Linguistic Polarity and Decision Architecture in Large Language Model–Based Abstract Screening in the Dental Field

Amir M. Behrouzian , Marco Meleti , Maria Teresa Colangelo , Elena Calciolari , Carlo Galli *

Posted Date: 18 March 2026

doi: 10.20944/preprints202603.1440.v1

Keywords: large language models; systematic reviews; negation; affirmation; pubmed screening

 doi:10.20944/preprints202603.1440.v1

Article

Linguistic Polarity and Decision Architecture in Large Language Model–Based Abstract Screening in the Dental Field

Amir M. Behrouzian [1], Marco Meleti [2], Maria Teresa Colangelo [1], Elena Calciolari [2,3] and Carlo Galli [1,*]

[1] Histology and Embryology Laboratory, Department of Medicine and Surgery, University of Parma, Via Volturno 39, 43126 Parma, Italy

[2] Department of Medicine and Surgery, Dental School, University of Parma, 43126 Parma, Italy

[3] Centre for Oral Clinical Research, Institute of Dentistry, Faculty of Medicine and Dentistry, Queen Mary University of London, London E1 2AD, UK

* Correspondence: carlo.galli@unipr.it

Abstract

Large language models (LLMs) are increasingly investigated for abstract screening in systematic reviews, yet it remains unclear whether screening errors attributed to linguistic complexity reflect intrinsic semantic limitations or the decision architecture in which the model is embedded. We investigated how five polarity variants of logically equivalent eligibility criteria—affirmative inclusion, antonymic exclusion, predicate negation, verb-level negation, and double negation—affect screening outcomes in a controlled biomedical task. Using 1,000 abstracts derived from a reconstructed Cochrane review corpus (50 eligible TARGET studies; 950 non-targets), we implemented four abstract-visible criteria within a sequential hard-gated pipeline, where failure at any step triggered irreversible exclusion. Under hard gating, linguistic polarity alone produced substantial sensitivity shifts. For GPT-5.1, recall ranged from 0.72 to 0.32 despite identical logical predicates and input data. Replication with GPT-3.5 Turbo yielded a similar polarity-dependent divergence (recall range 0.92–0.18), confirming that the effect generalizes across model generations. TARGET losses were highly concentrated at criteria frequently satisfied but inconsistently reported in abstracts, consistent with conservative exclusion under evidential under-specification. To assess whether this effect was semantic or architectural, we reimplemented screening using a scoring-based evidence-accumulation framework in which each criterion contributed graded support (YES/NO/UNCLEAR) and inclusion was determined by a tunable score threshold. Scoring substantially reduced polarity-driven recall divergence and transformed it into an explicit precision–recall trade-off. These findings indicate that negation sensitivity in LLM screening is strongly mediated by decision architecture: irreversible Boolean gating amplifies linguistic asymmetries under uncertainty, whereas cumulative scoring preserves uncertainty and enables controllable operating points.

Keywords: large language models; systematic reviews; negation; affirmation; pubmed screening

1. Introduction

Systematic reviews depend on rigorous title and abstract screening, yet this stage remains one of the most labor-intensive and cognitively demanding components of evidence synthesis [1]. Reviewers must evaluate thousands of records against multiple eligibility criteria, often under substantial time constraints [2]. Even with standardized protocols, screening decisions are vulnerable to fatigue, interpretive variability, and inconsistency, particularly when eligibility rules involve nuanced methodological or clinical distinctions [3].

 doi:10.20944/preprints202603.1440.v1

Automation has therefore been pursued through supervised machine learning, rule-based natural language processing, and hybrid approaches [4], with a few systems also being available beyond the experimental stage [5–7]. While these approaches can reduce screening workload, they typically require labelled training corpora and rely on feature representations derived from titles and abstracts. Such feature-based systems operate primarily on surface lexical information and may struggle when eligibility decisions depend on implicit assumptions, contextual interpretation, or information not explicitly stated in the abstract [8].

Large language models (LLMs) offer a qualitatively different paradigm [9]. Rather than learning task-specific classifiers, they can apply eligibility criteria directly from natural-language instructions, potentially enabling flexible, domain-adaptive screening without retraining [10]. Early applications in systematic review workflows have demonstrated promising sensitivity, but they also reveal substantial variability across prompt formulations and decision setups [11–15]. This variability raises a central methodological question: when LLM screening fails, are errors primarily semantic—reflecting difficulty with logical structures such as negation—or are they shaped by the procedural framework within which decisions are enforced?

Negation provides a natural test case. In biomedical writing, negation frequently alters truth conditions in ways that are straightforward for human reviewers yet challenging for automated systems [16]. Eligibility criteria such as "exclude studies that did not include adult participants" or "do not include trials without at least 6 months of follow-up" introduce scope-sensitive logical transformations. Prior work has treated such constructions primarily as a linguistic problem. However, screening is not merely a matter of sentence interpretation; it is also a decision process.

In typical systematic review workflows, criteria are implemented sequentially in a hard-gated manner: failure at any step results in immediate and irreversible exclusion [17]. This architecture imposes early commitment under uncertainty. When abstracts omit information that is commonly satisfied but not explicitly reported—for example, participant age or methodological details—an LLM must infer eligibility from incomplete evidence [18]. Under a fail-closed regime, absence of explicit confirmation may be treated as evidence of ineligibility. Small linguistic asymmetries, including negation scope or framing differences, may therefore be amplified into large recall losses by the irreversibility of the pipeline itself. More broadly, recent work on LLM benchmarking emphasizes the importance of systematic evaluation frameworks for identifying structural vulnerabilities in AI systems, including interactions between model behavior and evaluation design [19]. The present study contributes to this effort by demonstrating how linguistic formulation and decision architecture jointly determine screening outcomes in high-recall filtering tasks.

This study examines whether logically equivalent eligibility criteria expressed with different polarity—affirmative inclusion, exclusion framing, predicate negation, verb-level negation, and double negation—yield systematically different screening outcomes, and whether such polarity effects persist when irreversibility is removed. Using a reconstructed corpus derived from a published Cochrane review [20], we compare two architectures: a sequential hard-gated pipeline that mirrors conventional screening logic, and a scoring-based evidence-accumulation framework in which each criterion contributes graded support (YES/NO/UNCLEAR) and inclusion is determined by a tunable threshold.

By holding logical content constant while varying only surface formulation and decision architecture, this design allows us to disentangle linguistic sensitivity from procedural amplification — and to clarify how architectural design choices shape the reliability of high-recall filtering in evidence synthesis.

2. Materials and Methods

2.1. Reconstruction of the Reference Systematic Review and Corpus Assembly

This study evaluates how linguistic formulations specifically affirmative versus negated phrasing of eligibility criteria—affect the performance of large language models (LLMs) during

 doi:10.20944/preprints202603.1440.v1

abstract screening. To ground the evaluation in a realistic screening environment, we selected the 2024 Cochrane review on adjunctive antimicrobial photodynamic therapy for periodontal and peri-implant diseases as the reference standard, which includes 50 randomized controlled trials (RCTs) constituting the final set of eligible studies [20]. These 50 RCTs were designated as the *target articles* for all subsequent analyses (Table A1).

To recreate the initial decision space of the published review, we replicated the original search strategy as described in the Cochrane Methods section, though the search was limited to the Medline database. All retrieved citations were de-duplicated and merged into a master corpus. Each record was then assigned a binary label—1 for target articles included in the Cochrane review and 0 for all other records—based on title, author, and metadata verification. The reconstructed corpus thus contained the full set of non-target articles originally screened by the review authors, alongside the 50 known target studies.

2.2. Construction of Experimentally Controlled Screening Datasets

Given the size of the reconstructed corpus (approximately 6000 non-target citations), we created smaller, experimentally controlled datasets of 1,000 abstracts each — comprising the same 50 target articles and 950 non-target articles sampled from the full corpus — to enable tractable automated screening.

For the hard-gated screening experiments, a single dataset configuration was used to evaluate the effects of linguistic polarity under identical conditions. For the scoring-based experiments, additional datasets were generated by resampling the 950 non-target articles while keeping the 50 target articles fixed. These datasets served as development, validation, and test sets for selecting and evaluating score thresholds under recall constraints. This separation allowed threshold selection to be performed independently of the final evaluation dataset, preventing circular optimization of operating points.

2.3. Selection and Reformulation of Abstract-Visible Eligibility Criteria

The Cochrane review specifies a broad set of eligibility criteria, many of which concern methodological or clinical details that are not reliably reported in article abstracts [21]. To avoid penalizing the language model, we restricted the screening rules to criteria that (i) were essential for determining eligibility in the original review and (ii) would reasonably be expected to appear in the abstract of a biomedical publication.

Four abstract-visible criteria were therefore selected:

(1) randomized controlled trial design,

(2) adult participants (≥18 years),

(3) follow-up duration ≥1 month, and

(4) diagnosis of periodontitis or peri-implant disease.

Each criterion was reformulated into five linguistically distinct variants that differed only in polarity and scope of negation:

- AI – Affirmative Inclusion
- AE – Antonymic Exclusion
- PN – Predicate Negation
- VN – Verb Negation
- DN – Double Negation

These five formulations constituted the experimental prompt conditions. All other prompt components—including instruction structure, response format, and model parameters—were held constant to isolate the effect of linguistic polarity on screening decisions. Formal representations of the predicates and the construction of the polarity variants are provided in Appendix B.

 doi:10.20944/preprints202603.1440.v1

2.4. Iterative LLM Screening and Performance Evaluation Through Hard Gating

To emulate conventional systematic review workflows, we first implemented a sequential hard-gated screening pipeline. For each linguistic variant (Affirmative Inclusion, Antonymic Exclusion, Predicate Negation, Verb Negation, and Double Negation), the LLM evaluated each abstract against a given eligibility criterion and returned a binary decision ("include" or "exclude"). Responses were formatted as structured JSON outputs containing the decision label, a confidence estimate, and supporting evidence spans extracted from the abstract. Criteria were applied sequentially. The first criterion was evaluated on the full dataset, and only abstracts not excluded at that stage were passed to the next criterion. This process continued until all mandatory criteria had been evaluated. Because the pipeline was hard-gated, exclusion at any step resulted in irreversible removal from subsequent rounds.

For each linguistic formulation and dataset configuration, we recorded the number of TARGET abstracts retained after each criterion. Plotting these counts across screening rounds yielded retention curves illustrating how rapidly eligible studies were eliminated.

After the final screening step, model decisions were compared with the gold-standard TARGET labels. Confusion matrices were computed and standard performance metrics—accuracy, precision, recall, and F1-score—were derived. Because abstract screening prioritizes sensitivity, recall was treated as the primary safety metric [22].

2.5. Scoring-Based Screening Procedure

To disentangle the effects of linguistic polarity from those of irreversible Boolean filtering, we implemented an alternative scoring-based screening architecture. In contrast to hard gating—where failure of any single criterion results in immediate exclusion—the scoring framework evaluates all criteria independently and aggregates their evidential contributions before making an inclusion decision.

For each criterion, the LLM returned one of three labels: YES (criterion clearly satisfied), NO (criterion clearly not satisfied), or UNCLEAR (insufficient information). These labels were mapped to numerical values (YES = 2, UNCLEAR = 1, NO = 0) and summed across the four mandatory criteria, producing a total score ranging from 0 to 8.

This formulation converts eligibility from a strictly conjunctive Boolean rule into an evidence-accumulation process. Rather than triggering automatic exclusion, uncertainty reduces the cumulative score, allowing inclusion to be determined by a tunable threshold. Under this scheme, a score of 4 represents a natural normative threshold, corresponding to the minimal condition that no criterion is explicitly failed (NO = 0) and that each predicate is at least uncertain. This interpretation reflects the conservative rationale of abstract screening, in which absence of explicit confirmation should not automatically trigger exclusion.

For empirical evaluation, thresholds were selected on a development dataset under a predefined constraint of recall ≥ 0.90, reflecting the high-sensitivity priority of abstract screening in systematic reviews. The selected thresholds were then applied unchanged to independent validation and test datasets to assess generalization across varying abstract mixtures.

2.6. Model Configuration and Evaluation Metrics

All experiments were conducted using GPT-5.1 via the OpenAI API under fixed prompt templates and stateless calls. Each invocation evaluated a single abstract against a single eligibility criterion, ensuring independence across decisions.

To assess whether polarity effects observed under hard-gated screening were specific to a single model generation, the complete hard-gated experiment was replicated using GPT-3.5 Turbo. For GPT-3.5 Turbo, identical datasets, eligibility criteria, linguistic variants, prompt structures, decision pipelines, and evaluation metrics were employed. No model fine-tuning was performed. This

 doi:10.20944/preprints202603.1440.v1

replication enables isolation of polarity and architectural effects from model-specific calibration differences.

The prompt was as follows:

"""You are assisting in a systematic review.

You must apply exactly one eligibility criterion, stated below, to decide whether an article should be INCLUDED or EXCLUDED.

CRITERION:
{criterion}

INSTRUCTIONS:
- Read the title and abstract.
- Decide if, based on the information available in the abstract, the article SATISFIES the criterion.
- If the article clearly satisfies the criterion, respond: INCLUDE
- If the article clearly does not satisfy the criterion, respond: EXCLUDE
- If the abstract lacks enough information to decide, respond based on your best judgment, but still choose INCLUDE or EXCLUDE (do NOT answer "uncertain").

Title: {title}
Abstract: {abstract}

Respond with EXACTLY one word:
INCLUDE
or
EXCLUDE
"""

Scoring-based screening experiments were conducted using GPT-5.1 only. Unlike hard-gated screening—which yields a single binary operating point per linguistic variant—the scoring framework permits exploration of a continuous decision surface defined by score thresholds.

The prompt was as follows:

"""You are assisting in a systematic review.

Apply EXACTLY ONE eligibility criterion to the article below.

CRITERION:
{criterion}

Decide whether the abstract indicates that the article SATISFIES the criterion.

Return EXACTLY ONE of the following labels:
- YES (clearly satisfies the criterion)
- NO (clearly does not satisfy the criterion)
- UNCLEAR (insufficient information in the abstract)

Title: {title}
Abstract: {abstract}

Respond with ONLY one label: YES, NO, or UNCLEAR.
"""

 doi:10.20944/preprints202603.1440.v1

Mapping UNCLEAR to an intermediate score (YES = 2, UNCLEAR = 1, NO = 0) explicitly encodes tolerance for evidential under-specification. Abstracts lacking explicit confirmation are penalized relative to clearly eligible studies but are not automatically excluded. Although alternative encodings (e.g., treating UNCLEAR as 0 or 2) would alter threshold geometry, they would not eliminate the architectural distinction between cumulative evidence aggregation and irreversible Boolean filtering.

2.7. Computational Environment and Software Tools

All experiments were conducted in Google Colab [23] as the primary execution platform, running Python 3.12.12. The automated screening pipeline, dataset handling, and evaluation procedures were implemented in Python scripts that interacted with the OpenAI API for model inference.

Data manipulation and dataset construction were performed using the *Pandas* library [24], while numerical operations and score aggregation were handled using *NumPy* [25]. Performance metrics, confusion matrices, and threshold analyses were computed directly within the Python environment.

Visualizations of screening dynamics and model performance—including retention curves, threshold–recall plots, and precision–recall curves—were generated using *Matplotlib* [26]. These plots were subsequently exported for inclusion in the manuscript figures.

All experiments were executed in a stateless manner, with each model call evaluating a single abstract against a single eligibility criterion. The codebase used to generate datasets, run the screening pipelines, compute performance metrics, and produce the figures ensures full reproducibility of the reported results.

3. Results

This section reports the outcomes of abstract screening the hard-gated, sequential screening pipeline (Round 4 final decision and intermediate rounds) and, subsequently, for the evidence-accumulation scoring framework to assess whether the observed effects persist under an alternative decision architecture.

3.1. Final Classification Performance Under Hard-Gated Screening (GPT-5.1)

After completion of all four screening rounds under the hard-gated architecture, substantial performance differences emerged across linguistic variants despite identical logical predicates and identical input data. Overall accuracy remained uniformly high (0.948–0.959), reflecting the strong class imbalance of the corpus (50 TARGETs among 1,000 abstracts; 5% prevalence). However, accuracy proved insensitive to variant-level differences. In contrast, recall varied markedly and provides the most informative measure of screening safety.

As shown in Table 1, the Affirmative Inclusion (AI) formulation retained 36 of 50 TARGET abstracts (recall = 0.72; precision = 0.537; accuracy = 0.955). All exclusion-oriented or negation-based variants exhibited reduced recall relative to AI. Predicate Negation (PN) retained 26 TARGETs (recall = 0.52; precision = 0.481; accuracy = 0.948), while Double Negation (DN) retained 22 TARGETs (recall = 0.44; precision = 0.537; accuracy = 0.953). The largest recall reductions were observed for Antonymic Exclusion (AE), which retained 20 TARGETs (recall = 0.40; precision = 0.645; accuracy = 0.959), and Verb Negation (VN), which retained only 16 TARGETs (recall = 0.32; precision = 0.615; accuracy = 0.956).

 doi:10.20944/preprints202603.1440.v1

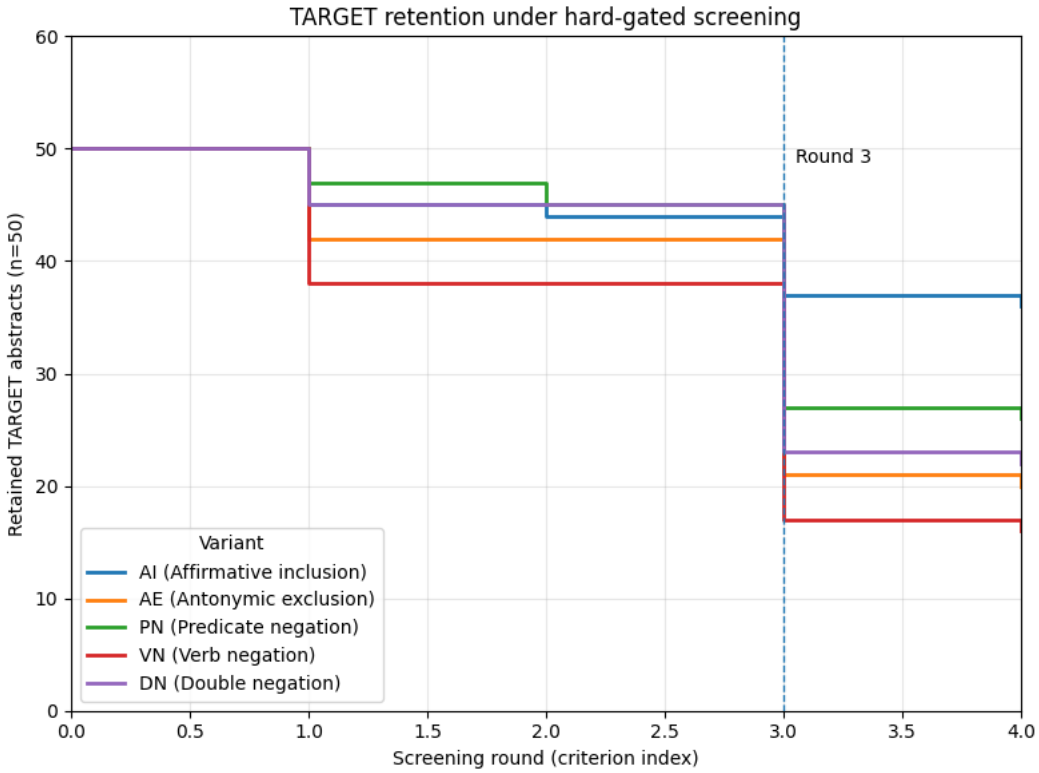
Table 1. Final abstract-level screening performance under hard-gated sequential filtering for five linguistic variants of logically equivalent eligibility criteria. Results are reported for GPT-5.1 on 1,000 abstracts (50 TARGET; 950 non-TARGET). Metrics reflect the final decision after application of all four criteria (Round 4).

Variant	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-score
AI – Affirmative Inclusion	36	31	14	919	0.955	0.537	0.720	0.615
PN – Predicate Negation	26	28	24	922	0.948	0.481	0.520	0.500
DN – Double Negation	22	19	28	931	0.953	0.537	0.440	0.484
AE – Antonymic Exclusion	20	11	30	939	0.959	0.645	0.400	0.494
VN – Verb Negation	16	10	34	940	0.956	0.615	0.320	0.421

Across variants, altering only the surface polarity of logically equivalent criteria produced a 40-percentage-point spread in recall (0.72 vs 0.32), corresponding to 20 additional eligible studies lost under the most exclusion-heavy formulation. Notably, this divergence occurred while overall accuracy varied by only 0.011, underscoring that class imbalance masks substantial differences in screening sensitivity under irreversible filtering.

3.2. Round-Wise Dynamics of TARGET Retention

To identify where eligible studies were lost, recall was computed after each sequential screening round under the hard-gated architecture. Figure 1 presents cumulative TARGET retention across the four criteria for the five linguistic variants (AI, AE, PN, VN, DN).



 doi:10.20944/preprints202603.1440.v1

Figure 1. TARGET retention across successive hard-gated screening rounds under five linguistically distinct formulations of logically equivalent eligibility criteria (AI, AE, PN, VN, DN). Attrition is modest during the first two criteria for most of the formulations but becomes concentrated at the third screening step (vertical dashed line), where divergence between variants emerges. Negation-based formulations (AE, PN, VN, DN) exhibit substantially greater TARGET loss than the affirmative formulation (AI), illustrating how linguistic polarity amplifies recall degradation within an irreversible sequential decision architecture.

TARGET retention followed a consistent structural pattern across all variants, as it remained relatively stable during the first two screening rounds, followed by a pronounced decline at Round 3 and only minor additional attrition at Round 4. This indicates that divergence in final recall originates primarily from a single intermediate decision step rather than gradual degradation across criteria.

For Affirmative Inclusion (AI), recall decreased modestly from 0.90 at Round 1 to 0.88 at Round 2, then dropped to 0.74 at Round 3 and stabilized at 0.72 after Round 4. The largest single-step loss therefore occurred between Rounds 2 and 3 (Δ recall = −0.14).

All negation-based variants exhibited the same discontinuity but with substantially larger Round-3 losses. Antonymic Exclusion (AE) declined from 0.84 at Round 2 to 0.42 at Round 3 (Δ = −0.42), Predicate Negation (PN) from 0.90 to 0.54 (Δ = −0.36), Verb Negation (VN) from 0.76 to 0.34 (Δ = −0.42), and Double Negation (DN) from 0.90 to 0.46 (Δ = −0.44). In every case, the Round-3 decrement exceeded the combined losses across all other transitions.

This pattern demonstrates that recall collapse is not uniformly distributed across criteria but is concentrated at a specific filtering step. Under irreversible hard-gating, polarity differences at that step are amplified into substantial differences in final retention.

3.3. Criterion-Level Analysis of TARGET Exclusions

To identify which predicates drove the recall collapse observed under hard-gated screening, TARGET abstracts excluded at each screening round were examined using the criterion-level exclusion logs. Table 2 summarizes the distribution of eligible studies lost at each predicate across the five linguistic variants (AI, AE, PN, VN, DN).

TARGET exclusions were highly concentrated in two predicates, with one clearly dominant. The criterion "Participants ≥ 18 years" accounted for the majority of TARGET losses across all variants: 8 of 14 losses under AI, 21 of 30 under AE, 20 of 24 under PN, 21 of 34 under VN, and 22 of 28 under DN. Importantly, 87–100% of these exclusions occurred at Round 3, corresponding exactly to the discontinuity identified in Section 3.2. For AE, VN, and DN, all adult-related exclusions occurred at Round 3.

A secondary source of TARGET loss was the criterion "Study design is a randomized controlled trial." This predicate contributed 5 losses under AI, 8 under AE, 3 under PN, 12 under VN, and 5 under DN, primarily at Round 1. In contrast, the criteria "Follow-up ≥ 1 month" contributed only isolated losses (one per variant), and "Diagnosis involves periodontitis or peri-implant disease" contributed no TARGET exclusions in any formulation.

Table 2. Distribution of TARGET abstracts excluded by each eligibility criterion under hard-gated screening for five linguistic variants. Counts indicate the number of eligible studies excluded at the criterion where failure occurred.

Failed eligibility criterion	AI	AE	PN	VN	DN
Study design is an RCT	5	8	3	12	5
Participants ≥ 18 years	8	21	20	21	22

 doi:10.20944/preprints202603.1440.v1

Follow-up ≥ 1 month	1	1	1	1	1
Periodontitis / peri-implant diagnosis	0	0	0	0	0
Total TARGET lost	14	30	24	34	28

These findings indicate that recall degradation is not diffusely distributed across predicates but instead driven by concentrated exclusions at a single dominant criterion. The adult-participant predicate encodes a property frequently satisfied in practice but often under-specified at the abstract level. Under irreversible hard gating, absence of explicit confirmation results in immediate exclusion. The magnitude of this effect varies systematically with linguistic formulation, indicating that polarity-sensitive interpretation interacts with evidential under-specification to produce substantial differences in retention. A qualitative inspection of the TARGET abstracts excluded under the Affirmative Inclusion formulation is provided in Supplementary File S1, which illustrates how omissions or indirect wording in abstracts can lead to exclusion despite the studies satisfying the eligibility criteria.

3.4. Cross-Model Replication Under Hard-Gated Screening (GPT-3.5 Turbo)

To evaluate whether the polarity-driven recall divergence observed under hard-gated screening is specific to GPT-5.1 or reflects a more general interaction between linguistic formulation and irreversible decision architecture, we replicated the hard-gated experiment using GPT-3.5 Turbo under identical prompts and datasets.

As shown in Table 3 and Figure C1, GPT-3.5 Turbo exhibited a distinct overall performance profile compared with GPT-5.1, including substantially lower precision across variants. However, the qualitative polarity pattern persisted. Under hard gating, recall varied markedly across linguistic formulations despite identical logical predicates and identical input data.

Affirmative Inclusion (AI) achieved the highest recall (0.92; 46/50 TARGET retained), followed by Double Negation (DN; recall = 0.74), Predicate Negation (PN; recall = 0.58), Verb Negation (VN; recall = 0.30), and Antonymic Exclusion (AE; recall = 0.18). Thus, recall ranged from 0.92 to 0.18—a 74-percentage-point spread—indicating pronounced sensitivity to surface polarity under irreversible filtering. The round-wise retention dynamics for GPT-3.5 Turbo are shown in Appendix Figure C1, which reproduces the same structural pattern observed for GPT-5.1: modest early losses followed by a polarity-sensitive collapse concentrated at the third screening step.

Notably, GPT-3.5 Turbo displayed a generally permissive inclusion bias under AI (FP = 264; precision = 0.148), reflecting limited discriminative specificity. However, as with GPT-5.1, exclusion-oriented formulations reduced false positives at the cost of substantial TARGET loss. AE and VN yielded higher specificity but sharply reduced recall, mirroring the conservative exclusion pattern observed with GPT-5.1.

Despite differences in absolute performance levels between models, both exhibited the same structural behavior: under sequential hard gating, linguistic polarity strongly modulated recall, with negation-heavy formulations amplifying false-negative rates. This cross-model replication supports the interpretation that polarity-driven recall collapse is not idiosyncratic to a specific LLM version but reflects a general interaction between negation-sensitive interpretation and irreversible decision architecture. A qualitative inspection of the TARGET abstracts excluded by GPT 3.5 turbo under the Affirmative Inclusion formulation is provided as Supplementary File S2.

 doi:10.20944/preprints202603.1440.v1

Table 3. Final abstract-level screening performance under hard-gated sequential filtering using GPT-3.5 Turbo (1,000 abstracts; 50 TARGET). Results reflect the final outcome after all four eligibility criteria. Recall varies markedly across polarity variants despite identical logical predicates, indicating substantial sensitivity to linguistic formulation under irreversible decision gating.

Variant	TP	FP	FN	TN	Accuracy	Precision	Recall
AI – Affirmative Inclusion	46	264	4	686	0.732	0.148	0.920
AE – Antonymic Exclusion	9	86	41	864	0.873	0.095	0.180
PN – Predicate Negation	29	119	21	831	0.860	0.196	0.580
VN – Verb Negation	15	135	35	815	0.830	0.100	0.300
DN – Double Negation	37	179	13	771	0.808	0.171	0.740

3.5. Scoring-Based Screening and Architectural Mediation

To assess whether the polarity-driven recall collapse observed under hard-gated screening reflects intrinsic semantic limitations of the model or properties of irreversible decision gating, we re-evaluated the same datasets using a scoring-based architecture. Importantly, neither the logical predicates nor the input abstracts were modified; only the decision aggregation rule was changed.

Under this alternative framework, all four eligibility criteria were applied independently to each abstract. Rather than triggering immediate exclusion upon failure of a single predicate, each criterion contributed to a cumulative score, and inclusion decisions were determined by applying a threshold to the total evidence accumulated across predicates.

This design removes sequential irreversibility while preserving the underlying logical structure. If recall divergence persists under scoring, it would suggest that polarity primarily affects semantic interpretation. Conversely, attenuation of divergence would indicate that the hard-gated architecture amplifies local polarity effects into global recall collapse. The scoring experiment therefore serves as an architectural mediation test, isolating the contribution of decision structure from that of linguistic formulation.

3.5.1. Threshold-Dependent TARGET Retention

Figure 2 shows TARGET retention as a function of the cumulative score threshold under scoring-based screening for the five linguistic variants (AI, AE, PN, VN, DN). Unlike hard-gated screening—which fixes each variant to a single operating point—scoring yields a continuous family of operating points parameterized by the inclusion threshold.

 doi:10.20944/preprints202603.1440.v1

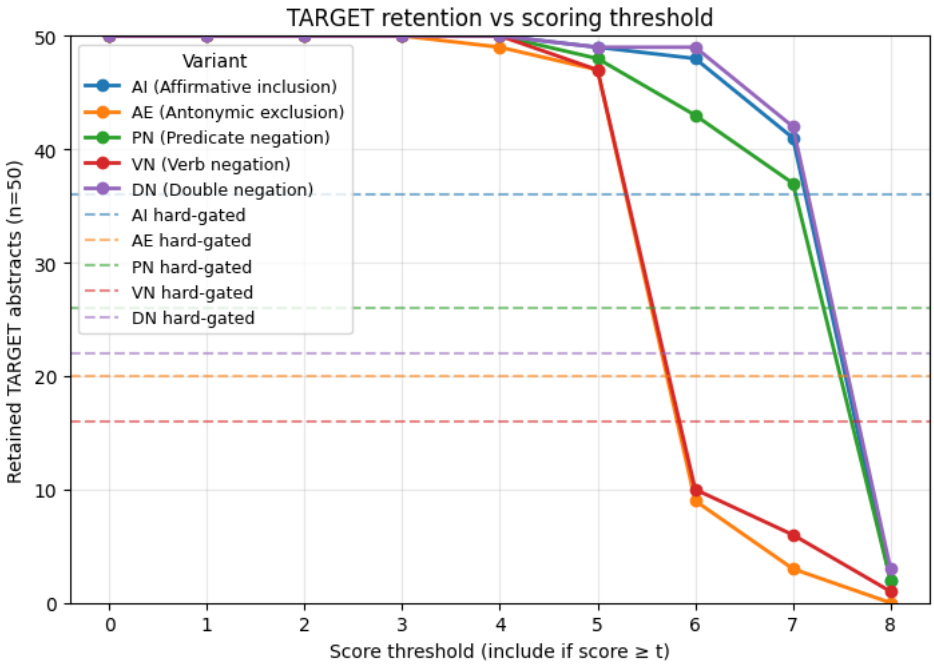


Figure 2. TARGET retention as a function of the cumulative score threshold under scoring-based screening (solid lines), with hard-gated retention shown as horizontal reference lines for the same linguistic variants (AI, AE, PN, VN, DN). Scoring converts screening into a threshold-controlled operating continuum, allowing recall to be tuned via the inclusion cutoff, whereas hard gating fixes each variant to a single operating point determined by irreversible sequential exclusion.

Across variants, TARGET retention increases monotonically as the threshold is lowered, approaching full retention at permissive cutoffs. Crucially, the abrupt recall collapse observed under hard-gated sequential filtering is not reproduced under scoring. Although variant differences remain visible at intermediate thresholds, they are substantially attenuated and manifest as smooth, threshold-controlled trade-offs rather than catastrophic, step-wise attrition. This indicates that evidence aggregation mitigates the amplification of local polarity effects by irreversible exclusion, converting recall loss into an explicit policy choice.

3.5.2. Precision–Recall Behavior Under Scoring-Based Screening

Under scoring-based screening, linguistic variants no longer correspond to fixed operating points but generate continuous precision–recall curves. Figure 3 presents the precision–recall profiles for the five formulations (AI, AE, PN, VN, DN) on the held-out test dataset. In contrast to hard-gated screening—where each variant yields a single precision–recall coordinate—scoring exposes the full operating surface defined by threshold selection.

The AI, PN, and DN variants maintain comparatively favorable precision across moderate-to-high recall levels and achieve areas under the precision–recall curve (AP) of approximately 0.495, 0.448, and 0.477, respectively. In contrast, AE and VN exhibit substantially lower precision across most of the recall range, with markedly reduced AP values (≈0.059 and ≈0.065). These variants therefore incur higher workload at comparable recall levels.

Importantly, however, all variants can achieve high recall under sufficiently permissive thresholds. The catastrophic recall collapse observed under hard-gated sequential screening is not inherent to the semantic content of the criteria; rather, it arises from irreversible exclusion. Under scoring, polarity-driven differences manifest as shifts in curve geometry and workload profile, not as structural barriers to high-sensitivity operation.

 doi:10.20944/preprints202603.1440.v1

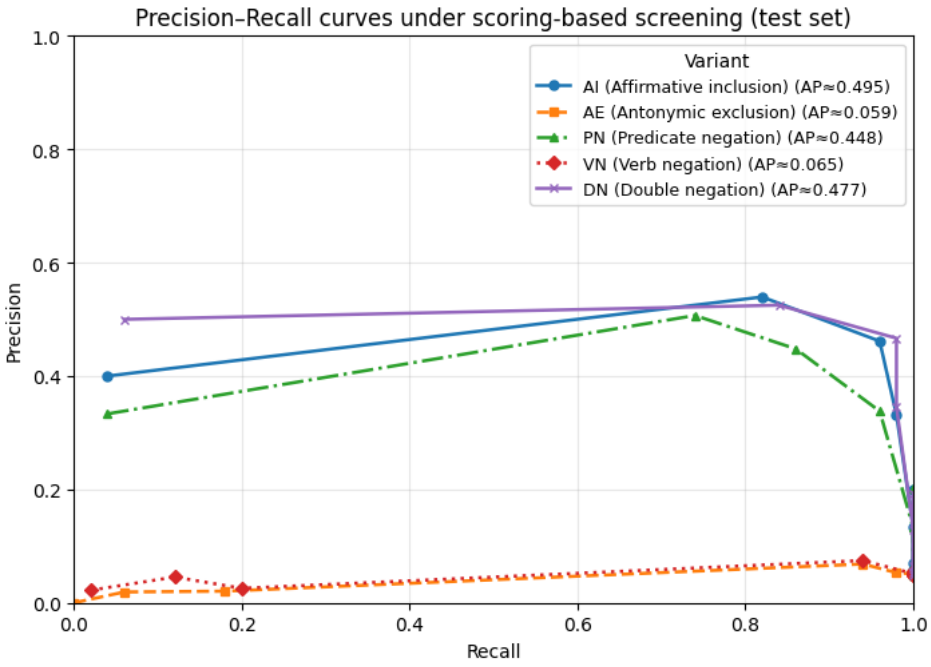


Figure 3. Precision–recall curves under scoring-based screening for five polarity variants of identical eligibility criteria (AI, AE, PN, VN, DN). Unlike hard-gated screening, which fixes each variant to a single operating point, scoring produces a continuous precision–recall spectrum determined by the inclusion threshold. Variants differ primarily in workload profile (precision at comparable recall levels), while high recall remains achievable across formulations.

To further examine how evidence accumulation separates eligible from non-eligible studies, we analyzed the distribution of cumulative scores assigned to TARGET and non-TARGET abstracts across linguistic variants. The corresponding score distributions are shown in Appendix Figure C2, which illustrates how the scoring framework produces a graded separation between TARGET and non-TARGET abstracts rather than a binary exclusion boundary. Across all variants, TARGET abstracts tend to accumulate higher total scores than non-TARGET abstracts, although the degree of separation varies by formulation. Affirmative Inclusion (AI), Predicate Negation (PN), and Double Negation (DN) show clearer separation between the two classes, whereas Antonymic Exclusion (AE) and Verb Negation (VN) exhibit substantial overlap, explaining their lower precision under recall-constrained thresholds. These distributions illustrate how scoring transforms eligibility assessment into a graded evidence signal rather than a sequence of irreversible exclusion decisions.

3.5.3. Direct Comparison with Hard-Gated Screening

Figure 4 provides a direct visual comparison of recall obtained under the hard-gated (HG) and scoring-based (SC) screening architectures for GPT-5.1 across the five linguistic variants (AI, AE, PN, VN, DN). Under hard gating, recall varied widely—from 0.72 for Affirmative Inclusion (AI) to 0.32 for Verb Negation (VN)—indicating strong sensitivity to surface polarity when decisions are irreversible. In contrast, scoring-based screening markedly increased recall across all variants and substantially reduced the divergence between formulations.

Table 4 reports the corresponding numerical values for this comparison. When thresholds were selected under a recall constraint, recall under scoring increased and converged across variants: all formulations achieved ≥0.86 recall, and four of five exceeded 0.90.

 doi:10.20944/preprints202603.1440.v1

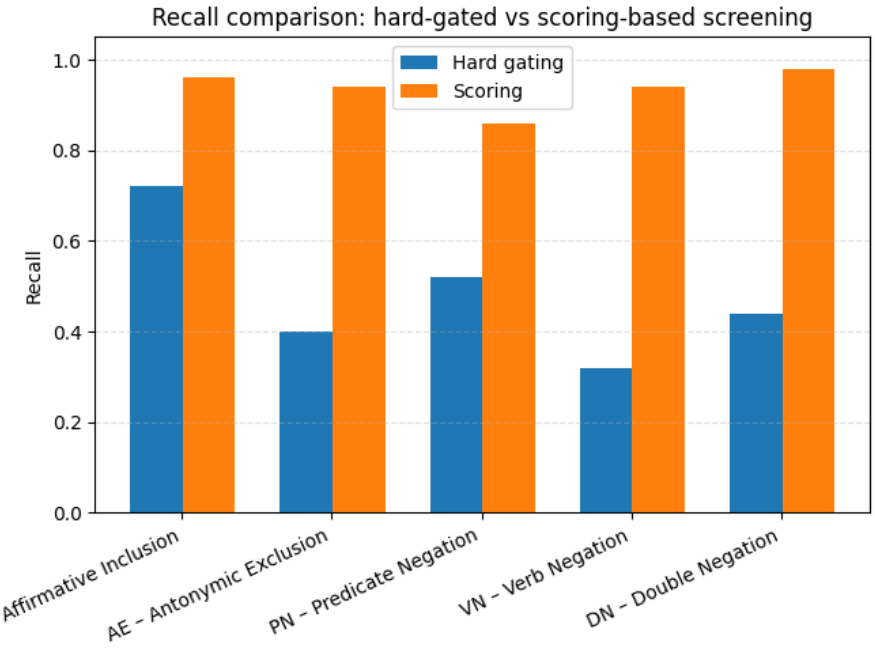


Figure 4. Recall comparison between hard-gated and scoring-based screening architectures. Bars show recall for five linguistic variants of identical eligibility criteria (AI – Affirmative Inclusion, AE – Antonymic Exclusion, PN – Predicate Negation, VN – Verb Negation, DN – Double Negation). Under sequential hard-gated filtering, recall varies widely across formulations because exclusion at any step is irreversible. Under scoring-based screening with recall-constrained thresholds, recall increases and converges across variants, indicating that evidence-accumulation architectures mitigate polarity-driven TARGET loss.

Under scoring, using thresholds selected under a recall constraint, recall increased markedly and converged across variants: all formulations achieved ≥0.86 recall, and four of five exceeded 0.90.

This comparison indicates that the polarity-induced recall collapse is strongly mediated by decision architecture. Linguistic formulation still affects evidence interpretation, but irreversible Boolean filtering amplifies those local differences into large false-negative counts. Evidence accumulation instead converts polarity sensitivity into an explicit operating trade-off: recall can be preserved across variants by adjusting the inclusion threshold, at the cost of increased workload (reduced precision), most prominently for the exclusion-oriented variants AE and VN.

Table 4. Direct comparison of hard-gated (HG) and scoring-based (SC) screening on the test dataset (50 TARGET abstracts) for five linguistic variants of identical eligibility logic (AI, AE, PN, VN, DN). Under hard gating, recall varies substantially across variants (0.32–0.72), reflecting strong polarity sensitivity under irreversible exclusion. Under scoring with recall-constrained thresholds, recall increases and converges across variants (0.86–0.98), substantially reducing false negatives. Precision decreases most sharply for the exclusion-oriented variants (AE and VN), indicating increased workload rather than irreversible loss of eligible studies. These results show that decision architecture strongly mediates the impact of linguistic formulation on screening sensitivity.

Variant	HG Recall	SC Recall	HG FN	SC FN	HG Precision	SC Precision
AI – Affirmative Inclusion	0.72	0.96	14	2	0.537	0.462
AE – Antonymic Exclusion	0.40	0.94	30	3	0.645	0.068

 doi:10.20944/preprints202603.1440.v1

PN – Predicate Negation	0.52	0.86	24	7	0.481	0.448
VN – Verb Negation	0.32	0.94	34	3	0.615	0.075
DN – Double Negation	0.44	0.98	28	1	0.537	0.467

3.6. Threshold Stability Across Datasets

Because the scoring-based screening procedure determines inclusion through a tunable score threshold, threshold selection represents a policy choice rather than an intrinsic property of the model. To avoid circular evaluation and to assess whether threshold choices are robust to variations in corpus composition, thresholds satisfying the predefined recall constraint (recall ≥ 0.90) were selected on a development dataset and subsequently applied unchanged to independently resampled validation and test datasets.

Table 5 summarizes the resulting performance across datasets. For each linguistic variant, the threshold determined on the development dataset produced comparable recall values when applied to the validation and test datasets, despite differences in the mixture of non-target abstracts. Deviations in recall were generally small, indicating that the operating points derived from the development dataset remain stable across alternative distractor compositions.

Table 5. Stability of score thresholds across independently sampled datasets. Thresholds satisfying the recall ≥ 0.90 constraint were selected on the development dataset and applied unchanged to validation and test datasets containing different mixtures of non-target abstracts. The resulting recall values remain comparable across datasets, indicating that the scoring-based screening policy generalizes across variations in distractor composition.

Variant	Threshold	Dev Recall	Validation Recall	Test Recall	Dev Precision	Test Precision
AI – Affirmative Inclusion	6	0.96	0.96	0.96	0.522	0.462
AE – Antonymic Exclusion	5	0.90	0.94	0.94	0.065	0.068
PN – Predicate Negation	6	0.90	0.82	0.86	0.484	0.448
VN – Verb Negation	5	0.94	0.94	0.94	0.071	0.075
DN – Double Negation	6	0.94	0.94	0.98	0.490	0.467

These results suggest that the scoring-based screening policy is not strongly dependent on the particular set of non-target abstracts used during threshold selection. In other words, once an operating threshold is chosen to satisfy a recall constraint, the resulting screening behavior generalizes across independently sampled datasets drawn from the same underlying corpus. This stability supports the use of scoring-based screening as a robust alternative to hard-gated pipelines, which produce fixed and often brittle decision trajectories determined solely by the sequential application of exclusion rules.

 doi:10.20944/preprints202603.1440.v1

4. Discussion

Recent surveys show rapidly expanding use of LLMs across systematic review workflows, particularly for screening and data extraction [27]. LLM-assisted citation screening has shown promising sensitivity and substantial time savings compared with manual screening [28], although reported performance varies considerably depending on task design and evaluation framework [14]. Recent large-scale evaluations of LLM-based abstract screening have likewise shown that model performance is highly sensitive to prompt wording. Experiments across multiple LLMs and prompt formulations demonstrate substantial variation in recall and precision depending solely on how screening instructions are phrased, with prompt bias toward inclusion systematically increasing sensitivity at the expense of precision [29,30].

This study examined how linguistic polarity and decision architecture jointly shape the behavior of large language models during abstract-level screening. By holding logical content constant and varying only the surface realization of eligibility criteria, we isolated the effect of negation independently of dataset composition and model parameters. Across two model generations (GPT-5.1 and GPT-3.5 Turbo), the results consistently showed that linguistic formulation is not a neutral wrapper around logical rules: under sequential hard-gated screening, it produces substantial and systematic differences in recall, error concentration, and exclusion dynamics. Recent evaluations of LLM-assisted literature screening similarly show that prompt design and model choice can substantially alter precision–recall trade-offs across screening tasks [31]. However, most existing studies treat prompting primarily as a way to optimize model performance. The present results suggested that part of this variability may instead arise from the interaction between linguistic form and the decision architecture used to enforce screening rules.

Under the hard-gating paradigm, linguistically equivalent variants yielded markedly different screening trajectories in both models. Affirmative Inclusion (AI) retained the largest proportion of eligible studies, whereas exclusion-oriented and negation-heavy formulations—particularly Antonymic Exclusion (AE) and Verb Negation (VN)—produced pronounced and irreversible TARGET losses. Although absolute calibration differed between GPT-5.1 and GPT-3.5 Turbo, the qualitative polarity pattern replicated across models. This cross-model consistency indicates that the observed recall divergence is not idiosyncratic to a specific LLM version but reflects a general interaction between linguistic polarity and irreversible decision structure.

Criterion-level analysis clarifies the mechanism underlying this effect. TARGET attrition was highly concentrated at a single predicate—participant age—that is frequently satisfied in practice but inconsistently reported at the abstract level. Qualitative inspection of the excluded TARGET abstracts (Supplementary Files S1 and S2) confirms that most false negatives arise from information that is typically implicit or abbreviated in abstracts—particularly participant age and explicit labeling of randomized design—illustrating how evidential under-specification interacts with hard-gated exclusion rules. When explicit confirmation is absent, negation-heavy prompts appear to induce a conservative, fail-closed decision policy. Under formulations such as “exclude if not X” or “do not include unless X,” absence of evidence is treated as evidence of ineligibility. In a hard-gated pipeline, this local bias is amplified by architectural irreversibility: once excluded at an intermediate step, an abstract cannot recover. The collapse therefore arises not from uniform semantic failure, but from the interaction between evidential under-specification and early commitment.

The scoring-based architecture demonstrates that this collapse is strongly mediated by decision structure [27]. When eligibility is treated as an evidence-accumulation problem rather than as a sequence of irreversible Boolean filters, recall becomes a continuous and controllable function of threshold selection. Across linguistic variants, scoring markedly reduced false-negative counts and attenuated polarity-driven divergence. Although differences between formulations persisted—primarily in precision and workload profile—the severe recall collapse observed under hard gating was largely eliminated. The dominant failure mode under negation therefore appears to be premature commitment under uncertainty rather than systematic misinterpretation of the criteria themselves. This architectural mediation is visible in the direct comparison between hard-gated and

 doi:10.20944/preprints202603.1440.v1

scoring-based screening (Figure 4), where recall converges across polarity variants once evidence accumulation replaces irreversible filtering.

Importantly, scoring does not eliminate policy decisions; it makes them explicit. In the experimental setting, thresholds were selected under recall constraints using labeled development data to allow controlled comparison with hard gating. In real-world screening scenarios, where ground-truth recall is unknown, threshold selection must rely on principled deployment strategies. One approach is normative thresholding: in the present scoring scheme (YES = 2, UNCLEAR = 1, NO = 0), a threshold of 4 corresponds to the minimal condition that no criterion is explicitly failed and that each predicate is at least uncertain. This aligns with the conservative philosophy of abstract screening, where absence of explicit confirmation should not automatically trigger exclusion [32]. This approach mirrors prior work on LLM-assisted screening pipelines, where prompt strategies were deliberately biased toward inclusion to minimize false negatives, reflecting the widely accepted principle that missing eligible studies is more damaging than passing additional records to downstream review [29]. Alternatively, scoring enables rank-based workflows in which abstracts are ordered by cumulative evidence and screened progressively, transforming threshold choice into an operational workload decision rather than a semantic inference. Unlike hard gating, threshold selection under scoring does not impose irreversible exclusion at intermediate stages.

These findings highlight a broader methodological distinction. Hard gating implicitly encodes an extreme threshold policy—requiring maximal support on every predicate—and collapses evidential interpretation and decision commitment into a single step. This observation aligns with recent work on LLM evaluation frameworks showing that evaluation criteria themselves often evolve during the inspection of model outputs, a phenomenon described as *criteria drift* [33]. In such settings, the behavior attributed to a model may partly reflect the structure of the evaluation pipeline used to measure it. Our results extend this perspective by demonstrating that architectural choices in screening workflows—such as irreversible gating versus cumulative scoring—can systematically shape observed model performance even when the underlying logical criteria remain unchanged. Evidence-accumulation architectures separate these components, allowing uncertainty to be preserved and managed rather than immediately resolved. In irreversible pipelines, small variations in surface form can produce qualitatively different screening outcomes. In cumulative architectures, the same linguistic sensitivity manifests as shifts along a precision–recall spectrum rather than as catastrophic exclusion.

Several limitations constrain the scope of these conclusions. The analysis was conducted within a single biomedical domain using a fixed set of abstract-visible criteria. Although polarity sensitivity replicated across two model generations, additional models and domains may reveal different quantitative patterns. More complex eligibility structures, interactive reviewer–model workflows, or alternative encodings of may alter the magnitude or direction of the observed trade-offs. Nonetheless, the architectural contrast between irreversible gating and cumulative scoring is independent of domain content and reflects a general property of sequential decision systems operating under linguistic uncertainty.

5. Conclusions

This study demonstrates that the impact of linguistic polarity on LLM-based abstract screening is strongly mediated by decision architecture. Across two model generations (GPT-5.1 and GPT-3.5 Turbo), negation-heavy and exclusion-oriented formulations produced substantial recall divergence when implemented within a sequential hard-gated pipeline. Despite identical logical predicates and dataset composition, irreversible Boolean filtering amplified evidential under-specification into false-negative exclusion, producing large differences in the number of eligible studies retained.

When the same criteria were evaluated within an evidence-accumulation framework, this structural effect was substantially attenuated. Under scoring-based screening, recall became an explicit and adjustable parameter rather than a fixed consequence of linguistic phrasing. Polarity differences persisted as shifts in precision–recall trade-offs and workload profiles but no longer

 doi:10.20944/preprints202603.1440.v1

produced the severe recall collapse observed under hard gating. These results indicate that the dominant failure mode under negation reflects premature commitment under irreversible filtering rather than intrinsic semantic limitations of the model.

More broadly, the findings highlight the importance of workflow design in LLM-assisted evidence synthesis. Hard-gated pipelines implicitly encode extreme threshold policies and amplify local interpretive biases, whereas cumulative scoring separates evidential interpretation from decision commitment and allows uncertainty to be preserved until the final inclusion decision. As a result, architectural choices can substantially influence screening sensitivity even when prompts, datasets, and logical criteria remain unchanged.

Taken together, these results suggest that linguistic form cannot be treated as a neutral interface to logical rules in language-mediated decision systems. Instead, surface structure interacts with procedural constraints to shape screening outcomes in systematic ways. Reliable deployment of LLMs for high-recall filtering tasks therefore requires attention not only to prompt formulation, but also to the design of decision architectures that manage uncertainty explicitly and avoid premature exclusion.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Supplementary File S1: Qualitative analysis of TARGET abstracts incorrectly excluded under the Affirmative Inclusion formulation during hard-gated screening. The file reports the 14 false-negative abstracts together with criterion-level annotations explaining which eligibility predicate triggered exclusion and how abstract-level under-specification contributed to the screening failure. Supplementary File S2: Qualitative analysis of TARGET abstracts incorrectly excluded under the Affirmative Inclusion formulation during hard-gated screening with GPT-3.5 Turbo. The file reports the false-negative abstracts together with criterion-level annotations identifying which eligibility predicate triggered exclusion and how abstract-level under-specification contributed to the screening failure.

Author Contributions: Conceptualization, C.G., M.M. and E.C.; methodology, C.G.; software, C.G.; formal analysis, C.G. and M.T.C.; data curation, A.M.B. and M.T.C.; writing—original draft preparation, C.G. and M.M.; writing—review and editing, A.M.B. and E.C.; All the authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Randomized controlled trials used as TARGET articles in the screening experiments. The table reports the 50 studies included in the reference Cochrane systematic review that were used as the positive class for evaluating LLM-based abstract screening performance. These studies were embedded within larger corpora of non-target abstracts to simulate realistic systematic review screening conditions.

Title	Year	Authors	Journal
Photodynamic therapy as adjunct to non-surgical periodontal treatment in patients on periodontal maintenance: a	2009	Chondros P; Nikolidakis D; Christodoulides N; Rossler R; Gutknecht N; Sculean A	Lasers in Medical Science

 doi:10.20944/preprints202603.1440.v1

randomized controlled clinical trial			
Short-term clinical effects of adjunctive antimicrobial photodynamic therapy in periodontal treatment: a randomized clinical trial	2008	Braun A; Dehn C; Krause F; Jepsen S	Journal of Clinical Periodontology
Photodynamic therapy as an adjunct to non-surgical periodontal treatment: a randomized controlled clinical trial	2008	Christodoulides N; Nikolidakis D; Chondros P; Becker J; Schwarz F; Rossler R; Sculean A	Journal of Periodontology
Clinical effectiveness of photodynamic therapy in the treatment of periodontitis	2009	Polansky R; Haas M; Heschl A; Wimmer G	Journal of Clinical Periodontology
Photodynamic therapy of persistent pockets in maintenance patients: a clinical study	2010	Ruhling A; Fanghanel J; Houshmand M; Kuhr A; Meisel P; Schwahn C; Kocher T	Clinical Oral Investigations
Photodynamic therapy in periodontal therapy: microbiological observations from a private practice	2010	Romanos GE; Brink B	General Dentistry
Clinical and microbiological effects of photodynamic therapy associated with nonsurgical periodontal treatment: a 6-month follow-up	2012	Theodoro LH; Silva SP; Pires JR; Soares GH; Pontes AE; Zuza EP; Spolidorio DM; de Toledo BE; Garcia VG	Lasers in Medical Science
Photodynamic therapy as an adjunctive to scaling and root planing in treatment of chronic periodontitis in smokers	2011	Al-Zahrani MS; Austah ON	Saudi Medical Journal
The adjunctive effect of photodynamic therapy for residual pockets in single-rooted teeth: a randomized controlled clinical trial	2013	Campos GN; Pimentel SP; Ribeiro FV; Casarin RC; Cirano FR; Saraceni CH; Casati MZ	Lasers in Medical Science
Photoactivated disinfection using light-emitting diode as an adjunct in the management of chronic periodontitis: a pilot double-blind split-	2013	Bassir SH; Moslemi N; Jamali R; Mashmouly S; Fekrazad R; Chiniforush N; Shamshiri AR; Nowzari H	Journal of Clinical Periodontology

 doi:10.20944/preprints202603.1440.v1

mouth randomized clinical trial			
Adjunctive effect of antimicrobial photodynamic therapy to nonsurgical periodontal treatment in smokers: a randomized clinical trial	2015	Queiroz AC; Suaid FA; de Andrade PF; Oliveira FS; Novaes AB Jr; Taba M Jr; Palioto DB; Grisi MF; Souza SL	Lasers in Medical Science
Long-term clinical effect of adjunctive antimicrobial photodynamic therapy in periodontal treatment: a randomized clinical trial	2015	Alwaeli HA; Al-Khateeb SN; Al-Sadi A	Lasers in Medical Science
Effects of photodynamic therapy on clinical and gingival crevicular fluid inflammatory biomarkers in chronic periodontitis: a split-mouth randomized clinical trial	2014	Pourabbas R; Kashefimehr A; Rahmanpour N; Babaloo Z; Kishen A; Tenenbaum HC; Azarpazhooh A	Journal of Periodontology
Efficacy of antimicrobial photodynamic therapy in the management of chronic periodontitis: a randomized controlled clinical trial	2014	Betsy J; Prasanth CS; Baiju KV; Prasanthila J; Subhash N	Journal of Clinical Periodontology
Clinical and microbiological effects of photodynamic therapy associated with nonsurgical treatment in aggressive periodontitis	2014	Chitsazi MT; Shirmohammadi A; Pourabbas R; Abolfazli N; Farhoudi I; Daghigh Azar B; Farhadi F	Journal of Dental Research, Dental Clinics, Dental Prospects
Effect of repeated adjunctive antimicrobial photodynamic therapy on subgingival periodontal pathogens in chronic periodontitis	2015	Petelin M; Perkic K; Seme K; Gaspirc B	Lasers in Medical Science
Antimicrobial photodynamic therapy as an adjunct to non-surgical treatment of aggressive periodontitis: a split-mouth randomized controlled trial	2015	Moreira AL; Novaes AB Jr; Grisi MF; Taba M Jr; Souza SL; Palioto DB; de Oliveira PG; Casati MZ; Casarin RC; Messora MR	Journal of Periodontology
Effect of a single session of antimicrobial photodynamic therapy	2015	Srikanth K; Chandra RV; Reddy AA; Reddy BH; Reddy C; Naveen A	Quintessence International

 doi:10.20944/preprints202603.1440.v1

using indocyanine green in the treatment of chronic periodontitis: a randomized controlled pilot trial			
Antimicrobial photodynamic therapy using diode laser activated indocyanine green as an adjunct in the treatment of chronic periodontitis: a randomized clinical trial	2016	Monzavi A; Chinipardaz Z; Mousavi M; Fekrazad R; Moslemi N; Azaripour A; Bagherpasand O; Chiniforush N	Photodiagnosis and Photodynamic Therapy
Efficacy of photodynamic therapy and lasers as an adjunct to scaling and root planing in aggressive periodontitis	2016	Annaji S; Sarkar I; Rajan P; Pai J; Malagi S; Bharmappa R; Kamath V	Journal of Clinical and Diagnostic Research
Microbiological efficacy of photodynamic therapy as an adjunct to non-surgical periodontal treatment: a clinical trial	2016	Talebi M; Taliee R; Mojahedi M; Meymandi M; Torshabi M	Journal of Lasers in Medical Sciences
Clinical effects of photodynamic and low-level laser therapies as an adjunct to scaling and root planing of chronic periodontitis	2016	Malgikar S; Reddy SH; Sagar SV; Satyanarayana D; Reddy GV; Josephin JJ	Indian Journal of Dental Research
Effect of photodynamic therapy adjunct to scaling and root planing in periodontitis patients: a randomized clinical trial	2016	Pulikkotil SJ; Toh CG; Mohandas K; Leong K	Australian Dental Journal
Effectiveness of adjunctive antimicrobial photodynamic therapy in reducing peri-implant inflammatory response in individuals vaping electronic cigarettes	2018	Al Rifaiy MQ; Qutub OA; Alasqah MN; Al-Sowygh ZH; Mokeem SA; Alrahlah A	Photodiagnosis and Photodynamic Therapy
Indocyanine green-mediated photothermal therapy in treatment of chronic periodontitis: a clinico-microbiological study	2018	Raut CP; Sethi KS; Kohale BR; Mamajiwala A; Warang A	Journal of Indian Society of Periodontology
Efficacy of adjunctive photodynamic therapy in the treatment of generalized aggressive	2019	Borekci T; Meseli SE; Noyan U; Kuru BE; Kuru L	Lasers in Surgery and Medicine

 doi:10.20944/preprints202603.1440.v1

periodontitis: a randomized controlled clinical trial			
Indocyanine green-based adjunctive antimicrobial photodynamic therapy for treating chronic periodontitis: a randomized clinical trial	2019	Hill G; Dehn C; Hinze AV; Frentzen M; Meister J	Photodiagnosis and Photodynamic Therapy
Clinical and microbiological effects of multiple applications of antibacterial photodynamic therapy in periodontal maintenance patients	2019	Grzech-Lesniak K; Gaspirc B; Sculean A	Photodiagnosis and Photodynamic Therapy
The effectiveness of photodynamic therapy as a complementary therapy to mechanical instrumentation on residual periodontal pocket clinical parameters	2020	Siva NTD; Silva DNA; Azevedo MLDS; Silva Junior FLD; Almeida ML; Longo JPF; Moraes M; Gurgel BCV; de Aquino Martins ARL	Photodiagnosis and Photodynamic Therapy
Antimicrobial photodynamic therapy using indocyanine green as a photosensitizer in treatment of chronic periodontitis	2019	Sethi KS; Raut CP	Indian Journal of Dental Research
Clinical effectiveness of indocyanine green mediated antimicrobial photodynamic therapy as an adjunct to scaling root planing	2020	Joshi K; Baiju CS; Khashu H; Bansal S	Photodiagnosis and Photodynamic Therapy
Effectiveness of adjunctive use of low-level laser therapy and photodynamic therapy after scaling and root planing in chronic periodontitis	2019	Gandhi KK; Pavaskar R; Cappetta EG; Drew HJ	International Journal of Periodontics & Restorative Dentistry
Antimicrobial photodynamic therapy with diode laser and methylene blue as an adjunct to scaling and root planning	2020	Derikvand N; Ghasemi SS; Safiaghdam H; Piriaei H; Chiniforush N	Photodiagnosis and Photodynamic Therapy
Efficacy of antimicrobial photodynamic therapy with chloro-aluminum	2020	de Araujo Silva DN; Silva NTD; Sena IAA; Azevedo MLDS; Junior	Photodiagnosis and Photodynamic Therapy

 doi:10.20944/preprints202603.1440.v1

phthalocyanine on periodontal clinical parameters		FLDS; Silva RCMD; Vasconcelos RC; de Moraes M; Longo JPF; de Araujo AA; de Aquino Martins ARL	
Impact of molar furcations on photodynamic therapy outcomes: a 6-month split-mouth randomized clinical trial	2020	Courval A; Harmouche L; Mathieu A; Petit C; Huck O; Severac F; Davideau JL	International Journal of Environmental Research and Public Health
Clinical and microbiological effects of adjunctive photodynamic diode laser therapy in chronic periodontitis	2020	Mallineni S; Nagarakanti S; Gunupati S; Bv RR; Shaik MV; Chava VK	Journal of Dental Research, Dental Clinics, Dental Prospects
Effects of adjunctive light-activated disinfection and probiotics on periodontal treatment	2021	Patyna M; Ehlers V; Bahlmann B; Kasaj A	Clinical Oral Investigations
Clinical and microbiological evaluation of local doxycycline and antimicrobial photodynamic therapy during supportive periodontal therapy	2021	Cosgarea R; Eick S; Batori-Andronescu I; Jepsen S; Arweiler NB; Rossler R; Conrad T; Ramseier CA; Sculean A	Antibiotics
Clinico-microbiological efficacy of indocyanine green as a novel photosensitizer for photodynamic therapy	2021	Karmakar S; Prakash S; Jagadeson M; Namachivayam A; Das D; Sarkar S	Journal of Pharmacy & Bioallied Sciences
Effectiveness of photodynamic therapy as an adjunct to periodontal scaling for treating periodontitis in geriatric patients	2022	Elsadek MF; Farahat MF	European Review for Medical and Pharmacological Sciences
Chloro-aluminum phthalocyanine-mediated photodynamic therapy in stage-II chronic periodontitis among smokers	2022	Al-Kheraif AA; Alshahrani OA; Al-Shehri AM; Khan AA	Photodermatology, Photoimmunology & Photomedicine
Antimicrobial photodynamic therapy using chloro-aluminum phthalocyanine for treating advanced stage-III periodontitis	2022	Al-Kheraif AA; Alshahrani OA; Al-Shehri AM; Khan AA	Photodermatology, Photoimmunology & Photomedicine

 doi:10.20944/preprints202603.1440.v1

The efficiency of photodynamic therapy in bacterial decontamination of periodontal pockets	2022	Munteanu IR; Luca RE; Mateas M; Darawsha LD; Boia S; Boia ER; Todea CD	Diagnostics
Photodynamic therapy as an adjunctive treatment for grade C periodontitis in molar teeth	2023	Coelho TDRC; Pinto Filho JM; Ribeiro Caponi LSFE; Soares JDM; Dos Santos JN; Cury PR	Quintessence International
Photodynamic therapy with tolonium chloride and a diode laser in non-surgical management of periodontitis	2023	El Mobadder M; Nammour S; Grzech-Lesniak K	Journal of Clinical Medicine
Photodynamic therapy as adjunctive treatment of single-rooted teeth in patients with grade C periodontitis	2023	Rodrigues RD; Araujo NS; Filho JMP; Vieira CLZ; Ribeiro DA; Dos Santos JN; Cury PR	Photodiagnosis and Photodynamic Therapy
Clinical comparison of using a single episode of photodynamic therapy vs ultrasonic scaling in periodontitis	2014	Amini S; Shirani S; Tahmourespour A	Research in Dental Sciences
Evaluation of nonsurgical periodontal treatment with antibiotics and photodynamic therapy in aggressive periodontitis	2018	Bechara Andere NM; Dos Santos NCC; Araujo CF; Mathias IF; Rossato A; de Marco AC; Santamaria MP	Photodiagnosis and Photodynamic Therapy
Comparison between scaling-root-planing and SRP/photodynamic therapy: six-month study	2012	Berakdar M; Callaway A; Fakhr Eddin M; Röss A; Willershausen B	Head & Face Medicine
Effects of laser and indocyanine-green mediated antimicrobial photodynamic therapy on red-complex bacteria	2023	Arya A; Srirangarajan S; Rao R; Prabhu S; Deepika O	Journal of the International Academy of Periodontology

Appendix B

B.1. Formal Representation of Criteria

Each eligibility condition was modeled as a Boolean predicate P_i, where:

$$P_i(\text{abstract}) = \begin{cases} 1 & \text{if the abstract satisfies criterion } i \\ 0 & \text{otherwise.} \end{cases}$$

The global eligibility decision for the abstract-level screen followed a conjunctive structure:

$$\text{Eligible} = P_1 \wedge P_2 \wedge \ldots$$

For the present review, four predicates formed this AND-block:

P_{RCT}: study is a randomized controlled trial

 doi:10.20944/preprints202603.1440.v1

P_{Adult}: participants are adults
P_{FollowUp}: follow-up ≥ 1 month
$P_{\text{Diagnosis}}$: diagnosis involves periodontitis or peri-implant disease

Because these properties are obligatory and not interchangeable, they were evaluated sequentially: an abstract failing any P_i did not advance to subsequent rounds.

A separate block of criteria dealt with acceptable intervention/comparison patterns relevant to antimicrobial photodynamic therapy (aPDT). These were not mutually required but functioned as alternatives. Thus, four specific comparison forms were grouped into a single composite OR-predicate:

$$P_{\text{aPDT}} = P_a \vee P_b \vee P_c \vee P_d,$$

where:

P_a: subgingival aPDT + scaling/instrumentation vs SI or sham
P_b: submucosal aPDT + scaling/instrumentation vs SI or sham
P_c: aPDT vs control during active/supportive therapy
P_d: aPDT vs control for peri-implant mucositis

This block was treated as one unified decision step in the screening pipeline:

$$\text{Eligible} = (P_{\text{RCT}} \wedge P_{\text{Adult}} \wedge P_{\text{FollowUp}} \wedge P_{\text{Diagnosis}}) \wedge (P_{\text{aPDT}})$$

This structure ensures that abstracts are not incorrectly excluded for failing one particular comparison pattern when they may satisfy another equivalent one—a problem that would arise if these alternatives were applied as sequential filters.

B.2. Linguistic Variants of Each Logical Predicate

For each predicate P_i, five linguistically distinct formulations were generated. These variants differed only in surface polarity and negation scope, while preserving logical intent.

1. **Affirmative Inclusion (AI)**
   - expresses the predicate directly as a positive inclusion requirement.
   - Example: "INCLUDE IF the follow-up period is at least one month."
   - Logical mapping: P_i
2. **Antonymic Exclusion (AE)**
   - expresses a clinically meaningful exclusion rule without explicit linguistic negation.
   - Example: "EXCLUDE IF the follow-up period is less than one month."
   - Logical mapping: $\neg P_i$ expressed via a non-negated antonym.
3. **Predicate Negation (PN)**
   - explicitly negates the predicate.
   - Example: "EXCLUDE IF the follow-up period is not at least one month."
   - Logical mapping: direct linguistic negation $\neg P_i$.
4. **Verb Negation (VN)**
   - negation is applied to the action ("do not include") while the condition remains in its positive or complementary form.
   - Logical mapping remains equivalent to exclusion but is linguistically distinct:

$$\text{DO NOT INCLUDE IF } Q_i$$

where Q_i denotes the positive or antonymically expressed condition associated with predicate P_i.

5. **Double Negation (DN)**
   - both the action and the predicate are negated: "DO NOT INCLUDE IF the follow-up period is not at least one month."
   - Logical mapping: still a form of $\neg P_i$, but syntactically more complex.

 doi:10.20944/preprints202603.1440.v1

These variants allow direct measurement of whether LLM performance changes solely due to the surface structure of negation, even when the logical content is nominally equivalent.

B.3. Special Considerations for the Composite Comparison Criterion

The composite aPDT criterion required special treatment because its logical structure differs from the atomic predicates. As defined in B.1, the composite aPDT criterion is the disjunction:

$$P_{\text{aPDT}} = P_a \vee P_b \vee P_c \vee P_d,$$

where each P_k corresponds to one acceptable aPDT comparison pattern. Unlike the mandatory AND-block criteria, this predicate specifies a disjunction over a finite set of admissible intervention structures.

A key consequence of this disjunctive structure is that its logical complement cannot be expressed as a concise positive antonym. The negation of an OR-predicate corresponds to the absence of all listed alternatives:

$$\neg P_{\text{aPDT}} = \neg P_a \wedge \neg P_b \wedge \neg P_c \wedge \neg P_d.$$

In an open biomedical domain, this complement set is effectively unbounded: it includes any intervention not conforming to the defined aPDT patterns. Expressing such a complement without explicit negation would require either enumerating all non-aPDT therapies or introducing vague domain-general exclusions, both of which undermine logical symmetry.

For this reason, the Antonymic Exclusion (AE) variant cannot constitute a strict logical complement of the Affirmative Inclusion (AI) formulation for the composite criterion. Instead, AE operationalizes exclusion by specifying a complementary intervention class (e.g., studies evaluating non-photodynamic therapies), which approximates but does not formally constitute the logical complement of the AI formulation. In contrast, the Predicate Negation (PN) and Double Negation (DN) variants more closely approximate the formal complement $\neg P_{\text{aPDT}}$, as they explicitly encode the absence of any of the admissible comparison structures.

The Verb Negation (VN) variant introduces an additional layer of asymmetry. Because negation is applied to the decision action rather than to the predicate itself, the formulation necessarily takes the form:

"*Do not include if the study evaluates therapies other than those in* $\{P_a, P_b, P_c, P_d\}$."

This construction does not provide a clean logical inversion of the affirmative rule. Instead, it shifts the scope of negation to the inclusion decision while relying on a domain-delimited description of alternatives. As a result, VN is semantically distinct from both AE and PN/DN, despite all functioning operationally as exclusion rules.

These asymmetries are not methodological artifacts but structural consequences of applying polarity transformations to a disjunctive predicate in an open intervention space. Accordingly, AE, PN, VN, and DN are not expected to behave identically for the composite criterion. Their differences allow examination of how LLMs respond to variations in negation scope—antonymic complement, predicate negation, action-level negation, and double negation—each imposing distinct interpretive demands.

 doi:10.20944/preprints202603.1440.v1

Appendix C

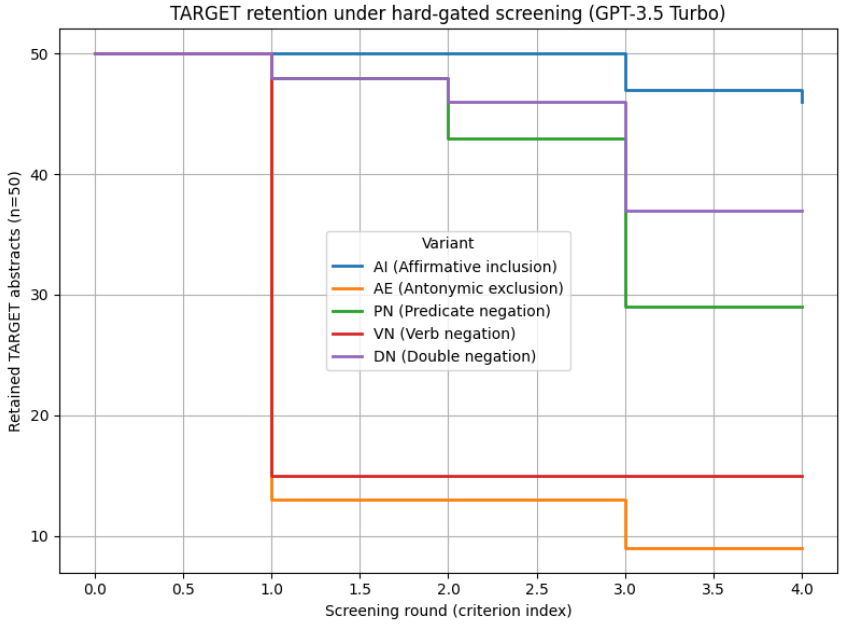
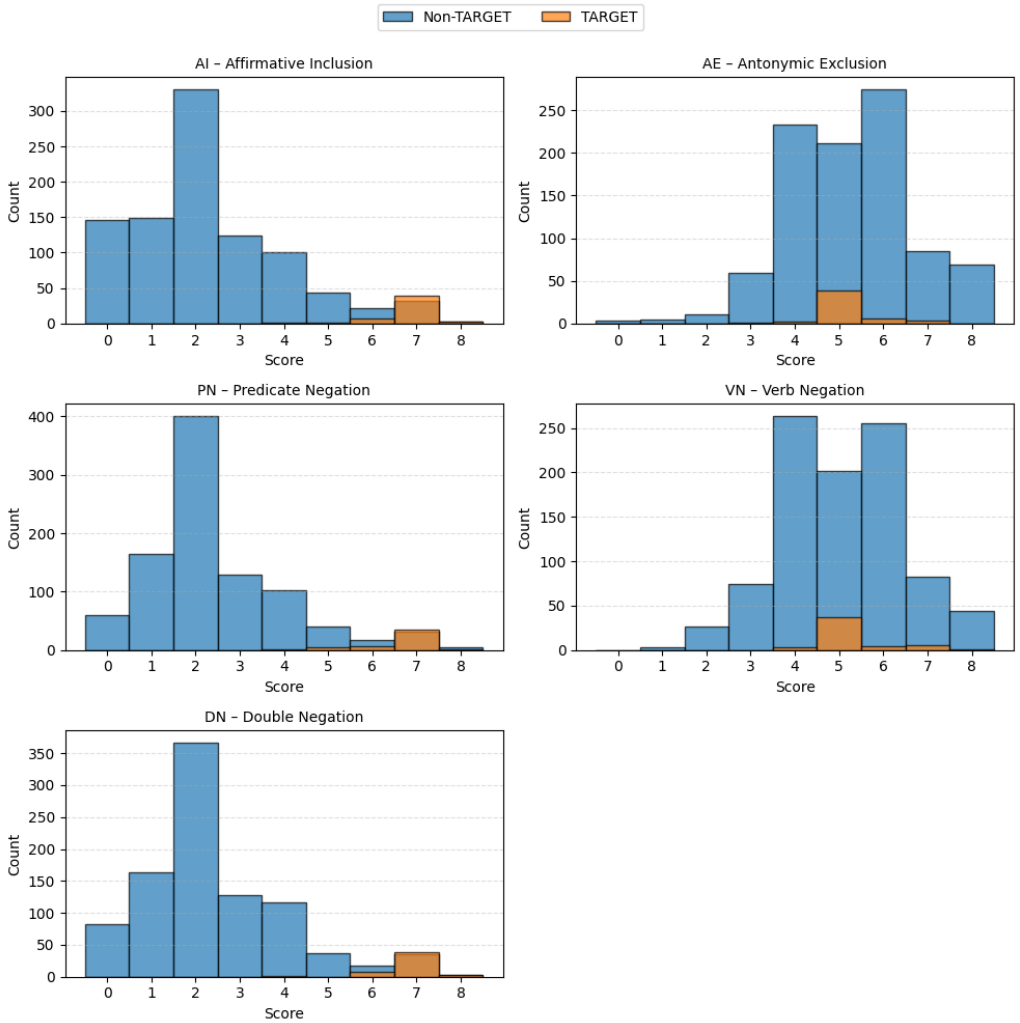


Figure C1. TARGET retention across successive hard-gated screening rounds using GPT-3.5 Turbo. The same polarity-dependent divergence and Round-3 concentration of losses observed for GPT-5.1 is replicated across model generations.



 doi:10.20944/preprints202603.1440.v1

Figure C2. Distribution of cumulative screening scores under the scoring-based architecture. Histograms show the cumulative score assigned to TARGET and non-TARGET abstracts across the four evaluated criteria for the five linguistic variants (AI – Affirmative Inclusion, AE – Antonymic Exclusion, PN – Predicate Negation, VN – Verb Negation, DN – Double Negation). Higher scores indicate stronger aggregate evidence of eligibility. TARGET abstracts tend to accumulate higher scores than non-TARGET abstracts, although the degree of separation varies across formulations. Variants with clearer score separation (AI, PN, DN) achieve higher precision at comparable recall levels, whereas AE and VN exhibit substantial score overlap and therefore require lower thresholds to maintain high recall.

References

1. Anderson, N.K.; Jayaratne, Y.S.N. Methodological Challenges When Performing a Systematic Review. *The European Journal of Orthodontics* **2015**, *37*, 248–250, doi:10.1093/ejo/cjv022.
2. Gupta, S.; Rajiah, P.; Middlebrooks, E.H.; Baruah, D.; Carter, B.W.; Burton, K.R.; Chatterjee, A.R.; Miller, M.M. Systematic Review of the Literature: Best Practices. *Acad. Radiol.* **2018**, *25*, 1481–1490, doi:10.1016/j.acra.2018.04.025.
3. Westgate, M.J.; Lindenmayer, D.B. The Difficulties of Systematic Reviews. *Conservation Biology* **2017**, *31*, 1002–1007.
4. Ofori-Boateng, R.; Aceves-Martins, M.; Wiratunga, N.; Moreno-Garcia, C.F. Towards the Automation of Systematic Reviews Using Natural Language Processing, Machine Learning, and Deep Learning: A Comprehensive Review. *Artif. Intell. Rev.* **2024**, *57*, 200.
5. Ouzzani, M.; Hammady, H.; Fedorowicz, Z.; Elmagarmid, A. Rayyan—a Web and Mobile App for Systematic Reviews. *Syst. Rev.* **2016**, *5*, 1–10.
6. Tsou, A.Y.; Treadwell, J.R.; Erinoff, E.; Schoelles, K. Machine Learning for Screening Prioritization in Systematic Reviews: Comparative Performance of Abstrackr and EPPI-Reviewer. *Syst. Rev.* **2020**, *9*, 73.
7. Hamel, C.; Kelly, S.E.; Thavorn, K.; Rice, D.B.; Wells, G.A.; Hutton, B. An Evaluation of DistillerSR's Machine Learning-Based Prioritization Tool for Title/Abstract Screening–Impact on Reviewer-Relevant Outcomes. *BMC Med. Res. Methodol.* **2020**, *20*, 256.
8. de la Torre-López, J.; Ramírez, A.; Romero, J.R. Artificial Intelligence to Automate the Systematic Review of Scientific Literature. *Computing* **2023**, *105*, 2171–2194, doi:10.1007/s00607-023-01181-x.
9. Sindhu, B.; Prathamesh, R.P.; Sameera, M.B.; KumaraSwamy, S. The Evolution of Large Language Model: Models, Applications and Challenges. In Proceedings of the 2024 International Conference on Current Trends in Advanced Computing (ICCTAC); IEEE, 2024; pp. 1–8.
10. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* **2023**.
11. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large Language Models in Medicine. *Nat. Med.* **2023**, *29*, 1930–1940.
12. Khraisha, Q.; Put, S.; Kappenberg, J.; Warraitch, A.; Hadfield, K. Can Large Language Models Replace Humans in the Systematic Review Process? Evaluating GPT-4's Efficacy in Screening and Extracting Data from Peer-Reviewed and Grey Literature in Multiple Languages. *arXiv preprint arXiv:2310.17526* **2023**.
13. Семеріков, С.О.; Мінтій, І.С. Automating Literature Screening with Large Language Models. **2024**.
14. Scherbakov, D.; Hubig, N.; Jansari, V.; Bakumenko, A.; Lenert, L.A. The Emergence of Large Language Models as Tools in Literature Reviews: A Large Language Model-Assisted Systematic Review. *Journal of the American Medical Informatics Association* **2025**, *32*, 1071–1086, doi:10.1093/jamia/ocaf063.
15. Colangelo, M.T.; Guizzardi, S.; Meleti, M.; Calciolari, E.; Galli, C. Performance Comparison of Large Language Models for Efficient Literature Screening. *BioMedInformatics* **2025**, *5*, 25, doi:10.3390/biomedinformatics5020025.
16. Rokach, L.; Romano, R.; Maimon, O. Negation Recognition in Medical Narrative Reports. *Inf. Retr. Boston.* **2008**, *11*, 499–538, doi:10.1007/s10791-008-9061-0.
17. Dickersin, K.; Scherer, R.; Lefebvre, C. Systematic Reviews: Identifying Relevant Studies for Systematic Reviews. *Bmj* **1994**, *309*, 1286–1291.

 doi:10.20944/preprints202603.1440.v1

18. Malaviya, C.; Chang, J.C.; Roth, D.; Iyyer, M.; Yatskar, M.; Lo, K. Contextualized Evaluations: Judging Language Model Responses to Underspecified Queries. *Trans. Assoc. Comput. Linguist.* **2025**, *13*, 878–900.
19. Ivanov, T.; Penchev, V. AI Benchmarks and Datasets for LLM Evaluation. *arXiv preprint arXiv:2412.01020* **2024**.
20. Jervøe-Storm, P.M.; Bunke, J.; Worthington, H. V.; Needleman, I.; Cosgarea, R.; MacDonald, L.; Walsh, T.; Lewis, S.R.; Jepsen, S. Adjunctive Antimicrobial Photodynamic Therapy for Treating Periodontal and Peri-Implant Diseases. *Cochrane Database of Systematic Reviews* 2024, *2024*.
21. McCrae, N.; Purssell, E. Eligibility Criteria in Systematic Reviews Published in Prominent Medical Journals: A Methodological Review. *J. Eval. Clin. Pract.* **2015**, *21*, 1052–1058.
22. Li, J.; Kabouji, J.; Bouhadoun, S.; Tanveer, S.; Filion, K.B.; Gore, G.; Josephson, C.B.; Kwon, C.-S.; Jette, N.; Bauer, P.R. Sensitivity and Specificity of Alternative Screening Methods for Systematic Reviews Using Text Mining Tools. *J. Clin. Epidemiol.* **2023**, *162*, 72–80.
23. Bisong, E. Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Bisong, E., Ed.; Apress: Berkeley, CA, 2019; pp. 59–64 ISBN 978-1-4842-4470-8.
24. Mckinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the Proceedings of the 9th Python in Science Conference; van der Walt, S., Millman, J., Eds.; 2010; pp. 51–56.
25. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362, doi:10.1038/s41586-020-2649-2.
26. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, doi:10.1109/MCSE.2007.55.
27. Huotala, A.; Kuutila, M.; Ralph, P.; Mäntylä, M. The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews. In Proceedings of the Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering; 2024; pp. 262–271.
28. Oami, T.; Okada, Y.; Nakada, T.A. Performance of a Large Language Model in Screening Citations. *JAMA Netw. Open* **2024**, *7*, doi:10.1001/jamanetworkopen.2024.20496.
29. Sanghera, R.; Thirunavukarasu, A.J.; El Khoury, M.; O'Logbon, J.; Chen, Y.; Watt, A.; Mahmood, M.; Butt, H.; Nishimura, G.; Soltan, A.A.S. High-Performance Automated Abstract Screening with Large Language Model Ensembles. *Journal of the American Medical Informatics Association* **2025**, *32*, 893–904.
30. Li, M.; Sun, J.; Tan, X. Evaluating the Effectiveness of Large Language Models in Abstract Screening: A Comparative Analysis. *Syst. Rev.* **2024**, *13*, doi:10.1186/s13643-024-02609-x.
31. Han, B.; Mathrani, A.; Susnjak, T. Evaluating Prompting Strategies and Large Language Models in Systematic Literature Review Screening: Relevance and Task-Stage Classification. *arXiv preprint arXiv:2510.16091* **2025**.
32. Gartlehner, G.; Affengruber, L.; Titscher, V.; Noel-Storr, A.; Dooley, G.; Ballarini, N.; König, F. Single-Reviewer Abstract Screening Missed 13 Percent of Relevant Studies: A Crowd-Based, Randomized Controlled Trial. *J. Clin. Epidemiol.* **2020**, *121*, 20–28.
33. Shankar, S.; Zamfirescu-Pereira, J.D.; Hartmann, B.; Parameswaran, A.; Arawjo, I. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. In Proceedings of the Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology; ACM: New York, NY, USA, October 13 2024; pp. 1–14.