot prompting increased EC citation counts across all models capable of criteria-based reasoning (Table 11), while Granite remained at zero. However, few-shot prompting also produced a slight FP increase across all models. The calibration examples promoted more thorough criteria matching but did not prevent inclusion of papers that superficially satisfy the inclusion criteria. The net effect is improved criteria articulation without consistent precision gains.

 doi:10.20944/preprints202603.2107.v1

Full corpus error categories. The five full corpus configurations produced 48 unique false positives on the GS-embedded set, of which 28 appeared in all five configurations. Table 12 categorises these persistent false positives by exclusion criterion. The categories are based on the exclusion criteria cited by both human reviewers in agreed cases, or by the third reviewer during disagreement resolution.

Table 12. Exclusion categories for false positives common to all five full corpus configurations.

Category	Count	Example
EC3: No original contribution (opinion, overview, survey)	19	"Disrupting the Ballot Box: Blockchain as a Catalyst for Innovation in Electoral Processes"
EC1: No blockchain implementation	6	"Privacy Preserving E-Voting System Using Homomorphic Encryption"
EC5: Non-English abstract	2	"Implementasi Sistem E-Voting Berbasis Blockchain..."
EC2: Non-electoral domain	1	"Controllable anonymous authentication scheme based on blockchain..."
Total	28	

Table 12 reveals that the dominant error category was EC3 (no original contribution), accounting for 19 of the 28 persistent false positives. These papers typically present surveys, tutorials, or position papers whose abstracts contain all expected inclusion keywords but offer no original system, framework, or empirical evaluation. The remaining categories — EC1 (no blockchain implementation), EC5 (non-English abstract), and EC2 (non-electoral domain) — together account for 9 papers. This distribution confirms that the primary precision bottleneck is not terminological ambiguity per se, but the inability of the models to distinguish the type of contribution from its topic based on title and abstract alone.

False negatives. Only two FN cases occurred across all 25 GS configurations, both in single-agent S1 (Granite ZS and LLaMA FS). All multi-agent configurations and all five full corpus configurations recorded zero false negatives. The error analysis confirms that false positive accumulation, not false negative risk, is the primary source of screening uncertainty.

These findings address RQ3 by identifying consistent error patterns across models and strategies.

4.5. Summary of Results

Table 13 consolidates the evaluation metrics for the top five configurations from the Gold Standard (GS) evaluation and the full corpus (FC) screening. The GS columns report performance on the fixed evaluation sets (N = 190 for few-shot, N = 200 for zero-shot), while the FC columns report performance on the same 190 GS papers embedded within the full corpus of 2,036 records.

Table 13. Summary comparison of top-ranked configurations on Gold Standard and full corpus.

		Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Str.		S1	S5	S5	S2	S4
Config		Qwen 2.5 7B	Q→M+L	M→L+Q	MLQ	MLQ
Mode		FS	FS	ZS	ZS	ZS
Rec%	GS	100.0	100.0	98.7	100.0	100.0
	FC	100.0	100.0	100.0	100.0	100.0
F1%	GS	82.6	76.2	74.5	75.1	75.1
	FC	82.6	76.2	75.8	75.8	75.8
WSS@95%	GS	43.4	36.0	34.0	33.5	33.5
	FC	43.4	36.0	35.5	35.5	35.5
FP	GS	29	43	49	49	49

 doi:10.20944/preprints202603.2107.v1

	FC	29	43	44	44	44
FN	GS	0	0	1	0	0
	FC	0	0	0	0	0

For the two few-shot configurations (Ranks 1–2), all GS and FC metrics are identical, confirming that screening behaviour remained stable when the evaluation set was embedded in the full corpus. For the three zero-shot configurations (Ranks 3–5), FC metrics differ from GS because 10 calibration papers — which contain 5 false positives and 1 false negative — were excluded from the FC evaluation (N = 190 vs N = 200). This exclusion accounts for the improved FP counts (49 → 44) and, for Rank 3 (S5 M→L+Q ZS), the increase in recall from 98.7% to 100.0%. This comparison is illustrated in Figure 6 (Section 5.1).

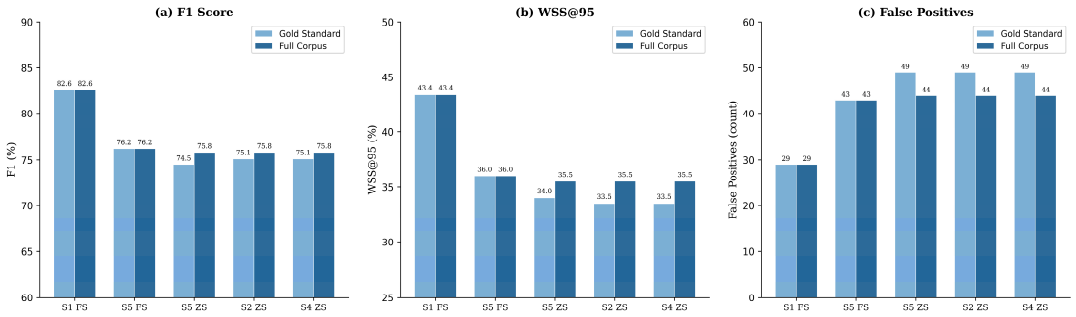


Figure 6. Comparison of top-ranked configurations on Gold Standard (GS) and full corpus (FC): (a) F1 score, (b) WSS@95, and (c) false positive counts.

Table 13 confirms that the ranking established on the Gold Standard was preserved in the full corpus evaluation. The single-agent baseline maintained its advantage across all metrics, while the multi-agent configurations produced comparable results among themselves. The persistent false positive pool identified in Section 4.4 represents a precision ceiling inherent to title-and-abstract screening, where distinguishing original contributions from general discussions exceeds the discriminative capacity of the tested models. These results address RQ1 and RQ2 by showing that model selection has a stronger impact on performance than coordination strategy.

5. Discussion

The experimental results addressed the three research questions posed in Section 1.3. This section interprets the findings, examines their implications, and acknowledges the boundaries of the current study.

5.1. Model Capability vs. Coordination Complexity

These findings should be interpreted within the context of the selected domain. Blockchain-based e-voting represents a terminologically overloaded setting, which amplifies ambiguity in title-and-abstract screening. The observed pattern is consistent across all tested configurations. Its generalisability to other domains remains to be validated. Configurations containing Qwen 2.5 7B consistently outperformed alternatives regardless of the coordination strategy applied. The single-agent baseline (S1) with Qwen in few-shot mode achieved the highest F1 (82.6%) and WSS@95 (43.4%), outperforming every multi-agent alternative. Figure 6 illustrates this comparison across the top five configurations. The identical GS and FC values for the two few-shot configurations confirm that screening behaviour remained stable at full corpus scale, while the single-agent baseline maintained a clear advantage across all metrics.

This outcome does not support the assumption that ensemble coordination compensates for individual model weaknesses. When one model performs better, adding weaker models reduces

 doi:10.20944/preprints202603.2107.v1

overall quality. S2 (majority voting) and S4 (confidence-weighted aggregation) produced identical results across all metrics. This indicates that self-reported confidence from 7–8B parameter models does not add discriminative value. This pattern is consistent with findings from a recent multi-agent screening study using API-based models, where majority voting also outperformed more complex adjudication and debate strategies [23]. S5 (two-stage debate) offered computational efficiency by filtering up to 38.4% of papers in Stage 1 (an 18% inference saving), but did not improve precision. S3 (recall-focused OR) amplified false positives without recall gains.

These results align with recent findings on multi-agent coordination with models of comparable size. A comparative evaluation of four coordination strategies against single-agent RAG baselines reported consistent performance degradation, with coordination overhead identified as the primary factor [22]. The present study extends this observation from question answering to binary screening. This suggests that coordination overhead is a general limitation at the 7–8B scale. This finding is consistent with a recent evaluation of 18 LLMs across three biomedical systematic reviews [27], where model rankings varied substantially across domains — Mistral achieved the highest inter-rater agreement (PABAK = 0.621) in clinical reviews, whereas Qwen dominated in the present interdisciplinary domain — suggesting that model superiority is domain-dependent rather than absolute. Notably, smaller models (llama3.1:8b, MCC = 0.302) outperformed their larger counterparts (llama3.1:70b, MCC = 0.242) in that study, reinforcing the observation that parameter count alone does not determine screening quality at this scale. Whether larger models (13B–70B) benefit from coordination remains an open question. Greater reasoning diversity among agents could make ensemble deliberation productive at higher parameter scales.

The 95% Wilson confidence intervals for precision overlapped across the top five configurations (Table 9), so the apparent ranking may not be statistically robust at N = 190. Given the sample size (N = 190), these differences should be interpreted with caution, as the available data do not provide sufficient statistical power to establish significant performance gaps.

5.2. Error Patterns and the Precision Ceiling

Error analysis (RQ3) showed that false positive accumulation was the dominant source of screening error. Only two false negatives were observed across all configurations, while a substantial number of papers were repeatedly classified as false positives across strategies.

The moderate inter-rater agreement ($\kappa = 0.515$) indicates that a portion of disagreement arises from inherent ambiguity in the classification task. This is consistent with prior observations that screening agreement decreases in interdisciplinary domains with overlapping terminology [24]. In such cases, differences between model predictions and ground truth may reflect uncertainty in the labels rather than purely model error.

The persistent false positives identified in Section 4.4 matched inclusion keywords but did not meet the contribution requirements. This suggests that distinguishing contribution type from topic may exceed the information available in titles and abstracts alone.

The relationship between exclusion criteria usage and false positive rates supports this interpretation. Models that explicitly applied exclusion criteria produced fewer false positives, while models that appeared to rely on surface-level matching showed limited discriminative power. This observation aligns with prior work showing that the formulation and application of inclusion and exclusion criteria strongly influence screening outcomes [27].

5.3. Implications for Practice

For practical applications of LLM-assisted screening, differences between models produced substantially larger performance variations than differences between strategies.

For well-defined domains, a single capable model with few-shot prompting may be sufficient. The observed workload reduction (WSS@95 = 43.4%) is comparable to ranges reported in prior studies of LLM-assisted screening [27], despite the absence of retrieval augmentation in the present framework.

 doi:10.20944/preprints202603.2107.v1

Multi-agent strategies remain relevant in cases where no single model achieves acceptable recall. In such settings, majority voting provides a simple approach, while two-stage strategies may offer computational savings.

The framework itself remains domain-independent. Applying it to a new domain requires redefining inclusion and exclusion criteria and constructing a domain-specific Gold Standard. Existing reporting frameworks such as RDAL [12] and PRISMA-trAIce [13] address transparency in AI-assisted reviews, while the blockchain-based audit mechanism proposed here complements these approaches by providing infrastructure-level decision traceability.

5.4. Limitations

The framework was validated on a single domain. The results should therefore be interpreted as domain-specific, and generalisability to other domains remains to be established.

The Gold Standard of 200 papers (190 for evaluation) limits the statistical power of precision comparisons. The few-shot calibration examples were drawn from the same corpus, which may constitute indirect data leakage despite the exclusion of all 10 calibration papers from evaluation. The sample size also limits statistical power, particularly when comparing configurations with overlapping confidence intervals.

All models tested were 7–8B parameter variants on consumer hardware. Larger models may benefit more from multi-agent coordination, where greater parameter capacity could support productive ensemble deliberation. Recent evidence from API-based models (GPT-4o Mini, Claude 3 Haiku, Gemini 1.5 Flash) supports this hypothesis, as multi-agent collaboration yielded consistent improvements over individual baselines at higher parameter scales [23]. The inter-rater agreement ($\kappa = 0.515$) introduces uncertainty in the ground truth labels, which may affect the interpretation of model performance. Reporting results separately on agreed and disputed labels would provide a clearer picture.

6. Conclusions

This study evaluated five LLM coordination strategies for automating title-and-abstract screening tasks – single-agent baseline (S1), majority voting (S2), recall-focused ensemble (S3), confidence-weighted aggregation (S4), and two-stage screening with debate (S5) – using four open-source 7–8B parameter models deployed locally on consumer hardware. The evaluation was conducted on a Gold Standard of 200 papers from a corpus of 2,036 records in the terminologically overloaded domain of blockchain-based e-voting.

The findings should be interpreted within the context of the selected domain and require validation in other application areas.

Three principal findings emerged, each addressing one of the research questions posed in Section 1.3.

RQ1 (coordination strategy effect): Model selection was the primary determinant of screening performance, outweighing strategy selection. The single-agent strategy with Qwen 2.5 7B in few-shot mode achieved the best overall performance (recall = 100.0%, precision = 70.4%, F1 = 82.6%, WSS@95 = 43.4%), outperforming all multi-agent alternatives. Confidence-weighted aggregation (S4) produced results identical to majority voting (S2), indicating that self-reported confidence from 7–8B parameter models does not provide additional discriminative value in this setting.

RQ2 (best strategy–model combination): The single-agent baseline (S1) with Qwen 2.5 7B in few-shot mode achieved the best balance between recall and effort reduction. No multi-agent configuration improved upon this result. The 95% Wilson confidence intervals for precision overlapped across the top five configurations, suggesting limited statistical separation. Multi-agent strategies introduced coordination overhead without measurable benefit under the conditions of this study.

RQ3 (systematic error patterns): The dominant source of screening error was false positive accumulation driven by terminological overlap. Models that actively applied exclusion criteria

 doi:10.20944/preprints202603.2107.v1

(Qwen) produced fewer false positives, while models relying on keyword matching (Granite) showed limited discriminative power. A persistent pool of 28 false positives – predominantly surveys and position papers (EC3) – appeared across all five full corpus configurations, suggesting a practical precision ceiling in title-and-abstract screening.

The blockchain-based audit mechanism combining private chain logging, OpenTimestamps anchoring, and Zenodo archival addresses a documented reproducibility gap in AI-aided screening [12], where decisions are typically recorded only in narrative form.

The main contributions of this study are: (1) a controlled comparison of single-agent and multi-agent LLM coordination strategies using 4-bit quantised 7–8B parameter models deployed locally on consumer hardware; (2) empirical evidence that, at this model scale, a single well-prompted model (Qwen 2.5 7B, few-shot) outperformed all multi-agent alternatives, and that self-reported confidence scores did not add discriminative value; (3) a three-tier blockchain-based audit mechanism combining private chain logging, Bitcoin-anchored temporal proofs, and DOI-based archival publication, addressing the documented reproducibility gap in AI-aided screening; and (4) application and evaluation of the framework in an interdisciplinary, terminologically overloaded domain.

For practitioners, investing in model selection and error-driven few-shot calibration appears more effective than designing complex multi-agent orchestration. A well-prompted single model achieved high recall for structured screening tasks while reducing manual workload by over 40%.

Future work should validate these findings across domains with different terminological characteristics and examine whether larger models benefit from multi-agent coordination. Two technical extensions could further address the persistent false positive pool: retrieval-augmented generation (RAG) could provide full-text access during screening, and active learning could iteratively refine the decision boundary by selecting the most informative papers for human review. The inclusion and exclusion criteria (Table 2) were designed specifically for the present domain. Future applications should examine which criteria contribute most to screening accuracy and whether the current set can be refined without loss of discriminative capacity. Both directions build directly on the PaSSER-SR infrastructure.

Author Contributions: Conceptualization, I.R. and T.N.; methodology, I.R.; software, I.R. and T.N.; validation, I.P., L.D. and T.N.; formal analysis, I.R.; investigation, T.N. and L.D.; resources, I.R., T.N. and L.D; data curation, I.R. and T.N.; writing—original draft preparation, I.R. and T.N.; writing—review and editing, I.R. and I.P; visualization, T.N.; supervision, I.R., I.P. and L.D.; project administration, L.D., I.R. and T.N.; funding acquisition, L.D. and T.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Centre of Competence Digitization of the economy in an environment of Big data, BG05M2OP001-1.002-0002-C05, OP SESG.

Data Availability Statement: The screening audit log, Gold Standard evaluation results, and blockchain verification files are available on Zenodo at https://doi.org/10.5281/zenodo.19182242 [35]. The source code for the PaSSER-SR platform and the experimental scripts are available on GitHub at https://github.com/scpdxtest/PaSSER-SR.

Acknowledgments: During the preparation of this manuscript, the authors used Claude (Anthropic, Opus 4.6) for the purposes of verifying data consistency between tables and source files, checking alignment between paper descriptions and source code implementations. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

 doi:10.20944/preprints202603.2107.v1

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	
CI	Confidence Interval
EC1–EC6	Exclusion Criteria 1–6
FC	Full Corpus
FN	False Negative
FP	False Positive
FS	Few-Shot
GS	Gold Standard
IC1–IC5	Inclusion Criteria 1–5
LLM	Large Language Model
MLX	Apple Machine Learning Framework
OTS	OpenTimestamps
PABAK	Prevalence-Adjusted Bias-Adjusted Kappa
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval-Augmented Generation
TN	True Negative
TP	True Positive
WSS@95	Work Saved over Sampling at 95% Recall
ZS	Zero-Shot

Appendix A. Audit Verification Instructions

The screening audit log was deposited on Zenodo [35]. The deposit contains two files: the full decision log (audit_export.json) and the OpenTimestamps proof (audit_export.json.ots). These files allow independent verification of the screening data without access to the private Antelope blockchain.

Requirements: The OpenTimestamps client is required for Step 2.
It is available at https://opentimestamps.org. No additional software is required.

Verification Steps:

1. **Download the files.** Both files (`audit_export.json` and `audit_export.json.ots`) are downloaded from the Zenodo record at https://doi.org/10.5281/zenodo.19182242.
2. **Verify the timestamp proof.** The .ots file is verified using the following command: ots verify `audit_export.json.ots`. The client automatically computes the SHA-256 hash of `audit_export.json`, applies the commitment operations encoded in the .ots proof, and checks the result against the Bitcoin blockchain. A successful result confirms that the file existed in its current form prior to the reported timestamp. An internet connection is required.
3. **Inspect the decision log.** The JSON file contains one record per screening decision. Each record includes the paper identifier, screener account, decision (INCLUDE, EXCLUDE, or UNCERTAIN), confidence level, criteria selections, timestamp, and Antelope blockchain transaction ID. Human and LLM decisions are stored in separate arrays within the same file.
4. ***(Optional)* Manual hash verification.** The SHA-256 hash of `audit_export.json` may be independently computed for additional assurance: `sha256sum audit_export.json`. This hash corresponds to the initial commitment in the .ots proof and may be compared against the value reported in the Zenodo record metadata.

Note on the Antelope Blockchain

The Antelope blockchain transaction IDs are included in the JSON log for reference. The Antelope blockchain is a private instance and is not publicly accessible. Independent verification of the data is performed through Step 2, which relies on the Bitcoin-anchored timestamp and the published Zenodo archive.

 doi:10.20944/preprints202603.2107.v1

References

1. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, doi:10.1136/bmj.n71.
2. Cohen, A.M.; Hersh, W.R.; Peterson, K.; Yen, P.-Y. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *Journal of the American Medical Informatics Association* **2006**, *13*, 206–219, doi:https://doi.org/10.1197/jamia.M1929.
3. Guo, E.; Gupta, M.; Deng, J.; Park, Y.-J.; Paget, M.; Naugler, C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *Journal of Medical Internet Research* **2024**, *26*, e48996, doi:10.2196/48996.
4. Akinseloyin, O.; Jiang, X.; Palade, V. A Question-Answering Framework for Automated Abstract Screening Using Large Language Models. *Journal of the American Medical Informatics Association* **2024**, *31*, 1939–1952, doi:10.1093/jamia/ocae166.
5. Wang, Z.; Nayfeh, T.; Tetzlaff, J.; O'Blenis, P.; Murad, M.H. Error Rates of Human Reviewers during Abstract Screening in Systematic Reviews. *PLOS ONE* **2020**, *15*, e0227742, doi:10.1371/journal.pone.0227742.
6. van de Schoot, R.; de Bruin, J.; Schram, R.; Zahedi, P.; de Boer, J.; Weijdema, F.; Kramer, B.; Huijts, M.; Hoogerwerf, M.; Ferdinands, G.; et al. An Open Source Machine Learning Framework for Efficient and Transparent Systematic Reviews. *Nature Machine Intelligence* **2021**, *3*, 125–133, doi:10.1038/s42256-020-00287-7.
7. Syriani, E.; Dávid, I.; Kumar, G.A. Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. *ArXiv* **2023**, *abs/2307.06464*.
8. 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*, 616–626, doi:https://doi.org/10.1002/jrsm.1715.
9. Ye, A.; Maiti, A.; Schmidt, M.; Pedersen, S.J. A Hybrid Semi-Automated Workflow for Systematic and Literature Review Processes with Large Language Model Analysis. *Future Internet* **2024**, *16*, doi:10.3390/fi16050167.
10. Galli, C.; Gavrilova, A.V.; Calciolari, E. Large Language Models in Systematic Review Screening: Opportunities, Challenges, and Methodological Considerations. *Information* **2025**, *16*, doi:10.3390/info16050378.
11. Matsui, K.; Utsumi, T.; Aoki, Y.; Maruki, T.; Takeshima, M.; Takaesu, Y. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. *Journal of Medical Internet Research* **2024**, *26*, doi:https://doi.org/10.2196/52758.
12. Lombaers, P.; de Bruin, J.; van de Schoot, R. Reproducibility and Data Storage for Active Learning-Aided Systematic Reviews. *Applied Sciences* **2024**, *14*, doi:10.3390/app14093842.
13. Holst, D.; Moenck, K.; Koch, J.; Schmedemann, O.; Schüppstuhl, T. Transparent Reporting of AI in Systematic Literature Reviews: Development of the PRISMA-trAIce Checklist. *JMIR AI* **2025**, *4*, e80247, doi:10.2196/80247.
14. Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N.V.; Wiest, O.; Zhang, X. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. In Proceedings of the Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24; Larson, K., Ed.; International Joint Conferences on Artificial Intelligence Organization, August 2024; pp. 8048–8057.
15. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.H.; Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In Proceedings of the ArXiv; 2022; Vol. abs/2203.11171.
16. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning; JMLR.org: Vienna, Austria, july 21 - 27, 2024, 2024; pp. 11733–11763.

 doi:10.20944/preprints202603.2107.v1

17. Yang, Y.; Ma, Y.; Feng, H.; Cheng, Y.; Han, Z. Minimizing Hallucinations and Communication Costs: Adversarial Debate and Voting Mechanisms in LLM-Based Multi-Agents. *Applied Sciences* **2025**, *15*, doi:10.3390/app15073676.
18. Yeow Lee, X.; Akatsuka, S.; Vidyaratne, L.; Kumar, A.; Farahat, A.; Gupta, C. Reliable Decision-Making for Multi-Agent LLM Systems. In Proceedings of the Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI-25); Philadelphia, PA, USA, March 25 2025.
19. Li, J.; Zhang, Q.; Yu, Y.; Fu, Q.; Ye, D. More Agents Is All You Need. *Transactions on Machine Learning Research* **2024**, 1–18.
20. Bernasconi, E.; Redavid, D.; Ferilli, S. Integrated Survey Classification and Trend Analysis via LLMs: An Ensemble Approach for Robust Literature Synthesis. *Electronics* **2025**, *14*, doi:10.3390/electronics14173404.
21. Mienye, I.D.; Swart, T.G. Ensemble Large Language Models: A Survey. *Information* **2025**, *16*, doi:10.3390/info16080688.
22. Radeva, I.; Popchev, I.; Doukovska, L.; Dimitrova, M. Multi-Agent Coordination Strategies vs. Retrieval-Augmented Generation in LLMs: A Comparative Evaluation. *Electronics* **2025**, *14*, doi:10.3390/electronics14244883.
23. Akinseloyin, O.; Jiang, X.; Palade, V. Large Language Model-Based Multiagent Collaboration for Abstract Screening toward Automated Systematic Reviews. *Biology Methods and Protocols* **2026**, *11*, bpag006, doi:10.1093/biomethods/bpag006.
24. Belur, J.; Tompson, L.; Thornton, A.; Simon, M. Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. *Sociological Methods & Research* **2021**, *50*, 837–865, doi:10.1177/0049124118799372.
25. Hanegraaf, P.; Wondimu, A.; Mosselman, J.J.; de Jong, R.; Abogunrin, S.; Queiros, L.; Lane, M.; Postma, M.J.; Boersma, C.; van der Schans, J. Inter-Reviewer Reliability of Human Literature Reviewing and Implications for the Introduction of Machine-Assisted Systematic Reviews: A Mixed-Methods Review. *BMJ Open* **2024**, *14*, doi:10.1136/bmjopen-2023-076912.
26. Langenhuijsen, L.F.S.; Janse, R.J.; Venema, E.; Kent, D.M.; van Diepen, M.; Dekker, F.W.; Steyerberg, E.W.; de Jong, Y. Systematic Metareview of Prediction Studies Demonstrates Stable Trends in Bias and Low PROBAST Inter-Rater Agreement. *Journal of Clinical Epidemiology* **2023**, *159*, 159–173, doi:10.1016/j.jclinepi.2023.04.012.
27. Delgado-Chaves, F.M.; Jennings, M.J.; Atalaia, A.; Wolff, J.; Horvath, R.; Mamdouh, Z.M.; Baumbach, J.; Baumbach, L. Transforming Literature Screening: The Emerging Role of Large Language Models in Systematic Reviews. *Proceedings of the National Academy of Sciences* **2025**, *122*, e2411962122, doi:10.1073/pnas.2411962122.
28. Kulothungan, V. Using Blockchain Ledgers to Record AI Decisions in IoT. *IoT* **2025**, *6*, doi:10.3390/iot6030037.
29. Radeva, I. *Blockchain Integration: Development and Implementation*; 1st ed.; Printing office of Prof. Marin Drinov Publishing House of Bulgarian Academy of Sciences: Sofia, 2024; ISBN 978-619-245-473-9.
30. Radeva, I.; Popchev, I.; Doukovska, L.; Dimitrova, M. Web Application for Retrieval-Augmented Generation: Implementation and Testing. *Electronics* **2024**, *13*, doi:10.3390/electronics13071361.
31. Kusa, W.; Lipani, A.; Knoth, P.; Hanbury, A. An Analysis of Work Saved over Sampling in the Evaluation of Automated Citation Screening in Systematic Literature Reviews. *Intelligent Systems with Applications* **2023**, *18*, 200193, doi:https://doi.org/10.1016/j.iswa.2023.200193.
32. Wilson, E.B. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association* **1927**, *22*, 209–212, doi:10.1080/01621459.1927.10502953.
33. Agresti, A.; Coull, B.A. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician* **1998**, *52*, 119–126, doi:10.1080/00031305.1998.10480550.
34. Radeva, I. Blockchains: Practical Approaches. *Engineering Sciences* **2022**, *59*, 79–92, doi:10.7546/EngSci.LIX.22.01.01.
35. Radeva, I.; Noncheva, T.; Doukovska, L.; Popchev, I. Blockchain-Verified Audit Trail: Automated Title-Abstract Screening. Zenodo **2025**. Available online: https://doi.org/10.5281/zenodo.19182242..

 doi:10.20944/preprints202603.2107.v1

36. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174.