# MetaBench: Benchmarking Two-Stage Reranking and Top-N Screening for Meta-Analyses

**Yining Huang**[1,2*], **Keke Tang**[1,3,4†], **Meilian Chen**[1‡]
[1]Meta Emergence Laboratory
[2]School of Politics and Public Administration, South China Normal University
[3]University of Chinese Academy of Sciences
[4]Shenyang institute of computing technology, Chinese academy of sciences
**Correspondence:** huangyining1987@gmail.com

## Abstract

Meta-analyses are central to evidence-based medicine, but their production remains slow because reviewers must search PubMed and screen large candidate sets under PICOS eligibility constraints. We introduce **MetaBench**, a large-scale benchmark for **early-stage PubMed title/abstract citation screening** in meta-analysis study selection, formulated as a **two-stage reranking and Top-$N$ screening** problem with **effort-aware** evaluation. MetaBench contains **7,801** PubMed meta-analyses, **120,930** PMID-linked included studies, and reproducible candidate pools with **4.2M** unique title/abstract records constructed from replayed PubMed strategies or expert-validated training-only reconstructions. To improve evaluation fidelity, we introduce a **coverage-based tiering** protocol and reserve **Tier 1** (**1,297** topics) as a high-integrity test set. MetaBench evaluates candidate reranking with Recall/nDCG and Effort@R/NormEffort@R, and budgeted Top-$N$ screening with sensitivity, specificity, precision, and F-scores. The main benchmark intentionally focuses on scalable title/abstract screening rather than full-text eligibility adjudication; PMCID metadata supports an auxiliary full-text track on an OA-constrained subset. Experiments show that MetaBench supervision improves dense reranking, reduces screening effort, and strengthens LLM-based inclusion decisions. MetaBench v1.0, the evaluation code, replay metadata, and released PMID-level candidate pools are available at `https://github.com/EmergencerOnEarth/MetaBench`.

## 1 Introduction

Meta-analyses and systematic reviews are central to evidence-based medicine: they inform clinical guidelines and synthesize evidence from primary studies. Reporting standards such as PRISMA reflect how structured and important these workflows are in practice (Page et al., 2021).

Yet producing high-quality meta-analyses remains slow and labor-intensive. Authors must design complex database queries, screen large retrieval outputs, and apply strict PICOS-based (Population, Intervention/Exposure, Comparator, Outcome and Study Design) eligibility criteria. Real-world registry analyses show substantial human workload, motivating continued efforts to accelerate evidence synthesis (Borah et al., 2017).

Accordingly, many studies have explored automation for literature search and citation screening, from classical IR and supervised ranking to LLM-based assistants. Early work on semi-automated screening highlighted both the promise and evaluation challenges of reducing review burden while maintaining high recall (Wallace et al., 2010). Shared tasks in high-recall / technology-assisted review settings further emphasize that retrieval must be optimized under explicit review-effort constraints (Kalphov et al., 2017).

However, existing benchmarks still have key limitations for meta-analysis-level citation screening. First, many treat screening as topical relevance classification rather than modeling the multi-dimensional PICOS constraints that define real inclusion decisions, where topically related studies may still be ineligible because of subtle mismatches in population, outcome, or study design. Second, existing large-scale resources often lack executable query replay, consistent candidate-pool construction, or standardized end-to-end protocols for PubMed-style screening (Kusa et al., 2023; Chan et al., 2025). Third, real screening is naturally two-stage (retrieve/rerank $\rightarrow$ inclusion decision), but many benchmarks evaluate stages in isolation and do not standardize effort-aware evaluation. Fourth, for meta-analysis-based supervision, "not included"

*huangyining1987@gmail.com
†tkk2012@gmail.com
‡523062863@qq.com



*Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 27010–27050
October 24-29, 2026 ©2026 Association for Computational Linguistics

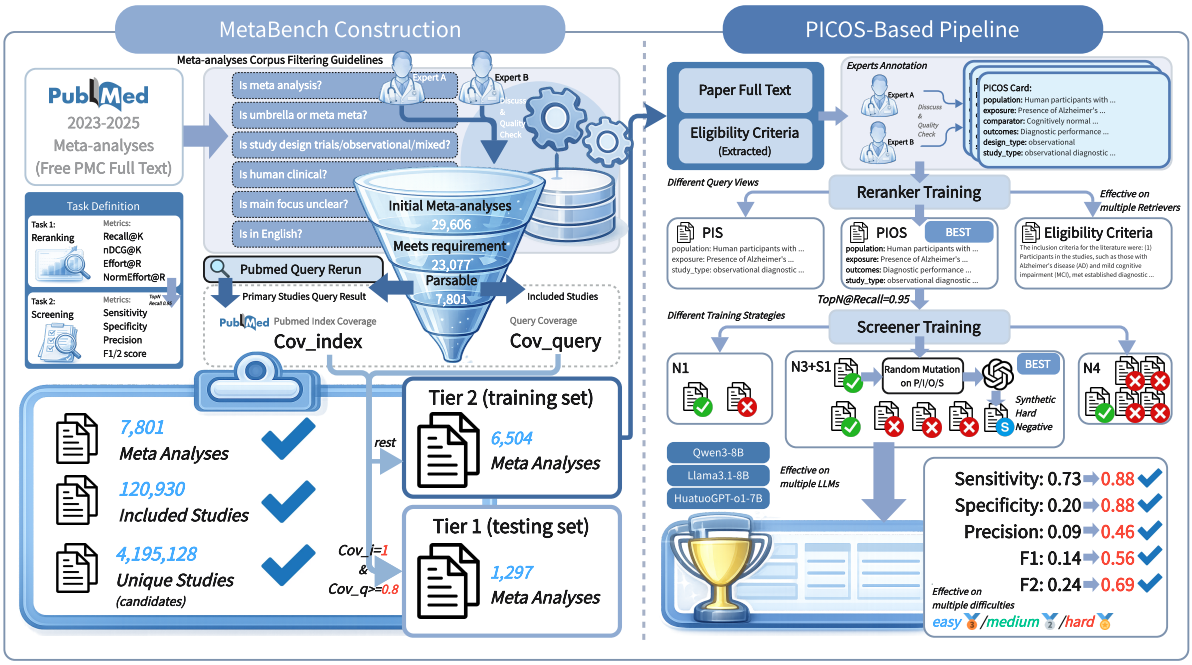



Figure 1: **Overview of MetaBench and the two-stage pipeline.** Left: we construct coverage-controlled tiered candidate pools by replaying PubMed queries and linking included studies for published meta-analyses. Right: we annotate PICOS cards and train a reranker (Task 1) and an LLM screener (Task 2) under a fixed Top-$N$ protocol, evaluated with effort-aware retrieval and sensitivity/precision-oriented screening metrics.

does not always mean "truly ineligible," so without coverage controls evaluation can confound model error with incompleteness of the human reference set.

To address these gaps, we introduce MetaBench, a large-scale benchmark for early-stage PubMed title/abstract citation screening in published meta-analyses, formalized as a two-stage retrieval–screening pipeline. MetaBench provides topic text, standardized query views including structured eligibility/PICOS, PMID-linked included studies, and author-reported or expert-validated PubMed strategies with replayed candidate pools. It standardizes (i) Task 1 reranking with effort-aware metrics under high-recall targets, and (ii) Task 2 inclusion decisions under a fixed Top-$N$/effort budget, while introducing coverage-based tiering to separate high-integrity evaluation topics from broader training-only topics.

MetaBench is therefore intended to standardize the scalable title/abstract retrieval and screening stages of evidence synthesis, rather than to replace full-text review, final eligibility adjudication, data extraction, or risk-of-bias assessment. To preserve reproducibility and avoid OA-only bias, the main benchmark is defined over title/abstract records, while PMCID metadata enables an auxiliary full-text track on an OA-constrained subset.

Empirically, training dense retrievers on MetaBench substantially improves ranking quality and reduces effort, while fine-tuning LLM screeners improves budgeted inclusion decisions. We further find that negative ratio controls the sensitivity–precision trade-off, and that PICOS-targeted synthetic hard negatives improve precision and robustness on reranker-hard topics. With this submission, we include anonymized supplementary code and dataset samples containing representative examples, schemas, configurations, and scripts for candidate-pool construction, tiering, training, and evaluation.

## 2 Related Work

Benchmark datasets for review screening include CSMED (Kusa et al., 2023), large-scale screening corpora (Chan et al., 2025), and high-recall Technology-Assisted Review setups such as CLEF eHealth TAR (Azzopardi et al., 2017). These resources have advanced citation-screening research, but they generally do not jointly provide executable PubMed query replay, PMID-linked published inclusions, coverage-controlled tiering, structured PICOS fields, and a standardized reranking-to-Top-$N$ screening protocol. Table 1 summarizes the comparison. Although MetaBench and Chan et al. (2025) cover similar numbers of topics, their candidate counts are not directly comparable: MetaBench constructs raw candidate pools by

PubMed replay, whereas prior large-scale resources mainly rely on curated screening artifacts or logged review records.

Adjacent resources link evidence-synthesis artifacts to biomedical corpora, such as trial-centric collections (Wang et al., 2025), or provide interactive screening systems with retrieval and feedback (Mao et al., 2025). These efforts are complementary: MetaBench instead defines a reproducible benchmark where each topic has replayable PubMed candidates, PMID-linked inclusions, structured eligibility representations, and coverage-based train/test tiering.

Prior work has studied query formulation, neural reranking, high-recall screening, and LLM-based eligibility prediction (Pourreza and Ensan, 2022; Wang et al., 2023; Mao et al., 2024; Wang et al., 2024b; thi Tran et al., 2024; Cao et al., 2025; Cai et al., 2025; Oami et al., 2024; Sanghera et al., 2024). We do not claim that two-stage screening, effort-aware evaluation, or structured eligibility representations are individually new. MetaBench's contribution is their integration into a PubMed-replayable, coverage-controlled, PICOS-structured benchmark with standardized Top-$N$/effort evaluation, validated query reconstruction, noise audits, and reusable artifacts.

## 3 Task Definition

We model early-stage PubMed title/abstract screening for meta-analysis study selection as a two-stage pipeline on topic-specific candidate pools. Task 1 is full-pool, high-recall prioritization; Task 2 is budgeted include/exclude screening over the near-miss region surfaced by the reranker. The two stages therefore evaluate different operating regimes rather than different information sources. This design reflects practical screening workflows: applying an LLM screener to every query-retrieved candidate is costly and, under low inclusion prevalence, can produce many false positives. A fixed Top-$N$ budget also makes evaluation comparable across topics. Full text is incomplete and non-universal at PubMed scale, and OA-only evaluation can introduce systematic bias; thus, MetaBench defines both main tasks over **title/abstract** for reproducibility and legal redistributability. MetaBench provides multiple topic/query views (e.g., title+abstract, criteria text, or structured PICOS variants). When a system conditions on explicit eligibility-oriented views such as PICOS cards, this is a property of the chosen input representation rather than a separately enforced task constraint.

### 3.1 Preliminaries and Notation

Let $M$ be the set of meta-analyses (topics). Each $m \in M$ is represented by a query view $q_m$ (e.g., title/abstract, structured PICOS). We construct a candidate pool $D_m$ by re-running the author-reported PubMed query in a matched time window (deduplicated), where each candidate $d \in D_m$ is represented by title/abstract. Let $I_m \subseteq D_m$ be the PMID-mapped included studies from the published meta-analysis, treated as reference positives:

$$y_m(d) = \begin{cases} 1, & d \in I_m, \\ 0, & d \in D_m \setminus I_m. \end{cases} \quad (1)$$

This uses the published inclusion list as a *pragmatic reference standard* for large-scale benchmarking, rather than a fully adjudicated eligibility gold standard. Retrieved-but-not-included records are therefore treated as *pragmatic negatives*: they indicate disagreement with the published inclusion list, but some may still be abstract-level PICOS-compatible or excluded for reasons not observable from title/abstract alone. We quantify this residual ambiguity through manual negative-noise audits and noise-aware evaluation, summarized in Section 4.4. A scoring function $s(q_m, d) \in \mathbb{R}$ induces a ranking $\pi_m$ over $D_m$.

**Document views.** Unless otherwise specified, each candidate $d \in D_m$ is represented by title/abstract. When available, we also store the candidate's PMCID to enable optional full-text retrieval from PMC OA for the auxiliary full-text track (Appendix J).

### 3.2 Task 1: Candidate Reranking

**Goal.** Given a chosen topic/query view $q_m$, the goal is to rank $D_m$ so that $I_m$ appears early, enabling high recall under limited inspection. Operationally, this stage is designed to concentrate likely inclusions near the top of the list and reduce downstream screening burden. Besides $\mathrm{Recall@}K$ and $\mathrm{nDCG@}K$, we quantify screening workload via effort-to-recall:

$$E_m(R) = \min\{k \mid \mathrm{Recall@}k(m) \geq R\}. \quad (2)$$

We further define the normalized version

$$\mathrm{NormEffort}_m(R) = \frac{E_m(R)}{|D_m|}. \quad (3)$$

$E_m(R)$ (and its normalized form) captures how many records must be reviewed to reach target

| Benchmark/Dataset | Scale | | Query | Incl. | Coverage | PICO(S) | Top-$N$ |
|---|---|---|---|---|---|---|---|
| | #Reviews/ Topics | #Docs (candidates) | Replay (exec.) | IDs (PMID) | Tiering /coverage | Structured fields | 2-stage Top-$N$/effort |
| **MetaBench** (ours) | 7,801 | 4.2M (unique) | Yes | Yes | Yes | Yes | Yes |
| CSMeD (2023) | 325 | 730,236 | Partial | Partial | No | No | No |
| Chan et al. (2025) | 8,608 | 544,157 | No | Partial | No | No | No |
| CLEF (2017) | 50 | 149,405 (dev) 117,562 (test) | Yes | Yes | No | No | Partial |

Table 1: **Positioning of MetaBench against closely related screening/retrieval benchmarks. Replay (exec.)** indicates whether executable search strategies are provided and used to construct candidate pools by query replay. **Coverage/tiering** indicates whether explicit coverage measures (e.g., index/query coverage) are used to define tiers/splits. **Top-$N$ 2-stage** indicates whether the benchmark standardizes a two-stage rerank→Top-$N$ screening setting and reports effort-aware metrics under an explicit budget. "Partial" denotes that the capability exists only for some subsets or is not consistently standardized across the dataset.

recall $R$, making ranking quality directly interpretable in terms of screening effort. We note that $E_m(R)$ is an offline oracle evaluation metric rather than a deployment-time stopping rule, since real reviewers do not know recall during screening. We use it to compare ranking quality against reference labels, and complement it with evaluation by instantiating Task 2 with $N$ derived from the topic-wise distribution of $E_m(0.95)$.

### 3.3 Task 2: Top-$N$ Screening / Inclusion Decision

**Goal.** Given the Top-$N$ candidates produced by Task 1, predict agreement with the published inclusion list under PICOS constraints using title/abstract evidence. This stage targets the harder near-miss cases after obvious non-relevant records have largely been filtered by reranking. We define the screening pool as the top-$N$ candidates from Task 1:

$$C_m = \mathrm{Top}N(\pi_m), \tag{4}$$

with $N$ chosen from Task 1 effort statistics (e.g., $N = \mathrm{p95}\{E_m(R)\}$, p95 is the 95th percentile across topics) to standardize budgets across topics. The screener predicts

$$\hat{y}_m(d) \in \{0, 1\}, \quad d \in C_m, \tag{5}$$

and is evaluated with macro-averaged Sensitivity (Recall), Specificity, Precision, and F-scores (F1/F2) under the fixed top-$N$ budget. Precision should therefore be interpreted as agreement with published inclusion decisions, not as absolute eligibility precision. Published inclusion may depend on full-text details, multi-database search results, or non-PICOS constraints that are not fully observable from title/abstract alone; thus Task 2 evaluates pragmatic title/abstract screening rather than final full-text adjudication.

## 4 MetaBench Dataset

### 4.1 Data Overview

We retrieved PubMed records (2023–2025) marked as *Free full text* and indicative of meta-analytic evidence, yielding 29,606 candidates. Two licensed medical annotators screened records with a shared 7-field eligibility guideline, 10% overlap agreement checking, rule-based adjudication, and a 5% post-audit. Applying eligibility filters retained 23,077 meta-analyses. Annotation details, IAA, and adjudication protocols are in Appendix A and A.3; supervision quality is further examined through negative-noise audits and noise-aware evaluation.

We processed PMC Open Access full texts to extract included-study lists and references, excluding non-machine-actionable cases. This produced 7,801 meta-analyses with a PubMed search component: 2,116 with executable queries in the paper, 3,475 in supplements, and 2,210 with narrative-only strategies. Narrative-only strategies were reconstructed with GPT-5.1 and medically audited. On the 10% stratified audit sample, 97.31% of the reconstructed queries passed without edits; all remaining cases were corrected or regenerated and re-verified for PubMed executability. Reconstructed-query topics are used only in Tier 2 training/analysis, while Tier 1 final evaluation uses author-provided executable PubMed strategies with high replay coverage.

Across the 7,801 meta-analyses, we link 120,930 included studies by PMID. For each topic, replayed PubMed candidates are unioned with included studies and deduplicated by PMID, yielding 4,195,128 unique title/abstract records in the global study library. We also retain source-review metadata such as journal, year, IF, and citation count for optional subgroup analysis.

For each included study and retrieved candidate,

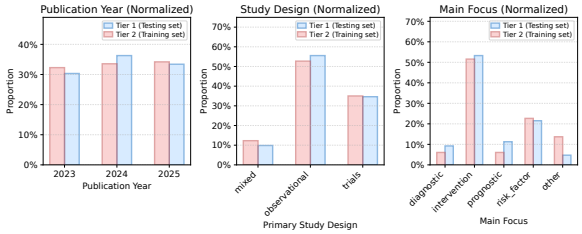



Figure 2: Dataset Statistics Comparison of Tier 1/2

we additionally record its **PMCID** (when available). This does not change MetaBench's PubMed-universe definition (PMID-based candidate generation), but enables subset-based analyses using publicly accessible PMC OA full texts.

### 4.2 Data Sources and Universe Definition

MetaBench is defined on a PubMed-only universe: meta-analyses and candidate studies are represented by PMIDs and constructed from executable PubMed queries, enabling standardized metadata and reproducible candidate generation. Full-text processing is restricted to PMC Open Access (PMCID-available) articles for legal accessibility and redistribution safety. Non-PubMed sources (e.g., Google Scholar and other databases) are excluded; thus both tasks operate on PubMed-indexed records and query-derived candidate pools unioned with PMID-linked inclusions.

### 4.3 Coverage-based Tiering

We tier topics using two coverage measures. *PubMed index coverage* is the fraction of included studies that are PMID-mapped and thus retrievable in PubMed; *Query coverage* is the fraction of included PMIDs recovered by re-running the paper-reported PubMed query. Together, these measures improve machine-actionability and enable controlled evaluation by grounding candidates and labels in a single verifiable corpus.

Topics with *PubMed index coverage* $= 1$ and *Query coverage* $\geq 0.8$ form Tier 1 (see Appendix F); all others form Tier 2. Query coverage can remain below 1 even for PMID-linked included studies because the published inclusion set may reflect multi-database search and manual additions, the reported PubMed strategy may be incomplete or under-specified, and present-day replay can differ from the historical search due to indexing drift and time-window mismatch, and the analysis of query coverage is shown in Table 9. Tier 1 is used only for final evaluation, while Tier 2 supports training and ablations. We report Tier 1–Tier 2 distributions in Figure 2, and provide a fuller panel with additional

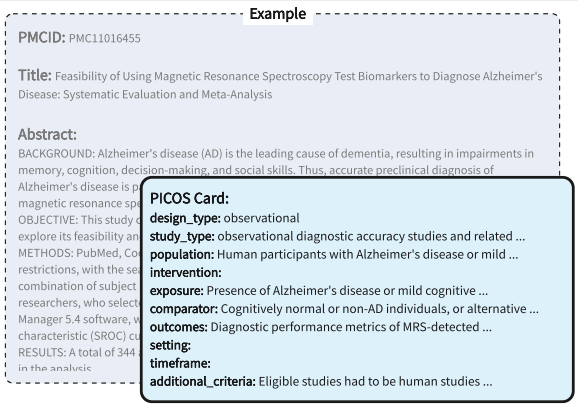



Figure 3: PICOS card example

top-level MeSH tree categories and included-to-retrieved ratios in the Appendix M. This coverage-based tiering is meant to control replay fidelity and machine-actionability; it is complementary to, not interchangeable with, source-review metadata such as venue/citation, whose subgroup effects are analyzed separately in Appendix K.

**PICOS Card Annotation.** For each meta-analysis, we create a compact PICOS card (Figure 3) summarizing eligibility (P/I(E)/C/O/S plus optional constraints such as setting or follow-up). Two medical annotators produce these cards with overlap auditing and evidence-based adjudication. The cards serve both as structured queries for retrieval/reranking and as screening prompts for inclusion decisions. Appendix A.4 provides annotator qualifications, training/calibration and workload estimates, a guideline snapshot, core field-level rules, and explicit no-leakage constraints that prohibit using results/discussion/conclusion content when writing PICOS fields.

### 4.4 Supervision Noise and Robustness

Because MetaBench uses published inclusion lists as pragmatic reference labels, retrieved-but-not-included candidates are not assumed to be perfectly ineligible. Under the same information setting as Task 2 (topic PICOS card plus candidate title/abstract), a stratified audit of 600 uniformly sampled Tier 1 negative pairs finds 2%–13% abstract-level PICOS compatibility/ambiguity across strata.

We construct cleaned Tier 1 pools by removing ambiguous/compatible negatives per topic. Re-auditing reduces estimated ambiguity from 6.40% to 5.22% in the paired audit, with an independent resample giving 5.37%. Under cleaned-pool evaluation, Task 2 precision for `n3+s1` at $N = 500$ improves from 0.4902 to 0.5081, while sensitivity is essentially unchanged (0.8564 vs. 0.8565); Task 1

is similarly stable. MetaBench should therefore be interpreted as a coverage-controlled PubMed benchmark with measured pragmatic-label noise, not a fully adjudicated eligibility benchmark.

## 5 Experiment

### 5.1 Experiment Setup

**Query views and baselines.** We evaluate Task 1 retrievers using multiple query views. Query views include TA (title+abstract), Cri (verbatim eligibility criteria), and PF (expert-curated PICOS-full). We compare Author-PubMed, Random, BM25, general dense retrievers (Contriever(Izacard et al., 2021), BGE(Xiao et al., 2023), E5(Wang et al., 2024a), GTE(Li et al., 2023)), and biomedical retrievers (MedCPT(Jin et al., 2023), BMRetriever(Xu et al., 2024), MedEmbed(Balachandran, 2024), BioLORD). For Tier 2 training, we use lightweight early-scoping views: picos-lite1 (PIS) and picos-lite2 (PIOS), and fine-tune BGE/MedEmbed under {Cri, PIS, PIOS}.

**Candidate pools.** For each topic $m$, we rerun the author-reported PubMed query within a matched time window to obtain $R_m$, then form $D_m = R_m \cup I_m$ (union with PMID-linked included studies) after deduplication. Tier 1 is used for evaluation; Tier 2 is used for training/ablations. Task 2 evaluates on the reranker-produced Top-$N$ candidates under the budgeted protocol.

In addition to the title/abstract (TA) setting used throughout the main experiments, we report an auxiliary full-text track on a PMCID-eligible subset. It is intended to quantify full-text *information gain* under explicit OA constraints and is not part of the primary benchmark results (see Appendix J).

**Evaluation metrics.** Task 1 uses Recall@K and nDCG@K, plus Effort@R and NormEffort@R $=$ Effort@R$/|D_m|$ (summarized by median/IQR and tail percentiles). Task 2 reports macro-averaged Sensitivity and e2e-Sensitivity (end-to-end sensitivity is sensitivity after composing both 2 stages), together with Specificity, Precision, and F1/F2 on Top-$N$ pools.

Full training, inference, and hardware settings for Task 1 and Task 2 are reported in Appendix L.

### 5.2 Task 1: Candidate Reranking

**Baselines.** We compare Author-PubMed, Random, BM25, and dense retrievers from general (Contriever/BGE/E5/GTE) and biomedical (MedCPT/BMRetriever/MedEmbed/BioLORD) families, plus BGE/MedEmbed variants fine-tuned on Tier 2.

**Training.** We fine-tune BGE and MedEmbed for one epoch on Tier 2 using query views {Cri, picos-lite1, picos-lite2}. Each training instance pairs a topic query with a candidate study title+abstract and is optimized with MultipleNegativesRankingLoss. Although TA can be strong in zero-shot retrieval, we avoid using TA as the training view because it is often not available or not reliable at early stage query time in real research workflows.

**Main results.** Task 1 targets *high-recall screening* rather than conventional top-$k$ retrieval. We therefore report both an *early-depth* cutoff ($K = 10$) and *screening-oriented* cutoffs ($K \in \{100, 200, 500\}$), together with $\mathrm{Effort@0.95}$ and $\mathrm{NormEffort@0.95}$, which directly quantify the inspection depth needed to reach target recall. Results on Tier 1 (Table 2 reports a subset of the retriever results; the full table is provided in Appendix B ) show:

**MetaBench supervision improves ranking quality and reduces effort.** Fine-tuned retrievers consistently outperform baselines at both shallow ($K = 10$) and screening-oriented depths, while the best variant reduces median `NormEffort@0.95` to $\approx$0.22 and maintains high recall.

**Lightweight PICOS helps.** Using **picos-lite2** (PIOS) as the query view gives the best nDCG and effort reductions, outperforming both criteria text and picos-lite1.

**Biomedical backbones remain advantageous.** Under matched settings, MedEmbed variants generally outperform BGE, suggesting benefits from domain pretraining for meta-analysis retrieval.

**Classical baselines are effort-inefficient.** Author-PubMed, Random, and BM25 require substantially deeper inspection to reach the same recall target, showing that topical matching alone is insufficient under screening-oriented budgets.

### 5.3 Task 2: Top-$N$ Screening / Inclusion Decision

**Fixed Top-$N$ protocol.** Task 2 evaluates Top-$N$ inclusion decisions on reranked candidates. A

| Model | Recall ↑ | | | | nDCG ↑ | | Effort@0.95 ↓ | | NormEffort@0.95 ↓ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | @10 | @100 | @200 | @500 | @10 | @200 | median [IQR] | p95 | median [IQR] | p95 |
| Author-Pubmed | 0.095 | 0.505 | 0.644 | 0.798 | 0.089 | 0.307 | 194.0 [60.5, 536.5] | 2859 | 0.80 [0.63, 0.92] | 0.99 |
| Random | 0.097 | 0.472 | 0.608 | 0.768 | 0.099 | 0.295 | 219.0 [76.0, 693.5] | 4147.5 | 0.94 [0.88, 0.97] | 1 |
| BM25(TA) | 0.371 | 0.801 | 0.866 | 0.915 | 0.443 | 0.627 | 80.0 [27.5, 222.0] | 1146.5 | 0.39 [0.21, 0.64] | 0.89 |
| BM25(Cri) | 0.333 | 0.761 | 0.848 | 0.906 | 0.404 | 0.593 | 91.0 [32.0, 258.0] | 1358 | 0.45 [0.25, 0.68] | 0.93 |
| BM25(PF) | 0.368 | 0.807 | 0.871 | 0.921 | 0.458 | 0.639 | 80.0 [27.5, 213.0] | 1083 | 0.39 [0.20, 0.61] | 0.87 |
| BGE(TA) | 0.428 | 0.849 | 0.900 | 0.931 | 0.534 | 0.701 | 56.0 [22.0, 143.0] | 593 | 0.27 [0.13, 0.48] | 0.79 |
| BGE(Cri) | 0.374 | 0.809 | 0.875 | 0.921 | 0.448 | 0.634 | 71.0 [25.5, 176.0] | 858.5 | 0.35 [0.18, 0.57] | 0.83 |
| BGE(PF) | 0.392 | 0.823 | 0.886 | 0.927 | 0.474 | 0.655 | 63.0 [25.0, 171.0] | 779.5 | 0.31 [0.17, 0.53] | 0.81 |
| ME(TA) | 0.428 | 0.852 | 0.901 | 0.932 | 0.528 | 0.698 | 57.0 [21.0, 143.0] | 585.5 | 0.28 [0.12, 0.47] | 0.80 |
| ME(Cri) | 0.382 | 0.813 | 0.878 | 0.925 | 0.451 | 0.636 | 68.0 [23.0, 186.0] | 815 | 0.34 [0.16, 0.54] | 0.80 |
| ME(PF) | 0.404 | 0.826 | 0.889 | 0.929 | 0.487 | 0.664 | 60.0 [23.0, 167.0] | 713 | 0.30 [0.14, 0.53] | 0.83 |
| ***BGE_cri** | 0.447 | 0.856 | 0.906 | 0.934 | 0.549 | 0.711 | 50.0 [18.0, 129.0] | 541.5 | 0.26 [0.12, 0.45] | 0.72 |
| ***BGE_lite1** | 0.455 | 0.863 | 0.909 | 0.935 | 0.558 | 0.716 | 48.0 [18.0, 124.0] | 510.5 | 0.26 [0.11, 0.45] | 0.72 |
| ***BGE_lite2** | 0.469 | 0.869 | 0.911 | 0.936 | 0.588 | 0.734 | 44.0 [17.0, 114.5] | 488.5 | **0.22 [0.10, 0.43]** | 0.71 |
| ***ME_cri** | 0.448 | 0.860 | 0.911 | 0.936 | 0.555 | 0.717 | 49.0 [17.0, 119.0] | 564.5 | 0.24 [0.11, 0.44] | 0.73 |
| ***ME_lite1** | 0.463 | 0.865 | 0.910 | 0.936 | 0.564 | 0.721 | 48.0 [19.0, 121.5] | 514 | 0.24 [0.11, 0.44] | **0.69** |
| ***ME_lite2** | **0.475** | **0.873** | **0.915** | **0.937** | **0.593** | **0.740** | **43.0 [18.0, 114.0]** | **462** | 0.23 [0.10, 0.40] | 0.71 |

Table 2: Retrieval performance and effort metrics across different retrievers on Tier 1. We report both early-depth metrics ($K = 10$) and screening-oriented cutoffs ($K = 100, 200, 500$). Higher is better for Recall and nDCG (↑); lower is better for Effort/NormEffort (↓). Models marked with * and in bold were trained on the MetaBench training set. The best performing models are highlighted in bold, while the second best are underlined.

| N | Model | Training/Inference Strategy | Sensitivity | | Spec./Prec. | | F-score | |
|---|---|---|---|---|---|---|---|---|
| | | | Sens. | Sens. (e2e) | Spec. | Prec. | F1 | F2 |
| N=500 | **GPT-5-nano** | GPT-5-nano (no reasoning) | 0.2140 | 0.2124 | **0.9920** | **0.5660** | 0.3110 | 0.2450 |
| | | GPT-5-nano (with reasoning) | 0.6570 | 0.6521 | 0.9570 | 0.4320 | 0.5210 | 0.5950 |
| | **Qwen3** | Qwen3-8B (base) | 0.3628 | 0.3608 | **0.8994** | 0.2714 | 0.2640 | 0.2962 |
| | | Qwen3-8B (n1) | **0.9704** | **0.9627** | 0.7342 | 0.2922 | 0.4124 | 0.5843 |
| | | Qwen3-8B (n4) | 0.8819 | 0.8754 | 0.8713 | 0.4385 | 0.5437 | 0.6731 |
| | | Qwen3-8B (n3+s1) | 0.8564 | 0.8504 | 0.8966 | **0.4902** | **0.5823** | **0.6932** |
| | **Llama3.1** | Llama-3.1-8B (base) | 0.7308 | 0.7233 | 0.1961 | 0.0906 | 0.1406 | 0.2363 |
| | | Llama-3.1-8B (n1) | 0.9451 | 0.9381 | 0.8153 | 0.3788 | 0.4987 | 0.6564 |
| | | Llama-3.1-8B (n4) | **0.9568** | **0.9493** | 0.7562 | 0.3100 | 0.4292 | 0.5967 |
| | | Llama-3.1-8B (n3+s1) | 0.8812 | 0.8753 | **0.8813** | **0.4626** | **0.5639** | **0.6883** |
| | **HuatuoGPT** | HuatuoGPT-o1-7B (base) | 0.6033 | 0.5996 | 0.8806 | 0.3518 | 0.3907 | 0.4642 |
| | | HuatuoGPT-o1-7B (n1) | **0.9534** | **0.9462** | 0.7842 | 0.3360 | 0.4560 | 0.6198 |
| | | HuatuoGPT-o1-7B (n4) | 0.8589 | 0.8528 | **0.8892** | **0.4719** | **0.5644** | 0.6807 |
| | | HuatuoGPT-o1-7B (n3+s1) | 0.8820 | 0.8755 | 0.8784 | 0.4529 | 0.5556 | **0.6819** |

Table 3: **Performance comparison of N=500 and training strategies.** Metrics include Sensitivity, e2e Sensitivity, Specificity, Precision, F1, and F2. We include GPT-5-NANO as a closed-source zero-shot reference. The best performing models are highlighted in **bold** for each group.

cheaper reranker first prioritizes the full candidate pool, and a stronger screener then makes include/exclude decisions in the high-yield near-miss region. The screener is a binary decision model rather than a full-pool ranker with globally comparable scores, so we separate recall-oriented prioritization from precision/calibration-oriented screening. For Tier 1 evaluation, we rerank with MEDEMBED_LITE2 and screen the top-$N$ candidates. We use $N = 500$ as the primary budget because MEDEMBED_LITE2 has `Effort@0.95` $p95 = 462$, and also report $N \in \{100, 200\}$. Inputs are `picos-lite2` plus candidate title/abstract; metrics are topic-level macro Sensitivity, e2e-Sensitivity, Specificity, Precision, F1, and F2. The reported end-to-end Task 2 results are therefore conditional on this upstream reranker. To support fixed-pool, screener-only comparisons, the MetaBench release includes the frozen Top-$\{100, 200, 500\}$ PMID lists, reranker scores and ranks, and full Tier 1 candidate-pool metadata used in this paper. This allows future screeners to be evaluated on identical candidate sets without re-running Task 1.

**Auxiliary full-text track.** Because candidate-side full text is incomplete and OA-skewed at PubMed scale, full text is not required for Task 2. Instead, we provide an auxiliary full-text track on a PMCID-eligible subset to test whether full text improves performance under the same labels. We select topics with at least 3 included studies and

20 non-included candidates with PMCID, retain all PMCID-available positives, build a PMCID-negative pool, and apply the same Tier 1/2 split. To control compute, we use the same negative-sampling strategy as in TA (e.g., n1) and compare **TA-input** vs. **full-text-input** screeners on the same subset. Details and OA-bias analysis are in Appendix J.

#### 5.3.1 Baseline Models

We evaluate three open-source LLMs spanning general and biomedical families: QWEN3-8B, LLAMA-3.1-8B, and HUATUOGPT-O1-7B. We also include a closed-source zero-shot reference, GPT-5-NANO, under two settings: without reasoning and with reasoning. For the open-source models, we test the base checkpoint and supervised fine-tuned variants trained on MetaBench Tier 2 under different strategies (Table 3 shows N=500 and full table with N=100/200/500 is shown in Appendix B). For GPT-5-NANO, we report zero-shot inference under the two reasoning settings. All models use the same top-$N$ reranked lists and the same `picos-lite2` query view.

#### 5.3.2 Training Strategies

We evaluate four settings (Table 3): **(base)** zero-shot include/exclude prediction, and three SFT variants where each positive is paired with negatives from the same meta-analysis pool: **(n1)** uses one negative, **(n4)** uses four, and **(n3+s1)** uses three sampled negatives plus one *synthetic hard negative*. The synthetic negative is generated from the annotated `picos-lite2` card by randomly mutating exactly one element from {P, I/E, O, S} and prompting GPT-5.1 to rewrite a title/abstract in the style of a sampled included study while violating only that element; outputs follow an IMRaD-like structure with realistic numbers and no copied text or real identifiers (prompt in Figure 10). According to our noise analysis, synthetic negatives are hard but clean (Appendix H) and serve as controlled stressors for fine-grained PICOS mismatch detection, although random target/template selection is a limitation. In all SFT settings, we iterate once over all Tier 2 meta-analyses and positives with the specified negative ratio; Appendix E provides additional analysis, including comparison with reranker-top mined negatives. We include this construction as an auxiliary benchmark baseline and a controlled analysis tool, rather than as a standalone hard-negative generation contribution.

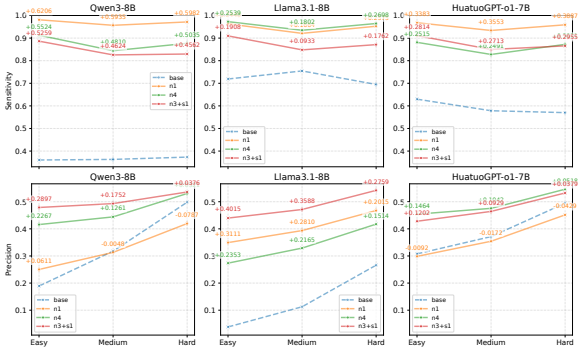



Figure 4: **Sensitivity and Precision across topic difficulty.** Performance of three LLM screeners under different SFT strategies on Tier 1, stratified by retrieval difficulty derived from the reranker's NormEffort.

| Setting | Sens. | Spec. | Prec. | F1 | F2 |
|---|---|---|---|---|---|
| TA | 0.940 | 0.739 | 0.644 | 0.764 | 0.861 |
| FT | 0.953 | 0.763 | 0.650 | 0.773 | 0.872 |
| TA(full-train) | 0.959 | 0.713 | 0.607 | 0.744 | 0.860 |

Table 4: **Auxiliary full-text track (PMCID-eligible subset; Tier 1 evaluation).** We compare paired TA vs. full-text (FT) screening under the same training strategy (n1) on a PMCID-eligible subset. The bottom row reports a distribution-mismatch reference: a TA model trained on the full corpus evaluated on the same subset.

#### 5.3.3 Main Results

All Task 2 results are evaluated on **Tier 1** and macro-averaged at the topic level. We report budgeted screening on Top-$N$ candidates with default $N$=500 and $N \in \{100, 200\}$ as references.

**Key findings.** Zero-shot models show different screening biases: Qwen3 is conservative, Llama3 aggressive, and HuatuoGPT-o1 more balanced. The closed-source GPT-5-NANO reference is highly specificity-oriented without reasoning; reasoning improves sensitivity and F-scores, but the best fine-tuned open screener, Qwen3 n3+s1, still has higher sensitivity, F1, and F2.

SFT on MetaBench consistently improves F1/F2 across backbones and budgets. Negative ratio controls the sensitivity–precision trade-off: n1 gives high sensitivity but low precision, while n4 improves precision and F-scores. n3+s1 usually further improves precision and achieves the best or near-best F1/F2, especially on reranker-hard topics (Figure 4). Because precision is measured against published inclusion lists, it should be read with the noise audits in Section 4.4: cleaned-pool evaluation improves Qwen3 n3+s1 precision from 0.4902 to 0.5081 with sensitivity essentially unchanged. Finally, the auxiliary PMCID subset shows that full text can improve screening, but this track remains OA-constrained and auxiliary (Table 4).

## 6 Conclusion

We presented **MetaBench**, a large-scale benchmark for early-stage PubMed title/abstract citation screening in meta-analysis study selection. Built with structured PICOS cards, replayed PubMed queries, and coverage-based tiering, MetaBench standardizes high-recall reranking and budgeted Top-$N$ screening at a scale not supported by prior resources. Experiments show that MetaBench supervision improves dense reranking, reduces screening effort, and strengthens LLM-based inclusion decisions; negative construction further shapes the sensitivity–precision trade-off. By keeping full text auxiliary, MetaBench preserves reproducibility while enabling future extensions toward deeper evidence-synthesis workflows.

## 7 Limitations

MetaBench supports reproducible evaluation of PubMed title/abstract retrieval and citation screening for meta-analysis, but it does not model the complete systematic-review workflow and still has several limitations and risks.

**(1) Search replay fidelity and candidate-pool quality.** Re-running author-reported PubMed queries cannot perfectly reproduce historical result sets because of incomplete reporting, multi-database review workflows, and database/indexing drift. This directly affects candidate-pool quality and the interpretation of downstream retrieval/screening results. We partially mitigate this issue through query reconstruction, audit, replay-coverage tiering, and paired sensitivity analyses; notably, reconstructed queries often reduce retrieval yield while improving coverage. However, these steps improve replay fidelity rather than fully recovering the original review-time search space (Appendix D, D.3, D.4, D.5).

**(2) Pragmatic negatives and residual label noise.** At scale, retrieved-but-not-included records are treated as pragmatic negatives, yet some may still be abstract-level PICOS-compatible or excluded for reasons not observable from title/abstract alone; conversely, even published included studies may contain residual source-review noise. We therefore conducted extensive negative-noise audits, source-specific negative audits, boundary-focused Tier 1 denoising, and noise-aware reevaluation, which reduce part of this ambiguity and show that our main conclusions are broadly robust. Still, the benchmark should be interpreted as a pragmatic large-scale supervision resource rather than a fully adjudicated gold-standard label set (Appendix G, H, I).

**(3) Metadata-based quality proxies are confounded.** Source-review metadata such as journal impact factor (IF) and citation count are imperfect quality proxies. Our stratified analyses show that differences across IF/citation bins are driven largely by changes in candidate-pool size and the number of included studies, rather than by source-review quality alone. For this reason, we release these metadata as optional analysis fields instead of using them as default benchmark filters, leaving users to apply stricter subset selection when appropriate for their use case (Appendix K, especially K.2, K.3, and K.4).

**(4) Full-text incompleteness, OA bias, and future extensions.** MetaBench's main-track screening uses title/abstract because candidate-side full text is incomplete and non-universal at PubMed scale, and restricting evaluation to Open Access (OA) full text would introduce systematic OA bias and reduce coverage. We therefore treat full text as an auxiliary analysis dimension rather than a core requirement. A natural next step is to use PMC OA/full-text subsets to support deeper noise detection and candidate-pool repair, as some inclusion/exclusion reasons are only recoverable from full text. More broadly, we view MetaBench as a foundation for future full-text tasks such as query reconstruction/improvement, data extraction and evidence synthesis, risk-of-bias assessment, and large-scale conclusion synthesis (Appendix J).

**(5) Model-dependent query reconstruction and external transfer.** Although paired sensitivity analyses show that reconstructed and original queries lead to comparable downstream conclusions, their retrieved candidate sets can differ substantially. Because the current reconstruction pipeline uses a single external LLM, model-specific reconstruction biases cannot be fully ruled out. Tier 1 evaluation therefore remains strictly anchored to author-provided executable queries. In addition, MetaBench has not yet been externally validated on general systematic-review corpora, many of which do not provide comparable structured PICOS annotations or replayable PubMed candidate construction. Future work will study reconstruction behavior across LLMs and develop transferable PICOS representations for evaluation beyond the meta-analysis-specific setting.

**(6) Upstream reranker dependence and pos-**

**sible pretraining exposure.** Task 2 performance is conditional on the candidate region produced by MEDEMBED_LITE2; alternative rerankers may expose different near-miss distributions and lead to different screening results. We mitigate this dependency by releasing the full candidate metadata and frozen Top-$\{100, 200, 500\}$ evaluation pools used in this paper, enabling controlled screener-only comparisons. Possible pretraining exposure of the evaluated LLMs to source reviews or primary studies also cannot be completely ruled out. Tier 1 topics are held out from MetaBench fine-tuning and models make decisions only from the supplied PICOS and title/abstract evidence, but these controls do not establish absence from their pretraining corpora.

## References


Leif Azzopardi, Vassil Kalphov, and Georgios Georgiadis. 2017. Sis at clef 2017 ehealth tar task. In *Conference and Labs of the Evaluation Forum*.

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models.

Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545.

Xiangming Cai, Yuanming Geng, Yiming Du, Bart Westerman, Duolao Wang, Chiyuan Ma, and Juan J. Garcia Vallejo. 2025. Utilizing large language models to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. *BMC Medical Research Methodology*, 25.

Christian Cao, J. Sang, Rahul K Arora, David Chen, Robbie Kloosterman, Matthew Cecere, J. Gorla, Richard Saleh, Ian Drennan, Bijan Teja, Michael Fehlings, Paul Everett Ronksley, Alexander A. Leung, Dany E Weisz, Harriet Ware, Mairead Whelan, David B. Emerson, Rahul K Arora, and Niklas Bobrovitz. 2025. Development of prompt templates for large language model–driven screening in systematic reviews. *Annals of Internal Medicine*, 178:389 – 401.

Gary C. K. Chan, Estrid He, Janni Leung, and Karin Verspoor. 2025. A comprehensive systematic review dataset is a rich resource for training and evaluation of ai systems for title and abstract screening. *Research Synthesis Methods*, 16:308 – 322.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Vassil Kalphov, Georgios Georgiadis, and Leif Azzopardi. 2017. Sis at clef 2017 ehealth tar task. In *CEUR workshop proceedings*, volume 1866, pages 1–5.

Wojciech Kusa, Óscar E. Mendoza, Matthias Samwald, Petr Knoth, and Allan Hanbury. 2023. Csmed: Bridging the dataset gap in automated citation screening for systematic literature reviews. *ArXiv*, abs/2311.12474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Xinyu Mao, Teerapong Leelanupab, Harrisen Scells, and G. Zuccon. 2025. Densereviewer: A screening prioritisation tool for systematic review based on dense retrieval. *ArXiv*, abs/2502.03400.

Xinyu Mao, Shengyao Zhuang, Bevan Koopman, and G. Zuccon. 2024. Dense retrieval with continuous explicit feedback for systematic review screening prioritisation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Takehiko Oami, Yohei Okada, and T. a. Nakada. 2024. Performance of a large language model in screening citations. *JAMA Network Open*, 7.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, and 1 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Mohammadreza Pourreza and Faezeh Ensan. 2022. Towards semantic-driven boolean query formalization for biomedical systematic literature reviews. *International journal of medical informatics*, 170:104928.

Rohan Sanghera, Arun J. Thirunavukarasu, Marc El Khoury, Jessica O'Logbon, Yuqing Chen, Archie Watt, Mustafa Mahmood, Hamid Butt, George Nishimura, and Andrew A S Soltan. 2024. High-performance automated abstract screening with large language model ensembles. *ArXiv*, abs/2411.02451.

Viet thi Tran, Gerald Gartlehner, Sally Yaacoub, Isabelle Boutron, Lukas Schwingshackl, Julia Stadelmaier, Isolde Sommer, Farzaneh Alebouyeh, Sivem Afach, Joerg J. Meerpohl, and Philippe Ravaud. 2024. Sensitivity and specificity of using gpt-3.5 turbo models for title and abstract screening in systematic reviews and meta-analyses. *Annals of Internal Medicine*, 177:791 – 799.

Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):55.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Shuai Wang, Harrisen Scells, Martin Potthast, Bevan Koopman, and G. Zuccon. 2023. Generating natural language queries for more effective systematic review screening prioritisation. *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*.

Shuai Wang, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and G. Zuccon. 2024b. Zero-shot generative large language models for systematic review screening automation. *ArXiv*, abs/2401.06320.

Zifeng Wang, Qiao Jin, Jiacheng Lin, Junyi Gao, Jathurshan Pradeepkumar, Pengcheng Jiang, Benjamin Danek, Zhiyong Lu, and Jimeng Sun. 2025. Trialpanorama: Database and benchmark for systematic review and design of clinical trials. *arXiv preprint arXiv:2505.16097*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May D. Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. Bmretriever: Tuning large language models as better biomedical text retrievers. *Preprint*, arXiv:2404.18443.

## Content of Appendix

A Medical-expert Topic Selection and Annotation Details
- A.1 Topic-level eligibility screening guideline
- A.2 Annotator qualifications, training, calibration, and workload
- A.3 Inter-annotator Agreement and Adjudication Protocols
- A.4 PICOS-card annotation protocol and no-leakage rules

B Full Experiments Result Table of Task 1 & 2

C Statistical Significance and Uncertainty Estimation

D Query Reconstruction Detail
- D.1 Audit Protocol
- D.2 Expert Verdict Schema and Error Taxonomy
- D.3 Tier 2 Reconstruction Quality: Recon vs.Original
- D.4 Reconstruction Sensitivity on Topics with Original Queries
- D.5 Sensitivity of Downstream Tasks to Query Reconstruction

E Analysis of Screener-Hard Negatives via Retrieval Mining and LLM Generation
- E.1 Include Rate of Negatives
- E.2 Screener Performance with Mined vs. Synthetic Hard Negatives
- E.3 Tail Behavior Across Topic Difficulty

F Sensitivity Analysis for the Query Coverage Threshold

G Negative Noise Audit on Tier 1
- G.1 Audit Design
- G.2 Results and Analysis

H Audit of Negative Sources: Random vs. Mined vs. Synthetic Hard Negatives
- H.1 Audit Design
- H.2 Results and Analysis

I Tier 1 Boundary Noise Audit and Noise-aware Evaluation
- I.1 Boundary-noise audit: sampling protocol and labels
- I.2 Boundary-noise labeling results and candidate-pool cleaning
- I.3 Re-auditing Tier 1 negative noise after cleaning (paired and resampled)
- I.4 Noise-aware evaluation: Task 1 and Task 2 comparisons
- I.5 Release and future work

J Auxiliary Full-text Track
- J.1 PMCID metadata and legal accessibility
- J.2 Full-text availability statistics and OA-skew risk
- J.3 PMCID-eligible subset construction
- J.4 Full-text retrieval and representation
- J.5 Training and evaluation protocol (paired TA vs. FT)
- J.6 Auxiliary full-text results

K Venue/Citation-based Stratification Analysis
- K.1 Distribution of IF and Citation Metadata
- K.2 Topic Scale under IF/Citation Stratification
- K.3 Task 1: Reranking under IF/Citation Stratification

K.4 Task 2: Top-$N$ Screening under IF/Citation Stratification
L Experimental Setup and Hyperparameters
M PubMed MeSH Tree-Number Distribution

# A Medical-expert Topic Selection and Annotation Details

This appendix provides the auditable annotation details that were abbreviated in the main text due to space limits. We describe: (i) the topic-level medical-expert screening protocol used to identify eligible source meta-analyses for MetaBench, and (ii) the PICOS-card annotation protocol used to convert each topic into a structured retrieval/screening representation. These procedures are intended to improve *auditability*, *machine-actionability*, and *benchmark reliability at scale*; they are not meant to serve as a full methodological appraisal of every source review at Cochrane-equivalent standards.

## A.1 Topic-level eligibility screening guideline

**Goal.** The goal of this stage is to identify source meta-analyses that are suitable for building a large-scale, reproducible PubMed-based benchmark for study retrieval and screening. This stage is a *topic-level gatekeeping procedure*: it determines whether a PubMed record corresponds to an eligible human-clinical meta-analysis for MetaBench, rather than retrospectively certifying the full methodological rigor of the original review. The guideline is shown in Figure 11.

**Interpretation.** The guideline is designed to ensure that MetaBench topics are: (i) genuine meta-analyses rather than umbrella/meta-meta reviews, (ii) grounded in human clinical evidence, (iii) compatible with PMID-based study linking and PubMed replay, and (iv) suitable for downstream retrieval and screening experiments.

## A.2 Annotator qualifications, training, calibration, and workload

**Annotator background.** All topic-level annotations were produced by two domain experts who both hold a Mainland China Physician Practicing Certificate and have more than 10 years of professional experience in medical knowledge curation and data organization.

- **Expert A:** BMed (Clinical Medicine); 10+ years of medical knowledge curation / data organization; multiple years of medical-AI dataset labeling and application development.
- **Expert B:** BMed (Clinical Medicine) + MSc (Public Health: Epidemiology & Biostatistics); strong evidence-synthesis and statistical foundation; 10+ years of medical data / knowledge curation; multiple years of medical-AI dataset labeling and application development.

**Guideline development and calibration.** The topic-level guideline was drafted jointly by the two medical experts and refined with in-lab AI researchers to ensure both annotation consistency and machine-actionable field design. Before large-scale annotation, the annotators completed pilot labeling, calibration discussion, onboarding tests, and quality checks (approximately 8 hours in total).

**Annotation material and workflow.** For topic-level eligibility screening, annotators used the source meta-analysis materials available to the project, with title/abstract evidence and gatekeeping fields serving as the primary basis for initial inclusion/exclusion decisions. A 10% overlap subset was double-labeled for agreement checking. Disagreements were resolved by rule-based adjudication grounded in the structured eligibility fields and explicit evidence. After the main pass, we additionally performed a 5% audit with targeted rework.

**Workload.** The total topic-level annotation effort was approximately 920 person-hours.

**Additional machine-actionability gates.** After topic-level expert screening, we further restricted the benchmark construction stage to PMC Open Access full texts so that included-study lists and references could be extracted and checked. We removed non-machine-actionable cases, including missing or invalid tables and count mismatches. Tier 1 further enforces PubMed index coverage $= 1$ and query coverage $\geq 0.8$, yielding a high-integrity subset with strong PubMed-only, replay-auditable properties.

## A.3 Inter-annotator Agreement and Adjudication Protocols

Table 5 reports inter-annotator agreement for the main double-labeled annotation and audit stages. Agreement is computed on the overlap subset for each task. Cohen's $\kappa$ is reported for categorical judgments.

| Task | Unit | N | Agreement | Cohen's $\kappa$ |
|---|---|---|---|---|
| Topic eligibility screening | Meta-analysis record | 2,684 | 97.09% | 0.879 |
| Query reconstruction audit | Query reconstruction case | 156 | 98.72% | 0.794 |
| Negative-noise audit | Topic–study pair | 600 | 97.33% | 0.786 |

Table 5: **Inter-annotator agreement for double-labeled annotation and audit stages.** Agreement is measured on overlap sets. For the negative-noise audit, labels follow the three-way abstract-level compatibility schema under topic PICOS card plus candidate title/abstract.

**Disagreement resolution.** For topic eligibility screening, disagreements were resolved by rule-based adjudication using the predefined eligibility fields and explicit evidence from the source record. The adjudication rules prioritized machine-actionable gatekeeping fields, including whether the record was a genuine meta-analysis, non-umbrella review, human clinical topic, eligible primary-study design, English-language source, and compatible with PMID-based benchmark construction.

For query reconstruction audit, disagreements were resolved through expert discussion and recorded in the audit log. Auditors checked whether each reconstructed query was executable under PubMed E-utilities, semantically aligned with the reported search intent, and free of disallowed time or publication-type constraints inside the core query. Non-pass cases were edited or regenerated and then re-verified for executability.

For negative-noise audits, two medical experts independently labeled sampled topic–study pairs under the same information setting as Task 2: topic PICOS card plus candidate title/abstract, without candidate full text. Disagreements were resolved through discussion using the predefined three-way schema: `definitely_incompatible`, `possibly_compatible`, and `likely_compatible`. For noise-aware evaluation, `possibly_compatible` and `likely_compatible` are treated as ambiguous/compatible negatives.

**Adjudicator limitation.** We did not use an independent third adjudicator. Instead, disagreements were resolved by discussion between the two medical experts using predefined schemas and audit logs. We therefore treat these agreement estimates as consistency checks for benchmark construction rather than as evidence of fully independent clinical adjudication.

### A.4 PICOS-card annotation protocol and no-leakage rules

**Goal.** For each eligible meta-analysis topic, we annotate a structured PICOS card for downstream retrieval and screening. The card is intended to summarize *eligibility criteria*, not empirical findings, and must therefore avoid any leakage from results-oriented sections of the paper.

**Annotators and workload.** The same two medical experts used for topic-level filtering also performed the PICOS-card annotation. Both hold Mainland China physician licensure and have more than 10 years of professional experience. The PICOS guideline was co-developed and iterated with in-house AI researchers to ensure both retrieval/screening usefulness and strict no-leakage compliance. Training and calibration for PICOS annotation required approximately 24 hours in total, including guideline walkthrough, pilot labeling, onboarding tests, and quality proofreading before production. The total PICOS-card annotation effort was approximately 2000 person-hours.

**Evidence priority and strict no-leakage rule.** Annotators use the **full text** of each meta-analysis. The highest-priority evidence comes from sections such as *Eligibility Criteria*, *Inclusion/Exclusion Criteria*, and *Study Selection*. Methods sections provide secondary support for study design, participants, interventions/exposures, comparators, and outcomes.

The following content must **not** be used to construct PICOS fields:

- Results / Findings / Synthesis / Meta-analysis results
- Discussion / Conclusion / Implications
- Abstract result/conclusion sentences

This prohibition prevents implicit label leakage such as "effective", "high accuracy", or any numerical summary of outcomes. The annotation guideline is shown in Figure 12 and Figure 13.

**Quality control.** During the training phase, pilot rounds and consistency checks were used to calibrate the annotators. During production, difficult or ambiguous cases were escalated for consensus. Outputs were checked for: (i) explicit evidence support in Methods/Eligibility sections, and (ii) violation of the no-leakage rule (e.g., result wording or numerical summaries).

**Examples.** We summarize two representative examples discussed during guideline development:

- **Diagnostic evaluation example (PMCID: PMC11016455).** The topic is labeled as `design_type=observational`; the biomarker assessment is placed under `exposure`; the `outcomes` field records the requirement for diagnostic accuracy data enabling TP/FP/TN/FN construction; no "high accuracy" wording or performance numbers are allowed.
- **Intervention example (PMCID: PMC11992344).** The topic is labeled as `design_type=trials`; acupuncture/acupressure is placed under `intervention`; sham/placebo is placed under `comparator`; `outcomes` list validated sleep measures such as PSQI/ISI; language/full-text constraints are placed in `additional_criteria`.

**Scope note.** These expert annotations are designed to make MetaBench auditable and reproducible at scale. They improve topic eligibility control and structured query/screening supervision, but they do not constitute a full review-by-review methodological appraisal of the source meta-analyses themselves.

# B Full Experiments Result Table of Task 1 & 2

The full experiments result tables of Task 1 and Task 2 mentioned in Section 5 are shown as Table 6 and Table 7.

# C Statistical Significance and Uncertainty Estimation

We quantify uncertainty for the comparison between `n3+s1` and `n4` under the primary Task 2 setting ($N = 500$). The two strategies use the same total number of negatives: `n4` uses four sampled negatives per positive, whereas `n3+s1` replaces one sampled negative with one PICOS-targeted synthetic hard negative.

We perform paired bootstrap resampling over Tier 1 meta-analysis topics with 10,000 repetitions. In each repetition, topic indices are sampled with replacement, the same sampled indices are used for both strategies, and the difference in macro-averaged F2 is recomputed. We report the observed difference, a 95% bootstrap confidence interval, and a two-sided $p$-value derived from the paired bootstrap distribution (Table 8).

The improvement from PICOS-targeted synthetic hard negatives is statistically significant for Qwen3-8B and Llama-3.1-8B. In contrast, the HuatuoGPT-o1-7B confidence interval includes zero, and the observed difference is not statistically significant. The qualitative conclusion is unchanged under a Bonferroni correction over the three backbone-level comparisons. These findings indicate that the benefit of synthetic hard negatives is backbone-dependent rather than universal.

# D Query Reconstruction Detail

## D.1 Audit Protocol

For all meta-analyses in MetaBench, including Tier 1 topics, Tier 2 topics with explicit query as well as those only provide narrative search descriptions (i.e., without an executable PubMed Boolean query), we reconstruct an executable PubMed query via an LLM-assisted pipeline and conduct a lightweight expert audit for quality assurance. The protocol is designed to ensure (i) *executability* under the official PubMed E-utilities interface and (ii) *auditable* documentation of the reconstruction process.

**Inputs.** For each meta-analysis topic, we provide the reconstruction model with (1) the meta-analysis title and abstract, and (2) the sections describing the PubMed search strategy (e.g., Eligibility Criteria, Search Strategy), when available.

**Reconstruction constraints.** The model is instructed (see Fig 8) to produce a single executable PubMed Boolean query string that: (i) uses standard PubMed syntax and fields (e.g., `[tiab]`, `[mh]`) when appropriate; (ii) is *free of* time filters and publication-type restrictions (these are applied separately by our replay protocol); (iii) avoids copying verbatim text from the source manuscript beyond terminology necessary for query formulation; and

| Model | Recall ↑ | | | | nDCG ↑ | | Effort@0.95 ↓ | | NormEffort@0.95 ↓ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | @10 | @100 | @200 | @500 | @10 | @200 | median [IQR] | p95 | median [IQR] | p95 |
| Author-Pubmed | 0.095 | 0.505 | 0.644 | 0.798 | 0.089 | 0.307 | 194.0 [60.5, 536.5] | 2859 | 0.80 [0.63, 0.92] | 0.99 |
| Random | 0.097 | 0.472 | 0.608 | 0.768 | 0.099 | 0.295 | 219.0 [76.0, 693.5] | 4147.5 | 0.94 [0.88, 0.97] | 1 |
| BM25(TA) | 0.371 | 0.801 | 0.866 | 0.915 | 0.443 | 0.627 | 80.0 [27.5, 222.0] | 1146.5 | 0.39 [0.21, 0.64] | 0.89 |
| BM25(Cri) | 0.333 | 0.761 | 0.848 | 0.906 | 0.404 | 0.593 | 91.0 [32.0, 258.0] | 1358 | 0.45 [0.25, 0.68] | 0.93 |
| BM25(PF) | 0.368 | 0.807 | 0.871 | 0.921 | 0.458 | 0.639 | 80.0 [27.5, 213.0] | 1083 | 0.39 [0.20, 0.61] | 0.87 |
| Contr(TA) | 0.399 | 0.820 | 0.878 | 0.920 | 0.497 | 0.665 | 64.0 [25.0, 181.5] | 987 | 0.34 [0.17, 0.55] | 0.83 |
| Contr(Cri) | 0.329 | 0.772 | 0.850 | 0.909 | 0.403 | 0.594 | 81.0 [31.0, 241.5] | 1334 | 0.42 [0.24, 0.63] | 0.87 |
| Contr(PF) | 0.364 | 0.796 | 0.865 | 0.914 | 0.445 | 0.624 | 75.0 [27.0, 211.0] | 1146 | 0.36 [0.20, 0.60] | 0.86 |
| BGE(TA) | 0.428 | 0.849 | 0.900 | 0.931 | 0.534 | 0.701 | 56.0 [22.0, 143.0] | 593 | 0.27 [0.13, 0.48] | 0.79 |
| BGE(Cri) | 0.374 | 0.809 | 0.875 | 0.921 | 0.448 | 0.634 | 71.0 [25.5, 176.0] | 858.5 | 0.35 [0.18, 0.57] | 0.83 |
| BGE(PF) | 0.392 | 0.823 | 0.886 | 0.927 | 0.474 | 0.655 | 63.0 [25.0, 171.0] | 779.5 | 0.31 [0.17, 0.53] | 0.81 |
| E5(TA) | 0.421 | 0.838 | 0.890 | 0.928 | 0.517 | 0.681 | 61.0 [24.0, 170.0] | 820 | 0.30 [0.15, 0.53] | 0.81 |
| E5(Cri) | 0.352 | 0.800 | 0.869 | 0.919 | 0.423 | 0.614 | 78.0 [27.5, 201.5] | 937 | 0.39 [0.21, 0.60] | 0.84 |
| E5(PF) | 0.386 | 0.823 | 0.884 | 0.926 | 0.473 | 0.656 | 64.0 [23.0, 177.5] | 867.5 | 0.31 [0.16, 0.53] | 0.80 |
| GTE(TA) | 0.430 | 0.853 | 0.903 | 0.933 | 0.529 | 0.697 | 52.0 [20.0, 135.0] | 599.5 | 0.26 [0.12, 0.50] | 0.78 |
| GTE(Cri) | 0.397 | 0.833 | 0.896 | 0.930 | 0.474 | 0.657 | 57.0 [22.0, 152.5] | 715.5 | 0.31 [0.13, 0.54] | 0.79 |
| GTE(PF) | 0.381 | 0.829 | 0.889 | 0.930 | 0.447 | 0.643 | 57.0 [22.5, 142.5] | 687 | 0.29 [0.14, 0.53] | 0.80 |
| MCPT(TA) | 0.326 | 0.780 | 0.856 | 0.909 | 0.382 | 0.585 | 85.0 [34.0, 250.0] | 1224 | 0.44 [0.24, 0.65] | 0.86 |
| MCPT(Cri) | 0.288 | 0.738 | 0.830 | 0.903 | 0.333 | 0.541 | 100.0 [37.0, 316.5] | 1293.5 | 0.51 [0.30, 0.72] | 0.91 |
| MCPT(PF) | 0.348 | 0.782 | 0.857 | 0.919 | 0.413 | 0.603 | 89.0 [31.0, 247.5] | 1132 | 0.44 [0.23, 0.67] | 0.88 |
| BMR(TA) | 0.409 | 0.826 | 0.882 | 0.924 | 0.498 | 0.663 | 64.0 [26.5, 193.0] | 838.5 | 0.31 [0.17, 0.54] | 0.83 |
| BMR(Cri) | 0.294 | 0.737 | 0.823 | 0.899 | 0.341 | 0.541 | 99.0 [34.0, 315.0] | 1441 | 0.51 [0.31, 0.69] | 0.92 |
| BMR(PF) | 0.307 | 0.765 | 0.851 | 0.912 | 0.356 | 0.563 | 93.0 [38.0, 283.0] | 1161.5 | 0.47 [0.27, 0.68] | 0.91 |
| ME(TA) | 0.428 | 0.852 | 0.901 | 0.932 | 0.528 | 0.698 | 57.0 [21.0, 143.0] | 585.5 | 0.28 [0.12, 0.47] | 0.80 |
| ME(Cri) | 0.382 | 0.813 | 0.878 | 0.925 | 0.451 | 0.636 | 68.0 [23.0, 186.0] | 815 | 0.34 [0.16, 0.54] | 0.80 |
| ME(PF) | 0.404 | 0.826 | 0.889 | 0.929 | 0.487 | 0.664 | 60.0 [23.0, 167.0] | 713 | 0.30 [0.14, 0.53] | 0.83 |
| BLORD(TA) | 0.336 | 0.776 | 0.846 | 0.905 | 0.400 | 0.593 | 83.0 [28.5, 257.5] | 1399.5 | 0.44 [0.24, 0.67] | 0.88 |
| BLORD(Cri) | 0.289 | 0.728 | 0.817 | 0.892 | 0.331 | 0.535 | 103.0 [34.0, 291.5] | 1775 | 0.53 [0.33, 0.72] | 0.90 |
| BLORD(PF) | 0.268 | 0.717 | 0.811 | 0.888 | 0.304 | 0.518 | 108.0 [34.0, 295.0] | 1768.5 | 0.53 [0.31, 0.72] | 0.89 |
| ***BGE_cri** | 0.447 | 0.856 | 0.906 | 0.934 | 0.549 | 0.711 | 50.0 [18.0, 129.0] | 541.5 | 0.26 [0.12, 0.45] | 0.72 |
| ***BGE_lite1** | 0.455 | 0.863 | 0.909 | 0.935 | 0.558 | 0.716 | 48.0 [18.0, 124.0] | 510.5 | 0.26 [0.11, 0.45] | 0.72 |
| ***BGE_lite2** | 0.469 | 0.869 | 0.911 | 0.936 | 0.588 | 0.734 | 44.0 [17.0, 114.5] | 488.5 | **0.22 [0.10, 0.43]** | 0.71 |
| ***ME_cri** | 0.448 | 0.860 | 0.911 | 0.936 | 0.555 | 0.717 | 49.0 [17.0, 119.0] | 564.5 | 0.24 [0.11, 0.44] | 0.73 |
| ***ME_lite1** | 0.463 | 0.865 | 0.910 | 0.936 | 0.564 | 0.721 | 48.0 [19.0, 121.5] | 514 | 0.24 [0.11, 0.44] | **0.69** |
| ***ME_lite2** | **0.475** | **0.873** | **0.915** | **0.937** | **0.593** | **0.740** | **43.0 [18.0, 114.0]** | **462** | 0.23 [0.10, 0.40] | 0.71 |

Table 6: Retrieval performance and effort metrics across different retrievers on Tier 1. We report both early-depth metrics ($K = 10$) and screening-oriented cutoffs ($K = 100, 200, 500$): the former reflects early ranking quality, while the latter better matches high-recall study screening. Higher is better for Recall and nDCG (↑); lower is better for Effort/NormEffort (↓). Models marked with * and in bold were trained on the MetaBench training set. The best performing models are highlighted in bold, while the second best are underlined.

(iv) outputs only the final query string in a machine-actionable format.

**Generation and normalization.** We use an external LLM (in our case, OpenAI GPT-5.1) to generate a candidate query. We then apply deterministic normalization to standardize formatting (e.g., whitespace and casing), and perform basic well-formedness checks (e.g., balanced parentheses/quotes and a whitelist of allowed field tags). Queries that fail these checks are automatically regenerated or repaired and re-checked.

**Executability verification via PubMed E-utilities.** Each reconstructed query is executed once using the official PubMed E-utilities endpoint (e.g., `esearch.fcgi`) to verify that it is accepted by the API and yields a parsable response. We log the API-returned result count for auditability. Queries that fail execution (e.g., API errors or unparsable responses) are regenerated/rewritten until they pass this executability check.

**Expert audit (sampling with overlap).** We conduct a manual expert audit on a stratified random sample comprising 10% of all topics. Two medically trained experts independently review an overlapping 2% subset (double-coding) to assess consistency. Reviewers inspect whether the reconstructed query is (a) executable as written, (b) aligned with the narrative search intent and eligibility description, and (c) free of disallowed constraints (time and publication-type filters). Discrepancies in the

| N | Model | Training Strategy | Sensitivity | | Spec./Prec. | | F-score | |
|---|---|---|---|---|---|---|---|---|
| | | | Sens. | Sens. (e2e) | Spec. | Prec. | F1 | F2 |
| N=100 | Qwen3 | Qwen3-8B (base) | 0.3769 | 0.3490 | **0.8626** | 0.3551 | 0.3240 | 0.3414 |
| | | Qwen3-8B (n1) | **0.9734** | **0.9023** | 0.5604 | 0.3210 | 0.4515 | 0.6316 |
| | | Qwen3-8B (n4) | 0.8981 | 0.8357 | 0.7530 | 0.4537 | 0.5643 | 0.6976 |
| | | Qwen3-8B (n3+s1) | 0.8749 | 0.8157 | 0.7979 | **0.5018** | **0.5994** | **0.7138** |
| | Llama3.1 | Llama-3.1-8B (base) | 0.7277 | 0.6711 | 0.2139 | 0.1413 | 0.2186 | 0.3521 |
| | | Llama-3.1-8B (n1) | 0.9525 | 0.8847 | 0.6683 | 0.3984 | 0.5243 | 0.6859 |
| | | Llama-3.1-8B (n4) | **0.9638** | **0.8939** | 0.5737 | 0.3312 | 0.4588 | 0.6337 |
| | | Llama-3.1-8B (n3+s1) | 0.8974 | 0.8354 | **0.7680** | **0.4746** | **0.5808** | **0.7083** |
| | HuatuoGPT | HuatuoGPT-o1-7B (base) | 0.6201 | 0.5754 | **0.7908** | 0.3896 | 0.4298 | 0.5022 |
| | | HuatuoGPT-o1-7B (n1) | **0.9609** | **0.8914** | 0.6176 | 0.3591 | 0.4870 | 0.6566 |
| | | HuatuoGPT-o1-7B (n4) | 0.8767 | 0.8157 | 0.7782 | **0.4836** | **0.5817** | 0.7015 |
| | | HuatuoGPT-o1-7B (n3+s1) | 0.8990 | 0.8365 | 0.7616 | 0.4652 | 0.5734 | **0.7037** |
| N=200 | Qwen3 | Qwen3-8B (base) | 0.3668 | 0.3571 | **0.8838** | 0.3065 | 0.2914 | 0.3177 |
| | | Qwen3-8B (n1) | **0.9725** | **0.9424** | 0.6560 | 0.3021 | 0.4267 | 0.6032 |
| | | Qwen3-8B (n4) | 0.8884 | 0.8627 | 0.8237 | 0.4434 | 0.5507 | 0.6818 |
| | | Qwen3-8B (n3+s1) | 0.8637 | 0.8396 | 0.8581 | **0.4937** | **0.5880** | **0.7003** |
| | Llama3.1 | Llama-3.1-8B (base) | 0.7304 | 0.7035 | 0.2042 | 0.1082 | 0.1700 | 0.2856 |
| | | Llama-3.1-8B (n1) | 0.9499 | 0.9218 | 0.7529 | 0.3848 | 0.5072 | 0.6676 |
| | | Llama-3.1-8B (n4) | **0.9605** | **0.9314** | 0.6747 | 0.3164 | 0.4386 | 0.6095 |
| | | Llama-3.1-8B (n3+s1) | 0.8892 | 0.8636 | **0.8372** | **0.4669** | **0.5702** | **0.6961** |
| | HuatuoGPT | HuatuoGPT-o1-7B (base) | 0.6082 | 0.5917 | 0.8428 | 0.3659 | 0.4059 | 0.4795 |
| | | HuatuoGPT-o1-7B (n1) | **0.9577** | **0.9289** | 0.7114 | 0.3436 | 0.4669 | 0.6341 |
| | | HuatuoGPT-o1-7B (n4) | 0.8665 | 0.8419 | **0.8459** | **0.4757** | **0.5705** | 0.6884 |
| | | HuatuoGPT-o1-7B (n3+s1) | 0.8899 | 0.8642 | 0.8322 | 0.4566 | 0.5616 | **0.6900** |
| N=500 | Qwen3 | Qwen3-8B (base) | 0.3628 | 0.3608 | **0.8994** | 0.2714 | 0.2640 | 0.2962 |
| | | Qwen3-8B (n1) | **0.9704** | **0.9627** | 0.7342 | 0.2922 | 0.4124 | 0.5843 |
| | | Qwen3-8B (n4) | 0.8819 | 0.8754 | 0.8713 | 0.4385 | 0.5437 | 0.6731 |
| | | Qwen3-8B (n3+s1) | 0.8564 | 0.8504 | 0.8966 | **0.4902** | **0.5823** | **0.6932** |
| | Llama3 | Llama-3.1-8B (base) | 0.7308 | 0.7233 | 0.1961 | 0.0906 | 0.1406 | 0.2363 |
| | | Llama-3.1-8B (n1) | 0.9451 | 0.9381 | 0.8153 | 0.3788 | 0.4987 | 0.6564 |
| | | Llama-3.1-8B (n4) | **0.9568** | **0.9493** | 0.7562 | 0.3100 | 0.4292 | 0.5967 |
| | | Llama-3.1-8B (n3+s1) | 0.8812 | 0.8753 | **0.8813** | **0.4626** | **0.5639** | **0.6883** |
| | HuatuoGPT | HuatuoGPT-o1-7B (base) | 0.6033 | 0.5996 | 0.8806 | 0.3518 | 0.3907 | 0.4642 |
| | | HuatuoGPT-o1-7B (n1) | **0.9534** | **0.9462** | 0.7842 | 0.3360 | 0.4560 | 0.6198 |
| | | HuatuoGPT-o1-7B (n4) | 0.8589 | 0.8528 | **0.8892** | **0.4719** | **0.5644** | 0.6807 |
| | | HuatuoGPT-o1-7B (n3+s1) | 0.8820 | 0.8755 | 0.8784 | 0.4529 | 0.5556 | **0.6819** |

Table 7: **Performance comparison across different sample sizes (N) and training strategies.** Metrics include Sensitivity, end-to-end Sensitivity (e2e), Specificity, Precision, F1, and F2. The best performing models are highlighted in **bold** for each group.

overlapping subset are resolved through discussion and documented as part of the audit log.

**Documentation and release.** For each reconstructed topic, we maintain a reconstruction record including the final executable query and the result count. For audited topics, we additionally store reviewer notes and edit annotations. The finalized reconstructed queries and audit documentation are released alongside the benchmark to support transparency and downstream reuse.

### D.2 Expert Verdict Schema and Error Taxonomy

**Verdict schema.** For the manual audit (10% stratified sample; 2% overlap double-coded), each reconstructed query is assigned one of four verdicts. The verdicts focus on (i) *executability* under PubMed E-utilities and (ii) *semantic alignment* with the paper-reported search intent and eligibility description. Note that time-window constraints and publication-type exclusions are *not* included in the reconstructed `core_query`; instead, they are appended by a unified replay template to ensure consistent evaluation. Concretely, the replay template: (a) applies an article-type exclusion block (e.g., excluding review-like and editorial publication types); and (b) applies a date window extracted from the meta-analysis when available, where the end date is set to the latest included-study publication date if the paper-reported end date is missing or earlier, and otherwise follows the paper-reported end date (so that included studies are not dropped due to overly restrictive date filters).

- **Pass (Accepted as-is).** The reconstructed

| Model | ΔF2 | 95% CI | $p$-value |
|---|---|---|---|
| Qwen3-8B | $+0.0200$ | $[0.0121,\ 0.0278]$ | $< 0.001$ |
| Llama-3.1-8B | $+0.0916$ | $[0.0806,\ 0.1028]$ | $< 0.001$ |
| HuatuoGPT-o1-7B | $+0.0012$ | $[-0.0067,\ 0.0092]$ | $0.7486$ |

Table 8: **Paired bootstrap comparison of `n3+s1` against `n4` at $N = 500$.** Positive ΔF2 favors `n3+s1`. Confidence intervals and $p$-values are computed from 10,000 paired bootstrap resamples over Tier 1 meta-analysis topics.

`core_query` is directly executable (accepted by the E-utilities API with a valid, parsable response), and is semantically consistent with the narrative search strategy and eligibility description. The query contains the core concepts implied by the review (at minimum the **Population** and **Intervention/Exposure** components; **Outcome/Setting** are included when explicitly specified in the review) and does not introduce forbidden constraints (e.g., time filters or publication-type restrictions inside `core_query`). No edits are required before replay.

- **Minor edit (Accepted with minor fixes).** The query is largely correct in intent and coverage, but requires small, non-substantive edits such as field-tag corrections (e.g., `[tiab]` vs. `[mh]`), bracket/quote/parenthesis fixes, minor synonym additions, spelling corrections, or local Boolean adjustments that do not materially change the search intent.

- **Major edit (Substantive revision required).** The query exhibits semantic issues that materially affect intent fidelity, such as missing a key PICOS component (P/I/E/O/S as required by the review description), being obviously too narrow or too broad, Boolean structure that causes intent drift, or mistakenly encoding exclusion criteria as inclusion (or vice versa). Substantial rewriting is required prior to replay.

- **Reject / Regenerate.** The query is either not executable (e.g., cannot be accepted by the API or produces an invalid/unparsable response) or is severely misaligned with the review-reported intent in a way that cannot be repaired with reasonable edits. The query must be regenerated under stricter constraints or rebuilt manually.

**Resolution policy for audited non-pass cases.** All audited instances labeled as **Minor edit**, **Major edit**, or **Reject / Regenerate** are corrected (edited) or regenerated and re-verified for executability via the PubMed E-utilities API. The finalized, executable versions are included in the released MetaBench data, together with audit annotations indicating the original verdict and the applied correction type.

**Audit outcome reporting.** On the audited 10% sample, we report the distribution of verdicts: Pass (97.31%), Minor edit (1.75%), Major edit (0.58%), and Reject / Regenerate (0.36%). These rates quantify semantic-fidelity issues in the initial reconstruction output; audited non-pass cases are resolved as described above prior to release.

**Semantic error taxonomy (retrieval intent).** To characterize why non-pass cases occur, we tag semantic errors into the following categories (not mutually exclusive):

- **Missing PICOS components.** The query omits a required eligibility element implied by the narrative strategy (commonly *Outcome* or *Setting*, or a key subpopulation qualifier), leading to intent mismatch.

- **Overly narrow formulation.** The query relies on overly restrictive terms (e.g., MeSH-only with limited synonyms, a single narrow term, or overly specific phrasing), which can miss PMID-linked included studies under replay.

- **Overly broad formulation.** The query uses broad OR expansions or weak conjunction structure (e.g., large OR blocks without sufficient AND constraints between core concepts), inflating the candidate pool with off-target records.

- **Inclusion/exclusion inversion.** The query incorrectly encodes exclusion criteria as inclusion terms (or vice versa), resulting in systematic intent drift.

### D.3 Tier 2 Reconstruction Quality: Recon vs. Original

**Query reconstruction quality and long-tail coverage.** To characterize the fidelity of reconstructed PubMed strategies in **Tier 2** (Recon; $n$=2210), we compare their *query coverage* ($\mathrm{Cov}_Q$, i.e., the fraction of PMID-linked included studies recovered by replay) against topics with executable author queries extracted from the **full text** (FT; $n$=2116) and **supplements** (Supp; $n$=3475). As shown in Table 9, reconstructed queries exhibit *higher high-coverage rate* and a *smaller low-coverage tail*: $\Pr(\mathrm{Cov}_Q \geq 0.8)$ is higher for Recon (0.5623) than FT (0.4778) and Supp (0.4979), while the fractions of low-coverage topics are consistently lower for Recon (e.g., $\Pr(\mathrm{Cov}_Q < 0.5)$: 0.1533 vs. 0.2670/0.1843; and $\Pr(\mathrm{Cov}_Q < 0.2)$: 0.0499 vs. 0.1621/0.0956). These results indicate that reconstruction reduces the long-tail of poorly reproducible topics under the same replay protocol, which is important for stabilizing supervision quality in Tier 2.

| $\mathrm{Cov}_Q$ | ratio | | |
|---|---|---|---|
| | **FT(2116)** | **Supp(3475)** | **Recon(2210)** |
| $\geq 0.8$ | 0.4778 | 0.4979 | 0.5623 |
| $< 0.5$ | 0.2670 | 0.1843 | 0.1533 |
| $< 0.2$ | 0.1621 | 0.0956 | 0.0499 |

Table 9: **Long-tail analysis of query coverage across query sources.** We report the fraction of topics whose replayed PubMed search achieves high query coverage ($\mathrm{Cov}_Q \geq 0.8$) and falls into low-coverage regimes ($\mathrm{Cov}_Q < 0.5$, $\mathrm{Cov}_Q < 0.2$), comparing author-provided executable queries extracted from full texts (FT; $n$=2116) and supplements (Supp; $n$=3475) against LLM-reconstructed queries used in Tier 2 (Recon; $n$=2210). Higher $\mathrm{Cov}_Q \geq 0.8$ and lower low-coverage fractions indicate reduced long-tail and better replay fidelity.

| **Stats** | $\mathrm{Cov}_Q \uparrow$ | | **# Query Result** $\downarrow$ | |
|---|---|---|---|---|
| | **Original** | **Recon** | **Original** | **Recon** |
| mean | 0.69 | **0.769** | 722.089 | **565.154** |
| p10 | 0.286 | **0.4** | **14** | 17 |
| Q1 | 0.5 | **0.625** | 48 | **47** |
| median | 0.75 | **0.833** | 190 | **148** |
| Q3 | 0.923 | **1** | 633.5 | **513.75** |
| p90 | **1** | **1** | 1933.8 | **1365.5** |

Table 10: **Tier 2 Original vs. Tier 2 Recon: query coverage and yield statistics.** Distributional statistics of query coverage ($\mathrm{Cov}_Q$; higher is better) and query result count (yield; lower is better) for Tier 2 topics with executable author queries that do not meet Tier 1 thresholds (Tier 2 Original) versus LLM-reconstructed queries (Tier 2 Recon). Recon achieves higher $\mathrm{Cov}_Q$ while retrieving fewer candidates on average, suggesting a favorable coverage–yield trade-off for Tier 2 training pools.

**Tier 2 pools: coverage–yield trade-off (Original vs. Recon).** We further contrast the Tier 2 subset that has author-provided executable queries but fails Tier 1 coverage thresholds (Tier 2 Original) against Tier 2 Recon, in terms of both replay query coverage and retrieval yield (result set size). Table 10 shows that Tier 2 Recon achieves higher coverage while retrieving fewer candidates: the mean $\mathrm{Cov}_Q$ increases from 0.69 to 0.769 (median 0.75 to 0.833), whereas the mean yield decreases from 722 to 565 (median 190 to 148; $p$90 1933.8 to 1365.5). Overall, reconstructed queries tend to recover a larger fraction of PMID-linked inclusions *with smaller candidate pools*, which is desirable for Tier 2 training as it reduces the prevalence of unknown negatives and limits noise from overly broad retrieval.

### D.4 Reconstruction Sensitivity on Topics with Original Queries

**Paired sensitivity of reconstructed vs. original search replay.** To further characterize reconstruction fidelity beyond the narrative-only subset, we also reconstruct queries for *all topics that already have executable author queries* (excluding topics without any original query), and replay both the original and reconstructed queries using the same PubMed API protocol. We compare (i) retrieval yield (number of retrieved PMIDs), (ii) query coverage $\mathrm{Cov}_Q$, and (iii) paired sensitivity statistics that quantify how candidate pools change under reconstruction.

**Yield reduction and long-tail compression.** Table 11 shows that reconstructed replay consistently reduces yield across **Overall**, **Tier 1**, and **Tier 2**. The reduction is especially pronounced in the tail (e.g., $p$90 decreases from 3904 to 1726 overall, and from 4822.5 to 1799.1 in Tier 2), indicating that reconstruction tends to avoid overly broad retrieval and compresses extreme candidate-pool sizes.

**Coverage effects differ by tier.** Table 12 shows that $\mathrm{Cov}_Q$ *improves overall* (mean 0.703 $\rightarrow$ 0.7675; median 0.8235 $\rightarrow$ 0.8462), but the gains are driven primarily by **Tier 2**: mean 0.6195 $\rightarrow$ 0.742 and median 0.70 $\rightarrow$ 0.8182, with substantial improvements in the lower quantiles ($p$10: 0 $\rightarrow$ 0.3333). **Tier 1** experiences a coverage decrease because of their query coverage threshold of 0.8.

**Pool sensitivity: $\Delta$ coverage, $\Delta$ yield, and Jaccard overlap.** To quantify how sensitive candidate pools are to query discrepancies, Table 13 reports paired statistics for $\Delta\mathrm{Cov}_Q$ (Recon$-$Orig), $\Delta$Yield (Recon$-$Orig), and the Jaccard overlap between the two retrieved PMID sets. Overall, reconstruction reduces yield (mean $\Delta\mathrm{Yield} = -540.3$; median $= -18$), while coverage changes are heterogeneous (overall median $\Delta\mathrm{Cov}_Q = 0$). Importantly, Jaccard overlaps are modest (overall mean 0.2488; Tier 2 mean 0.2262), indicating that *pool composition is highly sensitive* to differences between reconstructed and original queries, especially in Tier 2. We therefore (i) exclude reconstructed queries from Tier 1, and (ii) release paired replay statistics and retrieved PMID lists to make reconstruction effects auditable.

### D.5 Sensitivity of Downstream Tasks to Query Reconstruction

To address concerns that GPT-based query reconstruction is under-described and may introduce bias in Tier 2 training, we conduct a controlled sensitivity analysis on topics that have *both* an author-provided executable query and a reconstructed (recon) query. Our goal is *not* to optimize query writing of Meta-analysis, but to quantify (i) how candidate pools change under reconstruction, and (ii) whether such changes materially affect the two-stage pipeline performance.

**Stage-1 retrieval and reranking: recon reruns vs. author queries.** Table 14 compares retrieval effectiveness and effort when candidate pools are generated by author queries versus reconstructed queries. First, simply *re-running* reconstructed queries as a candidate generator (ReconRerun) yields higher recall and ranking quality than AuthorPubMed and reduces screening effort, suggesting that reconstruction can recover missing or under-specified terms in narrative-only strategies. Second, on a paired random sampled subset of $n$=1500 topics with *both* pools available, we train the same reranker using either the reconstructed pools (TrainRecon1500) or the original pools (TrainOrig1500). We observe that (a) training substantially improves over the untrained baseline regardless of which pool is used, and (b) TrainRecon1500 and TrainOrig1500 achieve very similar Recall/nDCG/Effort, with recon being marginally better overall.

**Stage-2 screening: counterfactual training on recon vs. original pools.** To directly test downstream sensitivity, we further fine-tune an LLM screener on the same paired $n$=1500 topics, using either the reconstructed pools (SFT-Recon) or the original pools (SFT-Orig), while keeping the prompt format, screener backbone, and evaluation protocol fixed. As shown in Table 17, both fine-tuned variants yield large gains over the zero-shot base model across budgets ($N \in \{100, 200, 500\}$). Importantly, the differences between SFT-Recon and SFT-Orig are small across Sensitivity/Specificity/Precision and F-scores, indicating that our Task 2 conclusions (e.g., SFT improves screening quality under budgeted pools) are *robust* to the choice of candidate pool generator on topics where both are available.

**Takeaway.** Overall, these results support the design choice that reconstructed queries are appropriate for *Tier 2 training/analysis* while keeping *Tier 1 evaluation* strictly grounded in author-provided executable queries. Even though reconstruction can alter candidate pools, the end-to-end retrieval+screening conclusions remain stable under a counterfactual swap between reconstructed and original pools on paired topics.

## E Analysis of Screener-Hard Negatives via Retrieval Mining and LLM Generation

### E.1 Include Rate of Negatives

To assess whether *reranker-hard* negatives are also challenging for *LLM screeners*, we measure the **include rate** (false-include rate, since all are negatives) on two paired negative sets per topic: (i) **mined** negatives selected as top-ranked non-included studies under the best encoder-based reranker, and (ii) **random** negatives sampled from the same candidate pool. Table 15 shows that, across the Qwen3 base model and all SFT checkpoints, mined negatives exhibit substantially higher include rates than random negatives (e.g., $\Delta$ ranges from $+0.2167$ to $+0.6231$), indicating that **reranker-hard negatives are genuinely near-miss cases for LLM screening** and concentrate false positives.

We further compare **LLM-generated** PICOS-targeted hard negatives against random negatives (Table 16). For models not trained on synthetic negatives (base/n1/n4), generated negatives are also frequently (and consistently more than random)

| | Yield (Query Result Number) ↓ | | | | | |
|---|---|---|---|---|---|---|
| | Overall | | Tier 1 | | Tier 2 | |
| stats | Original | Recon | Original | Recon | Original | Recon |
| mean | 1275.82 | **735.503** | 1120.62 | **622.358** | 1330.8 | **775.587** |
| p10 | **17** | 18 | 38.7 | **20** | **13** | 17 |
| Q1 | 61.75 | **47** | 101 | **52** | 49 | **46** |
| median | 245 | **145** | 325 | **150.5** | 214 | **143** |
| Q3 | 906.25 | **532.5** | 1044.25 | **490.5** | 849.5 | **552** |
| p90 | 3904 | **1726** | 3082.7 | **1564** | 4822.5 | **1799.1** |

Table 11: **Yield (query result number) statistics across Overall, Tier 1, and Tier 2 for Original vs. Recon.** For topics with executable author queries, we replay both the original and reconstructed PubMed strategies under the same API protocol and report summary statistics of retrieval yield (number of retrieved PMIDs). Lower is better as it indicates fewer candidates to screen; reductions in upper-tail quantiles reflect long-tail compression.

| | Query Coverage ↑ | | | | | |
|---|---|---|---|---|---|---|
| | Overall | | Tier 1 | | Tier 2 | |
| stats | Original | Recon | Original | Recon | Original | Recon |
| mean | 0.703 | **0.7675** | **0.9402** | 0.8395 | 0.6195 | **0.742** |
| p10 | 0.125 | **0.375** | **0.8333** | 0.5385 | 0 | **0.3333** |
| Q1 | 0.5233 | **0.6492** | **0.8824** | 0.7867 | 0.3846 | **0.6** |
| median | 0.8235 | **0.8462** | **0.9677** | 0.9167 | 0.7 | **0.8182** |
| Q3 | **1** | **1** | **1** | **1** | 0.9 | **0.963** |
| p90 | **1** | **1** | **1** | **1** | **1** | **1** |

Table 12: **Query coverage statistics across Overall, Tier 1, and Tier 2 for Original vs. Recon.** We report summary statistics of query coverage $\mathrm{Cov}_Q$ (fraction of PMID-linked included studies recovered by replay) for original vs. reconstructed strategies under identical replay settings. Higher is better. Coverage gains are primarily driven by Tier 2, while Tier 1 shows a decrease, motivating our design choice to keep Tier 1 evaluation anchored to author-provided strategies and restrict reconstruction to Tier 2 usage.

misclassified as `include`, confirming that **LLM-generated negatives can act as screener-hard examples**. Thus, this result should be interpreted as **training effectiveness on the injected synthetic distribution**, rather than evidence that generated negatives are intrinsically easier than mined negatives.

Taken together, these findings justify using **two complementary hard-negative sources** in MetaBench: (1) **reranker-mined** negatives that are top-ranked near-misses and empirically hard for LLM screening, and (2) **LLM-generated** PICOS-targeted negatives that serve as controlled, high-signal counterexamples and are challenging for models that have not been explicitly trained on them. This dual design supports both realism (mined near-misses) and controllability (PICOS-targeted synthesis) when constructing hard negatives for screener training.

### E.2 Screener Performance with Mined vs. Synthetic hard negatives

Table 19 reports Qwen3-8B screening results under the budgeted Top-$N$ protocol with additional experiments in which reranked-mined negatives are used. Among the hard-negative strategies, `n3+s1` (injecting PICOS-targeted LLM-generated hard negatives) achieves the strongest overall F1/F2 across budgets, showing that controlled PICOS violations are effective training signals for reducing false positives without collapsing recall. We additionally evaluate `n3+m1`, which replaces the synthetic hard negative with an encoder-mined top-ranked negative (a near-miss from the reranker head). Compared to `n3+s1`, `n3+m1` tends to be more conservative (higher specificity but lower sensitivity), leading to slightly lower overall F2. This trade-off aligns with our include-rate diagnostics: encoder-mined negatives are genuinely near-miss and thus concentrate false-includes for both base and fine-tuned screeners, while synthetic negatives provide a more controllable signal that can more consistently improve average performance.

### E.3 Tail Behavior Across Topic Difficulty.

Table 20 further breaks down F1/F2 by retrieval difficulty (Easy/Medium/Hard). Across all $N$, hard topics exhibit substantially higher screening difficulty, and training provides the largest gains there, reflecting the practical importance of robustness in the high-effort tail. Comparing `n3+s1` vs. `n3+m1`, we observe a consistent pattern: `n3+s1` is generally best overall, while `n3+m1` can be slightly more competitive on the Hard subset in some set-

| stats | Paired | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Δ Query Coverage | | | Δ Yield | | | Jaccard | | |
| | Overall | Tier 1 | Tier 2 | Overall | Tier 1 | Tier 2 | Overall | Tier 1 | Tier 2 |
| mean | 0.0641 | -0.1007 | 0.1225 | -540.314 | -498.262 | -555.213 | 0.2488 | 0.3126 | 0.2262 |
| p10 | -0.3 | -0.3892 | -0.25 | -2730 | -2118.4 | -3193.2 | 0.0075 | 0.0313 | 0.0039 |
| Q1 | -0.0719 | -0.1538 | -0.0249 | -395 | -547.5 | -350.75 | 0.0452 | 0.08847 | 0.0347 |
| median | 0 | 0 | 0 | -18 | -60 | -8 | 0.1624 | 0.2333 | 0.1363 |
| Q3 | 0.1818 | 0 | 0.2935 | 80.25 | 19 | 113 | 0.3802 | 0.4768 | 0.3448 |
| p90 | 0.5385 | 0.0667 | 0.667 | 732 | 480.2 | 814.4 | 0.6539 | 0.7533 | 0.591 |

Table 13: **Paired statistics for ΔQuery Coverage, ΔYield, and Jaccard across Overall, Tier 1, and Tier 2.** For each topic, we compute paired differences between reconstructed and original replay: $\Delta\mathrm{Cov}_Q = \mathrm{Cov}_Q^{\mathrm{Recon}} - \mathrm{Cov}_Q^{\mathrm{Orig}}$, $\Delta$Yield = $\mathrm{Yield}^{\mathrm{Recon}} - \mathrm{Yield}^{\mathrm{Orig}}$, and the Jaccard overlap between the retrieved PMID sets from the two strategies. Lower (more negative) ΔYield indicates fewer retrieved candidates; Jaccard captures pool composition sensitivity to query discrepancies.

| Model | Recall ↑ | | | nDCG ↑ | Effort@0.95 ↓ | | NormEffort@0.95 ↓ | |
|---|---|---|---|---|---|---|---|---|
| | @100 | @200 | @500 | @200 | median [IQR] | p95 | median [IQR] | p95 |
| AuthorPubMed | 0.5048 | 0.6438 | 0.7977 | 0.3069 | 194.0 [60.5, 536.5] | 2859 | 0.80 [0.63, 0.92] | 0.99 |
| Random | 0.4717 | 0.6081 | 0.7683 | 0.2945 | 219.0 [76.0, 693.5] | 4147.5 | 0.94 [0.88, 0.97] | 1 |
| ReconRerun | 0.5429 | 0.672 | 0.777 | 0.3389 | 107.0 [38.0, 286.5] | 1495.6 | 0.79 [0.63, 0.92] | 1 |
| Untrained | 0.8264 | 0.8894 | 0.929 | 0.6642 | 60.0 [23.0, 167.0] | 713 | 0.30 [0.14, 0.53] | 0.828 |
| FullTrain | 0.873 | 0.9148 | 0.9371 | 0.7404 | 43.0 [18.0, 114.0] | 462 | 0.23 [0.10, 0.40] | 0.7113 |
| TrainRecon1500 | 0.8606 | 0.905 | 0.9332 | 0.7074 | 48.0 [20.0, 127.0] | 597.5 | 0.26 [0.11, 0.47] | 0.7568 |
| TrainOrig1500 | 0.8455 | 0.9021 | 0.9324 | 0.6902 | 55.0 [21.0, 144.0] | 604.5 | 0.28 [0.12, 0.51] | 0.7939 |

Table 14: **Task 1 sensitivity to query reconstruction.** Retrieval and effort metrics under different candidate pool generators and training settings. RECONRERUN denotes using reconstructed queries to rerun retrieval; TRAINRECON1500/TRAINORIG1500 train the same reranker on $n$=1500 paired topics using reconstructed vs. original pools. Higher is better for Recall/nDCG (↑); lower is better for Effort/NormEffort (↓).

| Qwen3-8B | Rand. | Mined | Δ (Mined–Rand.) |
|---|---|---|---|
| Base | 0.0948 | 0.3116 | +0.2167 |
| n1 | 0.1155 | 0.7386 | +0.6231 |
| n4 | 0.0526 | 0.5578 | +0.5052 |
| n3+s1 | 0.0669 | 0.6335 | +0.5665 |

Table 15: **Include rate (false-include rate) on encoder-mined hard negatives vs. random negatives.** All items are negatives; higher include rate indicates harder negatives for the screener.

| Qwen3-8B | Rand. | Gen. | Δ (Gen.–Rand.) |
|---|---|---|---|
| Base | 0.0940 | 0.2805 | +0.1865 |
| n1 | 0.1155 | 0.5912 | +0.4757 |
| n4 | 0.0494 | 0.4837 | +0.4343 |

Table 16: **Include rate (false-include rate) on LLM-generated PICOS-targeted negatives vs. random negatives.** For base/n1/n4, generated negatives behave as hard stressors.

tings, supporting the intuition that *reranker-mined near-misses better match the hardest real candidates produced by the reranker*. Together with the include-rate results (mined ≫ random in false-include rate), these findings motivate treating **both** encoder-mined negatives and PICOS-targeted synthetic negatives as meaningful "hard negatives" for screener training: the former emphasize realism and near-miss hardness, and the latter offer controlled, PICOS-focused counterexamples that improve average precision and stability.

## F Sensitivity Analysis for the Query Coverage Threshold

We use query coverage ($\mathrm{Cov}_Q$) as a *replay-fidelity* proxy: it measures the fraction of PMID-linked included studies that are recovered when re-running an executable PubMed search strategy. A higher $\mathrm{Cov}_Q$ reduces the risk that "missing positives" are due to query replay mismatch rather than model behavior, and therefore improves the integrity of downstream evaluation.

To justify the default cutpoint, we sweep $\mathrm{Cov}_Q$ thresholds on all topics with *author-provided executable queries* (i.e., Tier 1 + Tier 2_Original) and report the resulting subset size and coverage statistics (Table 18), together with the empirical CDF of $\mathrm{Cov}_Q$ (Figure 5). Raising the threshold monotonically improves replay fidelity but reduces the evaluation set size: moving from $\mathrm{Cov}_Q \geq 0.7$ to $\geq 0.8$ removes 152 topics (1449 → 1297), while moving from $\geq 0.8$ to $\geq 0.9$ removes an additional 426 topics (1297 → 871). In exchange, the typical replay fidelity increases from mean/median $0.9102/0.931$ at $\geq 0.7$ to $0.9403/0.9677$ at $\geq 0.8$, and further to $0.9786/1.0$ at $\geq 0.9$. This indicates

| N | Training Strategy | Sensitivity | | Spec./Prec. | | F-score | |
|---|---|---|---|---|---|---|---|
| | | Sens. | Sens. (e2e) | Spec. | Prec. | F1 | F2 |
| **N=100** | Qwen3-8B (base) | 0.3769 | 0.3490 | 0.8626 | 0.3551 | 0.3240 | 0.3414 |
| | SFT-RECON | 0.7728 | 0.7023 | 0.5630 | 0.3194 | 0.4249 | 0.5551 |
| | SFT-ORIG | 0.7714 | 0.7139 | 0.5555 | 0.3201 | 0.4253 | 0.5564 |
| **N=200** | Qwen3-8B (base) | 0.3668 | 0.3571 | 0.8838 | 0.3065 | 0.2914 | 0.3177 |
| | SFT-RECON | 0.7687 | 0.7310 | 0.6288 | 0.3074 | 0.4101 | 0.5399 |
| | SFT-ORIG | 0.7709 | 0.7372 | 0.6247 | 0.3070 | 0.4082 | 0.5370 |
| **N=500** | Qwen3-8B (base) | 0.3628 | 0.3608 | 0.8994 | 0.2714 | 0.2640 | 0.2962 |
| | SFT-RECON | 0.7595 | 0.7482 | 0.6768 | 0.3007 | 0.4002 | 0.5266 |
| | SFT-ORIG | 0.7619 | 0.7518 | 0.6773 | 0.3008 | 0.3990 | 0.5245 |

Table 17: **Task 2 counterfactual screening with reconstructed vs. original pools.** On $n$=1500 paired topics, we fine-tune the screener using candidates from reconstructed pools (SFT-RECON) or original pools (SFT-ORIG) and evaluate on Top-$N$ budgets. Both substantially outperform the zero-shot base model, while SFT-RECON and SFT-ORIG achieve comparable Sensitivity/Precision and F-scores across $N \in \{100, 200, 500\}$.

| Metric | Query Coverage | | | |
|---|---|---|---|---|
| | $\geq$ **0.7** | $\geq$ **0.8** | $\geq$ **0.9** | $=$ **1** |
| #topic | 1449 | 1297 | 871 | 602 |
| mean cov | 0.9102 | 0.9403 | 0.9786 | 1 |
| median cov | 0.931 | 0.9677 | 1 | 1 |

Table 18: **Query coverage statistics under different $\mathrm{Cov}_Q$ thresholds.** We report the number of topics retained and the mean/median query coverage after applying thresholds $\mathrm{Cov}_Q \geq \{0.7, 0.8, 0.9, 1.0\}$ on the set of topics with author-provided executable queries (Tier 1 + Tier 2_Original).

a clear *marginal trade-off*: tightening from 0.8 to 0.9 yields relatively limited gains for an already high-coverage subset, yet incurs a substantially larger reduction in evaluation breadth and statistical power.

The CDF in Figure 5 provides an intuitive view of this trade-off. The distribution exhibits a heavy tail of low-coverage topics and a dense mass in the high-coverage regime; thus, a threshold at 0.8 effectively filters the low-fidelity tail while retaining a sizable, diverse set of topics for robust benchmarking, whereas 0.9 begins to exclude a large fraction of topics concentrated in the $[0.8, 0.9)$ band. In our benchmark, Tier 1 further applies additional gatekeeping constraints beyond $\mathrm{Cov}_Q$ (e.g., PMID-linkage and other machine-actionability checks), and we release scripts to regenerate tiers under alternative thresholds for ablation or use-case-specific operating points.

# G Negative Noise Audit on Tier 1

## G.1 Audit Design

To quantify potential label noise induced by treating "retrieved-but-not-included" records as negatives, we conducted a small-scale manual audit on **Tier 1** under the same information setting as Task 2 (topic PICOS card + study title/abstract; no full-text access assumed). Our goal is to estimate the *abstract-level PICOS compatibility/ambiguity rate* among negatives, i.e., the fraction of candidate studies labeled as negatives ($d \in D_m \setminus I_m$) that may still appear compatible with the topic's PICOS description at the title/abstract level.

**Stratified topic sampling.** We stratified Tier 1 topics along two dataset-intrinsic dimensions: (i) **query coverage level** (**High**: $\mathrm{Cov}_Q \geq 0.9$ vs. **Low**: $0.8 \leq \mathrm{Cov}_Q < 0.9$), and (ii) **candidate pool size** $|D_m|$ (**A**: $\leq p33$, **B**: $(p33, p66]$, **C**: $> p66$ within Tier 1). From each of the $3 \times 2$ strata, we uniformly sampled 10 topics, resulting in 60 topics in total.

**Negative sampling and labeling.** For each sampled topic, we uniformly sampled 10 negatives from $D_m \setminus I_m$ (600 topic–study pairs total). Two medically trained annotators independently assessed each pair and assigned one of three **abstract-level** labels: `definitely_incompatible` (clear PICOS mismatch), `possibly_compatible` (borderline or insufficient information in title/abstract), or `likely_compatible` (no obvious PICOS mismatch at the abstract level). We report the *compatibility/ambiguity rate* as the fraction of audited negatives judged `possibly_compatible` or `likely_compatible` (i.e., an upper-bound estimate of "suspected false negatives" under abstract-level screening).

## G.2 Results and Analysis

Table 21 summarizes the estimated abstract-level compatibility/ambiguity rates across strata. Overall, the measured rates range from **2% to 13%**, indicating that while pragmatic negatives are

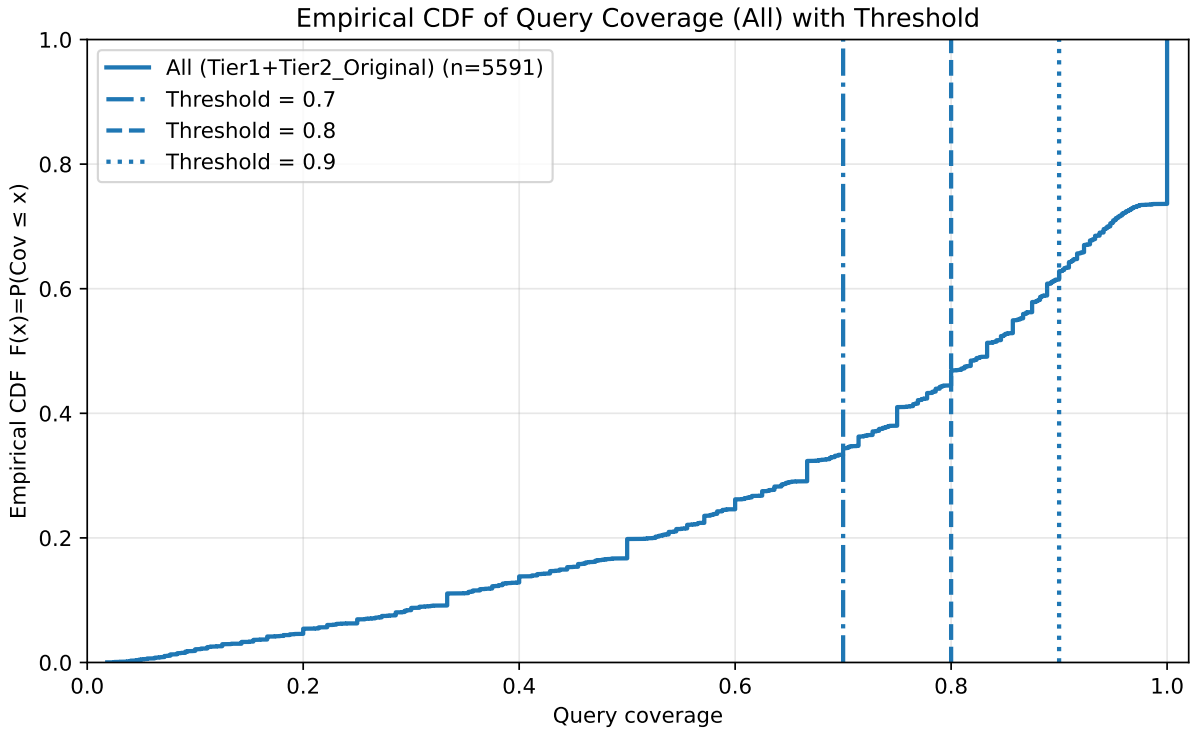



Figure 5: **Empirical CDF of query coverage with candidate thresholds.** Empirical CDF of query coverage $\mathrm{Cov}_Q$ over all topics with author-provided executable queries (Tier 1 + Tier 2_Original; $n = 5591$), with vertical reference lines at $\mathrm{Cov}_Q \in \{0.7, 0.8, 0.9\}$. The plot visualizes the fraction of topics below each threshold and highlights the trade-off between replay fidelity (higher $\mathrm{Cov}_Q$) and evaluation set size.

not perfectly clean (as expected in meta-analysis pipelines), the majority of negatives are judged clearly incompatible in most regimes.

**Pool size dominates the ambiguity rate.** Ambiguity is highest in small candidate pools (A: 9%–13.47%) and substantially lower in large pools (C: 2.0%–3.0%). This pattern is consistent with the intuition that narrower pools concentrate more topically similar, borderline candidates, whereas larger pools contain a higher proportion of obviously mismatched records.

**Query coverage is not a monotonic driver of negative ambiguity within Tier 1.** Within Tier 1, splitting topics by $\mathrm{Cov}_Q$ at 0.9 yields no consistent monotonic effect across pool-size strata (e.g., A shows higher ambiguity in Low vs. High coverage, while B shows the opposite, and C shows only small differences). This supports our clarification that coverage-based tiering primarily improves *query replay fidelity and metric interpretability*, rather than guaranteeing "clean" negatives.

# H Audit of Negative Sources: Random vs. Mined vs. Synthetic Hard Negatives

## H.1 Audit Design

To better understand how *negative source* affects label noise and downstream screening behavior, we conducted a manual audit comparing three negative constructions used in our Task 2 training: (i) **Random** negatives uniformly sampled from $D_m \setminus I_m$ (the same pool used by `n4`); (ii) **Mined** negatives (`m1`) selected as the reranker top-ranked non-included study for each topic (experiment detail can be seen in Appendix E); and (iii) **Synthetic** PICOS-targeted hard negatives (`s1`) generated by mutating exactly one PICOS element while keeping other elements compatible (Section 5.3.2).

**Stratified topic sampling.** We sampled **360** Tier 2 topics via stratified sampling across **pool-size** and **query-coverage** strata. Specifically, we binned topics by candidate pool size $|D_m|$ into three sections using within-Tier 2 quantiles: **A** ($|D_m| \leq p33$), **B** ($p33 < |D_m| \leq p66$), and **C** ($|D_m| > p66$). We also binned topics by query coverage $\mathrm{Cov}_Q$ into three levels (**low/medium/high**) using within-Tier 2 quantiles. This yields $3 \times 3$=9 strata, from each

| N | Training Strategy | Sensitivity | | Spec./Prec. | | F-score | |
|---|---|---|---|---|---|---|---|
| | | Sens. | Sens. (e2e) | Spec. | Prec. | F1 | F2 |
| **N=100** | Qwen3-8B (n1) | **0.9734** | **0.9023** | 0.5604 | 0.3210 | 0.4515 | 0.6316 |
| | Qwen3-8B (n4) | 0.8981 | 0.8357 | 0.7530 | 0.4537 | 0.5643 | 0.6976 |
| | Qwen3-8B (n3+s1) | 0.8749 | 0.8157 | 0.7979 | **0.5018** | **0.5994** | **0.7138** |
| | **Qwen3-8B (n3+m1)** | 0.8210 | 0.7666 | **0.8082** | 0.4994 | 0.5807 | 0.6816 |
| **N=200** | Qwen3-8B (n1) | **0.9725** | **0.9424** | 0.6560 | 0.3021 | 0.4267 | 0.6032 |
| | Qwen3-8B (n4) | 0.8884 | 0.8627 | 0.8237 | 0.4434 | 0.5507 | 0.6818 |
| | Qwen3-8B (n3+s1) | 0.8637 | 0.8396 | 0.8581 | **0.4937** | **0.5880** | **0.7003** |
| | **Qwen3-8B (n3+m1)** | 0.8152 | 0.7925 | **0.8627** | 0.4886 | 0.5677 | 0.6685 |
| **N=500** | Qwen3-8B (n1) | **0.9704** | **0.9627** | 0.7342 | 0.2922 | 0.4124 | 0.5843 |
| | Qwen3-8B (n4) | 0.8819 | 0.8754 | 0.8713 | 0.4385 | 0.5437 | 0.6731 |
| | Qwen3-8B (n3+s1) | 0.8564 | 0.8504 | 0.8966 | **0.4902** | **0.5823** | **0.6932** |
| | **Qwen3-8B (n3+m1)** | 0.8090 | 0.8036 | **0.8995** | 0.4830 | 0.5604 | 0.6603 |

Table 19: **Qwen3-8B budgeted screening performance under different Top-$N$ budgets.** We compare negative-sampling strategies for supervised fine-tuning: `n1` (1 random negative per positive), `n4` (4 random negatives), `n3+s1` (3 random + 1 PICOS-targeted LLM-generated hard negative), and `n3+m1` (3 random + 1 encoder-mined top-ranked hard negative). Metrics are macro-averaged over Tier 1 topics.

| N | Training Strategy | Easy | | Medium | | Hard | | Overall | |
|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 |
| **N=100** | Qwen3-8B (n1) | 0.4172 | 0.6053 | 0.4688 | 0.6430 | 0.5544 | 0.7178 | 0.4515 | 0.6316 |
| | Qwen3-8B (n3+s1) | **0.6007** | **0.7242** | **0.5909** | **0.6974** | 0.6268 | **0.7273** | **0.5994** | **0.7138** |
| | **Qwen3-8B (n3+m1)** | 0.5767 | 0.6846 | 0.5727 | 0.6665 | **0.6325** | 0.7270 | 0.5807 | 0.6816 |
| **N=200** | Qwen3-8B (n1) | 0.3867 | 0.5700 | 0.4458 | 0.6173 | 0.5508 | 0.7135 | 0.4267 | 0.6032 |
| | Qwen3-8B (n3+s1) | **0.5879** | **0.7111** | **0.5791** | **0.6816** | 0.6243 | 0.7207 | **0.5880** | **0.7003** |
| | **Qwen3-8B (n3+m1)** | 0.5612 | 0.6703 | 0.5603 | 0.6526 | **0.6299** | **0.7226** | 0.5677 | 0.6685 |
| **N=500** | Qwen3-8B (n1) | 0.3682 | 0.5457 | 0.4334 | 0.6005 | 0.5504 | 0.7130 | 0.4124 | 0.5843 |
| | Qwen3-8B (n3+s1) | **0.5817** | **0.7040** | **0.5731** | **0.6737** | 0.6222 | 0.7167 | **0.5823** | **0.6932** |
| | **Qwen3-8B (n3+m1)** | 0.5531 | 0.6619 | 0.5529 | 0.6439 | **0.6273** | **0.7182** | 0.5604 | 0.6603 |

Table 20: **F1/F2 stratified by topic difficulty for Qwen3-8B.** Topics are grouped into Easy/Medium/Hard by the reranker's retrieval difficulty (NormEffort-based bins). We report macro-averaged F1/F2 for each training strategy and Top-$N$ budget, highlighting tail robustness where screening effort is highest.

of which we sampled 40 topics (360 total).

**Per-topic paired negatives and labeling.** For each sampled topic, we obtained exactly one negative from each source (Random/Mined/Synthetic), resulting in **three paired negatives per topic**. Two medically trained annotators labeled each topic–study pair under the same information setting as Task 2 (topic PICOS card + study title/abstract; abstract-level only). We used a three-way abstract-level compatibility label: `definitely_incompatible`, `possibly_compatible` (borderline/insufficient abstract evidence), and `likely_compatible`.

### H.2 Results and Analysis

Table 22 reports, for each stratum, the mean ambiguity/noise rate (with Wilson 95% confidence intervals) for each negative source. The pattern is highly consistent across strata: **(i) mined negatives are frequently ambiguous**, **(ii) random negatives are moderately ambiguous**, and **(iii) synthetic PICOS negatives are almost always unambiguous at the abstract level**.

**Mined hard negatives are *hard but noisy*.** Across all strata, the mined negative source exhibits very high ambiguity rates (0.275–0.475), indicating that reranker top-ranked non-included candidates are often *abstract-level compatible/ambiguous* even though they are not in the inclusion list. This supports the interpretation that mined hard negatives concentrate near the screening boundary and can include substantial "compatible-but-not-included" cases (e.g., excluded for non-PICOS reasons or due to insufficient abstract evidence), thus injecting noisy supervision.

**Synthetic PICOS negatives are *hard but clean*.** In contrast, the synthetic negative source has very little ambiguity. While the Wilson upper bounds reflect finite sample size, the observed ambiguity is minimal, consistent with the intended construction of synthetic negatives as controlled PICOS-violating near-misses.

| Pool Size Stratum | Query Coverage Level | |
|---|---|---|
| | **High** ($\mathrm{Cov}_Q \geq 0.9$) | **Low** ($0.8 \leq \mathrm{Cov}_Q < 0.9$) |
| **A (Small;** $\leq p33$**)** | 0.0900 | 0.1347 |
| **B (Medium;** $(p33, p66]$**)** | 0.0900 | 0.0200 |
| **C (Large;** $> p66$**)** | 0.0200 | 0.0300 |

Table 21: **Tier 1 negative noise audit (manual).** We report the abstract-level PICOS compatibility/ambiguity rate among audited negatives, defined as the fraction labeled `possibly_compatible` or `likely_compatible` under topic PICOS cards and study title/abstracts. Topics are stratified by candidate pool size $|D_m|$ (A/B/C using Tier 1 $p33/p66$ cutoffs) and query coverage level (High: $\mathrm{Cov}_Q \geq 0.9$; Low: $0.8 \leq \mathrm{Cov}_Q < 0.9$). Each of the $3 \times 2$ strata contains 10 topics, and each topic contributes 10 uniformly sampled negatives (600 pairs total).

**Paired comparisons confirm large source effects.** Because each topic contributes a paired triple (random/mined/synthetic), we additionally compute paired deltas and McNemar tests (Table 23). Synthetic negatives have substantially lower ambiguity than mined negatives in every stratum (mean deltas $-0.250$ to $-0.475$, all $p \ll 0.01$), and mined negatives are more ambiguous than random negatives (mean deltas $+0.175$ to $+0.400$). These findings provide direct evidence that *hard-negative mining can amplify pragmatic-negative ambiguity*, and that our PICOS-targeted synthetic negatives offer a cleaner form of hard supervision.

# I Tier 1 Boundary Noise Audit and Noise-aware Evaluation

This appendix reports (i) a *boundary-noise* audit that targets reranker-head near-miss negatives, (ii) paired and resampled re-audits following our original Tier 1 negative-noise protocol (Table 21), and (iii) downstream Task 1/Task 2 comparisons under IGNORE-AMBIGUOUS and CLEAN-POOL evaluation.

## I.1 Boundary-noise audit: sampling protocol and labels

**Stratification.** We reuse the same Tier 1 stratification as Table 21: (i) **pool-size stratum** $|D_m|$ using Tier 1 $p33/p66$: **A** (small), **B** (medium), **C** (large); and (ii) **query-coverage level**: **High** ($\mathrm{Cov}_Q \geq 0.9$) vs. **Low** ($0.8 \leq \mathrm{Cov}_Q < 0.9$). Table 24 reports the number of Tier 1 topics in each cell.

**Reranker-head sampling (Top-$k$).** To concentrate on *reranker-head* near-misses (the region that most affects budgeted Top-$N$ screening precision), we first rank each Tier 1 candidate pool with our best Task 1 reranker (MEDEMBED_LITE2 with `picos-lite2` query view), and then sample top-ranked *negatives* (non-included studies) per topic. Based on the original manual audit (Table 21), three strata exhibit higher ambiguity rates; we therefore sample **Top-3** negatives for these strata and **Top-1** negatives for the remaining strata. This yields a total of **2,791** topic–study pairs for manual labeling:

$$(280 + 148 + 319) \times 3 + (108 + 368 + 74) \times 1 = 2791.$$

**Labeling guideline.** All sampled topic–study pairs are labeled under the same information setting as Task 2 (topic `picos-lite2` card + study title/abstract; no full-text), using the same three-way abstract-level compatibility schema as Table 21: `definitely_incompatible`, `possibly_compatible`, and `likely_compatible`. We treat `possibly` and `likely` as *ambiguous/compatible* negatives (upper-bound suspected false negatives under abstract-only screening).

## I.2 Boundary-noise labeling results and candidate-pool cleaning

Among the 2,791 reranker-head negative pairs, we label **630** as `likely_compatible` (22.57%), **225** as `possibly_compatible` (8.06%), and the remaining **1,936** as `definitely_incompatible` (69.37%). Therefore, the reranker-head **ambiguity/compatibility rate** is

$$\frac{630 + 225}{2791} = 0.3063 \quad (30.63\%).$$

We remove the ambiguous/compatible negatives (all `possibly` and `likely` pairs) *per topic* from Tier 1 candidate pools, producing **cleaned Tier 1 pools** for noise-aware evaluation.

## I.3 Re-auditing Tier 1 negative noise after cleaning (paired and resampled)

To quantify how boundary cleaning affects the *global* negative-noise estimate (uniform sampling from $D_m \setminus I_m$), we re-run the original Tier 1 negative-noise audit protocol (Table 21) on cleaned

| Section | Cov. level | n | Synthetic | Random | Mined |
|---|---|---|---|---|---|
| A | high | 40 | 0.025 [0.004, 0.129] | 0.100 [0.040, 0.231] | 0.475 [0.329, 0.625] |
| A | low | 40 | 0.025 [0.004, 0.129] | 0.125 [0.055, 0.261] | 0.300 [0.181, 0.454] |
| A | medium | 40 | 0.000 [0.000, 0.088] | 0.150 [0.071, 0.291] | 0.475 [0.329, 0.625] |
| B | high | 40 | 0.025 [0.004, 0.129] | 0.075 [0.026, 0.199] | 0.475 [0.329, 0.625] |
| B | low | 40 | 0.025 [0.004, 0.129] | 0.000 [0.000, 0.088] | 0.275 [0.161, 0.428] |
| B | medium | 40 | 0.050 [0.014, 0.165] | 0.100 [0.040, 0.231] | 0.400 [0.263, 0.554] |
| C | high | 40 | 0.050 [0.014, 0.165] | 0.025 [0.004, 0.129] | 0.325 [0.201, 0.480] |
| C | low | 40 | 0.000 [0.000, 0.088] | 0.000 [0.000, 0.088] | 0.350 [0.221, 0.505] |
| C | medium | 40 | 0.050 [0.014, 0.165] | 0.025 [0.004, 0.129] | 0.350 [0.221, 0.505] |

Table 22: **Manual audit of ambiguity/noise rates by negative source (Tier 2; stratified).** Each cell reports the mean of the binary ambiguity indicator $\texttt{noise}(d)$ with Wilson 95% CI. Topics are stratified by candidate pool size section (**A/B/C** via within-Tier 2 $p33/p66$ of $|D_m|$) and query-coverage level (low/medium/high via within-Tier 2 quantiles). From each of the 9 strata we sampled 40 topics (360 total). For each topic, we labeled one paired negative from each source: Random, Mined (reranker top-ranked non-included), and Synthetic (PICOS-targeted).

| Section | Cov. level | n | Δ(Syn–Mined) | $p$ | Δ(Mined–Rand) | $p$ |
|---|---|---|---|---|---|---|
| A | high | 40 | $-0.450$ | $7.6 \times 10^{-6}$ | $+0.375$ | $2.7 \times 10^{-4}$ |
| A | low | 40 | $-0.275$ | $3.4 \times 10^{-3}$ | $+0.175$ | $1.1 \times 10^{-1}$ |
| A | medium | 40 | $-0.475$ | $3.8 \times 10^{-6}$ | $+0.325$ | $2.3 \times 10^{-3}$ |
| B | high | 40 | $-0.450$ | $4.0 \times 10^{-5}$ | $+0.400$ | $1.4 \times 10^{-4}$ |
| B | low | 40 | $-0.250$ | $6.3 \times 10^{-3}$ | $+0.275$ | $9.8 \times 10^{-4}$ |
| B | medium | 40 | $-0.350$ | $5.1 \times 10^{-4}$ | $+0.300$ | $1.8 \times 10^{-3}$ |
| C | high | 40 | $-0.275$ | $3.4 \times 10^{-3}$ | $+0.300$ | $4.8 \times 10^{-4}$ |
| C | low | 40 | $-0.350$ | $1.2 \times 10^{-4}$ | $+0.350$ | $1.2 \times 10^{-4}$ |
| C | medium | 40 | $-0.300$ | $4.1 \times 10^{-3}$ | $+0.325$ | $9.7 \times 10^{-4}$ |

Table 23: **Paired deltas of ambiguity/noise rates (manual; McNemar test).** For each topic, we obtain a paired triple of negatives (Synthetic, Mined, Random) and compute per-topic deltas of the binary ambiguity indicator $\texttt{noise}(d)$. We report the mean delta within each stratum and the $p$-value from McNemar's test on discordant pairs (two-sided exact test). Across all strata, Synthetic is significantly less ambiguous than Mined. Mined has a higher ambiguity rate than Random in every stratum, with the difference reaching statistical significance in eight of nine strata.

pools. We report: (i) a **paired** re-audit on the same sampled topics used previously, and (ii) an **independent resample** under the same stratification.

**Paired re-audit (same topics).** On the original audit topics, the overall ambiguity rate decreases from 6.40% to 5.22%. Table 26 reports per-stratum rates before vs. after cleaning.

**Resampled re-audit (new topics).** On an independently resampled set of topics under the same stratification, the overall ambiguity rate is 5.37%, consistent with the paired estimate (Table 27).

### I.4 Noise-aware evaluation: Task 1 and Task 2 comparisons

We evaluate two noise-aware settings: **(i) IGNORE-AMBIGUOUS:** ambiguous/compatible negatives are treated as *unknown* for metric computation (not counted as false positives); **(ii) CLEAN-POOL:** ambiguous/compatible PMIDs are removed from Tier 1 candidate pools and evaluation is rerun on the cleaned pools. For Task 1, removing ambiguous negatives directly changes ranking positions and therefore effort/nDCG; thus Task 1 uses the cleaned-pool computation. For Task 2, we report both IGNORE-AMBIGUOUS (recompute metrics) and CLEAN-POOL (remove then re-select Top-$N$ and re-run screening).

**Task 1.** Table 28 compares the best reranker on the original vs. cleaned Tier 1 pools. All metrics improve slightly, consistent with removing reranker-head ambiguous negatives that occupy early ranks.

**Task 2.** Table 29 reports overall screening results under the best training strategy (`n3+s1`) across $N \in \{100, 200, 500\}$. Both IGNORE-AMBIGUOUS and CLEAN-POOL improve precision/specificity (and thus F-scores) while keeping sensitivity nearly unchanged, indicating that precision-oriented conclusions are robust under conservative noise handling.

### I.5 Release and future work

**Release plan.** In the next Tier 1 release, we will include: (i) the list of Tier 1 PMIDs labeled as `possibly_compatible` or `likely_compatible` (per topic), and (ii) the corresponding cleaned can-

| Pool-size stratum | #Topics in Tier 1 | | Boundary sampling | |
|---|---|---|---|---|
| | **High** ($\mathrm{Cov}_Q \geq 0.9$) | **Low** ($0.8 \leq \mathrm{Cov}_Q < 0.9$) | **Top-$k$** | **#Pairs** |
| A (Small; $\leq p33$) | 280 | 148 | 3 | $(280 + 148) \times 3$ |
| B (Medium; $(p33, p66]$) | 319 | 108 | High: 3, Low: 1 | $319 \times 3 + 108 \times 1$ |
| C (Large; $> p66$) | 368 | 74 | 1 | $(368 + 74) \times 1$ |
| | | **Total** | | **2,791** |

Table 24: **Tier 1 stratification counts and reranker-head sampling plan.** Topics are stratified by pool size (A/B/C) and query-coverage level (High/Low). We sample top-ranked *negatives* from the best reranker: Top-3 for three higher-ambiguity strata (A_high, A_low, B_high) and Top-1 for the remaining strata, resulting in 2,791 topic–study pairs for manual labeling.

| Label | #Pairs | Rate |
|---|---|---|
| `likely_compatible` | 630 | 0.2257 |
| `possibly_compatible` | 225 | 0.0806 |
| `definitely_incompatible` | 1936 | 0.6937 |
| **Ambiguous** (`likely+possibly`) | 855 | **0.3063** |

Table 25: **Boundary-noise labeling results on reranker-head negatives (Tier 1).** We labeled 2,791 reranker-head negative pairs using the abstract-level compatibility schema. Ambiguous/compatible negatives (`possibly` or `likely`) account for 30.63% and are removed (per topic) to construct cleaned Tier 1 candidate pools.

| Stratum | Original | Cleaned | Δ (cleaned–orig.) |
|---|---|---|---|
| A_high | 0.0900 | 0.0500 | −0.0400 |
| A_low | 0.1347 | 0.1220 | −0.0127 |
| B_high | 0.0900 | 0.0500 | −0.0400 |
| B_low | 0.0200 | 0.0300 | +0.0100 |
| C_high | 0.0200 | 0.0000 | −0.0200 |
| C_low | 0.0300 | 0.0300 | +0.0000 |
| **Overall** | **0.0640** | **0.0522** | −0.0118 |

Table 26: **Paired re-audit of Tier 1 negative noise after boundary cleaning.** We re-run the original Tier 1 negative-noise audit protocol (uniform sampling of negatives) on the same audited topics, and report the abstract-level ambiguity/compatibility rate (fraction of `possibly+likely`). Overall noise decreases from 6.40% to 5.22%.

didate pools with these PMIDs removed, enabling reproducible noise-aware evaluation.

**Future work.** We will continue Tier 1 denoising in subsequent releases, targeting a global abstract-level ambiguity rate below **5%** under our audit protocol. For Tier 2, due to its scale, we plan to expand this line into an explicit **noise identification** task: developing and benchmarking models that detect and filter ambiguous/compatible negatives (and other pragmatic-negative noise), thereby improving the quality and usability of large-scale supervision for evidence-synthesis retrieval and screening.

| Stratum | Resampled (cleaned) |
|---|---|
| A_high | 0.1122 |
| A_low | 0.0500 |
| B_high | 0.0700 |
| B_low | 0.0300 |
| C_high | 0.0300 |
| C_low | 0.0300 |
| **Overall** | **0.0537** |

Table 27: **Resampled re-audit of Tier 1 negative noise after boundary cleaning.** We independently resample topics under the same Tier 1 stratification and re-run the original negative-noise audit protocol on cleaned pools. The overall ambiguity/compatibility rate is 5.37%, consistent with the paired re-audit estimate.

# J Auxiliary Full-text Track

This appendix details an *auxiliary* full-text (FT) track that quantifies the information gain of candidate-side full text under explicit PMC Open Access (OA) constraints. Because candidate-side full text is *incomplete and non-universal* at PubMed scale and is OA-skewed, this track is reported as an auxiliary analysis rather than a core benchmark requirement.

## J.1 PMCID metadata and legal accessibility

MetaBench is defined on a PubMed (PMID) universe, where both included studies and retrieved candidates are PMID-linked. When available, we additionally store each study's PMCID to enable optional retrieval of publicly accessible full texts from PMC OA. We do *not* assume that full text exists or is redistributable for all candidates, and we do not require FT input for the main tasks.

## J.2 Full-text availability statistics and OA-skew risk

**Per-topic FT ratio.** To characterize candidate-side FT availability, for each topic $m$ we compute the ratio

$$r_m = \frac{\#\{d \in D_m \ : \ \texttt{PMCID}(d) \text{ available}\}}{|D_m|}, \quad (6)$$

| Model | Recall ↑ | | | nDCG ↑ | Effort@0.95 ↓ | | NormEffort@0.95 ↓ | |
|---|---|---|---|---|---|---|---|---|
| | @100 | @200 | @500 | @200 | median [IQR] | p95 | median [IQR] | p95 |
| **Original** | 0.8730 | 0.9148 | 0.9371 | 0.7404 | 43.0 [18.0, 114.0] | 462 | 0.23 [0.10, 0.40] | 0.7113 |
| **Clean-Pool** | 0.8737 | 0.9149 | 0.9371 | 0.7588 | 42.0 [17.0, 113.0] | 462 | 0.23 [0.09, 0.39] | 0.7012 |

Table 28: **Task 1 noise-aware evaluation on Tier 1 (original vs. cleaned pools).** We remove reranker-head ambiguous/compatible negatives (`possibly+likely`) per topic to form a cleaned candidate pool and recompute Task 1 metrics. Higher is better for Recall/nDCG (↑); lower is better for Effort/NormEffort (↓).

| N | Setting | Sens. | Sens. (e2e) | Spec. | Prec. | F1 | F2 |
|---|---|---|---|---|---|---|---|
| 100 | Original | 0.8749 | 0.8157 | 0.7979 | 0.5018 | 0.5994 | 0.7138 |
| | IGNORE-AMBIGUOUS | 0.8744 | 0.8159 | 0.8077 | 0.5198 | 0.6112 | 0.7202 |
| | CLEAN-POOL | 0.8745 | 0.8159 | 0.8078 | 0.5200 | 0.6114 | 0.7203 |
| 200 | Original | 0.8637 | 0.8396 | 0.8581 | 0.4937 | 0.5880 | 0.7003 |
| | IGNORE-AMBIGUOUS | 0.8637 | 0.8397 | 0.8661 | 0.5115 | 0.5997 | 0.7067 |
| | CLEAN-POOL | 0.8638 | 0.8397 | 0.8662 | 0.5117 | 0.5998 | 0.7068 |
| 500 | Original | 0.8564 | 0.8504 | 0.8966 | 0.4902 | 0.5823 | 0.6932 |
| | IGNORE-AMBIGUOUS | 0.8564 | 0.8504 | 0.9040 | 0.5080 | 0.5939 | 0.6996 |
| | CLEAN-POOL | 0.8565 | 0.8504 | 0.9041 | 0.5081 | 0.5941 | 0.6997 |

Table 29: **Task2 noise-aware evaluation (overall; Tier1).** We compare the original evaluation against IGNORE-AMBIGUOUS (recompute metrics by treating ambiguous/compatible negatives as unknown) and CLEAN-POOL (remove ambiguous PMIDs, re-select Top-$N$, and re-run screening). Noise-aware settings consistently improve specificity/precision and F-scores with nearly unchanged sensitivity.

| Split | **median**(FT ratio $r_m$) | **mean**(FT ratio $r_m$) |
|---|---|---|
| Tier 1 | 0.37698 | 0.40337 |
| Tier 2 | 0.39470 | 0.42284 |

Table 30: **Tier-level FT ratio.** $r_m$ is the fraction of candidates with PMCID (PMC OA full text available) for topic $m$.

where $D_m$ is the topic-specific candidate pool defined in Section 3 (union of query-retrieved candidates and PMID-linked inclusions after deduplication). $r_m$ summarizes the fraction of candidates whose full texts are accessible via PMC OA.

**Tier-level comparison.** Figure 6 compares the distribution of $r_m$ for Tier 1 vs. Tier 2 (each tier normalized to sum to 1). We also summarize the mean and median FT ratios (see Table 30). The two tiers exhibit highly overlapping distributions and similar summary statistics, suggesting that FT availability is not a primary driver of the Tier 1/Tier 2 split. However, the ratio is far from 1 for most topics, reflecting that candidate-side FT access is *incomplete* and varies across topics.

**Why this motivates an auxiliary track.** Even when some candidates have PMC OA full texts, requiring FT for *all* candidates (or symmetrically for both positives and negatives) would: (i) shrink evaluation to a small OA-only slice, (ii) introduce systematic OA skew (e.g., venue/time/subfield bias), and (iii) reduce reproducibility in settings where full texts cannot be legally redistributed. Therefore, we keep TA as the main-track representation and treat FT as an auxiliary, explicitly OA-constrained analysis dimension.

### J.3 PMCID-eligible subset construction

**Motivation for topic-level thresholds.** To avoid degenerate per-topic evaluation (e.g., topics with only 1 FT positive or too few FT negatives), we define a PMCID-eligible subset using minimal stability constraints:

- **FT-positive count**: the number of included studies with PMCID is at least 3;
- **FT-negative count**: the number of non-included candidates with PMCID is at least 20.

The intuition is that $\geq$ 3 FT positives reduces topic-level metric degeneracy (0/1 sensitivity), and $\geq$ 20 FT negatives permits a non-trivial specificity/precision estimate per topic under sampling-based evaluation.

**Split protocol.** After filtering topics by the above criteria, we apply the same Tier 1/Tier 2 split mechanism as in the main benchmark: Tier 2 is used for training and Tier 1 for evaluation, ensuring the auxiliary track remains compatible with the main experimental pipeline. We emphasize that this auxiliary subset is OA-constrained and may not represent the full PubMed distribution.

### J.4 Full-text retrieval and representation

**Retrieval.** For each candidate with PMCID, we retrieve its PMC OA full text (e.g., via PMC APIs)

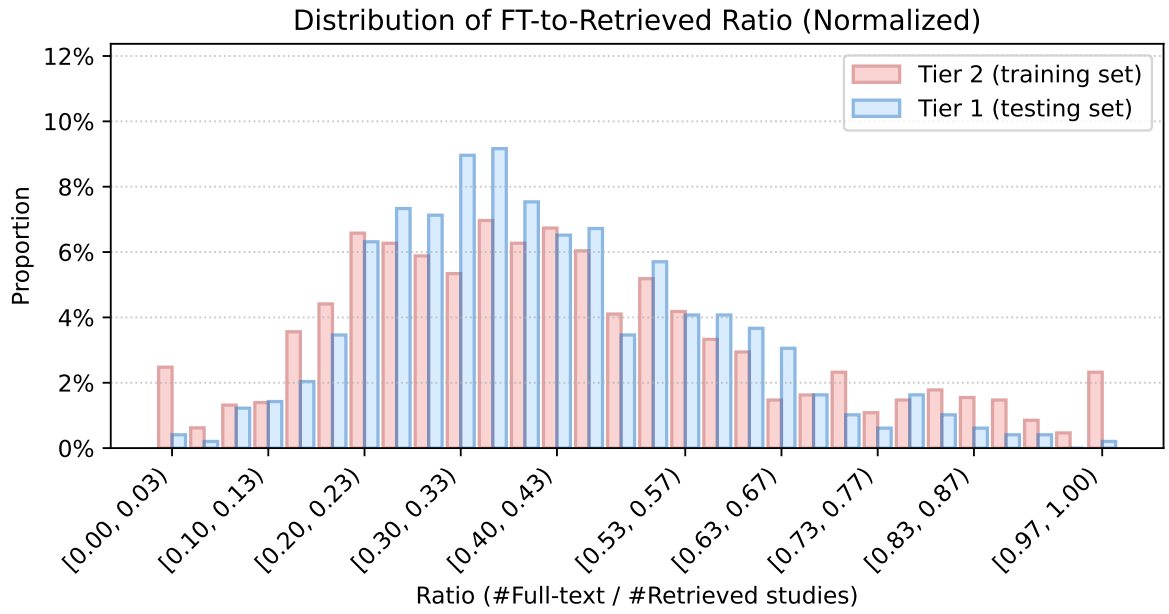



Figure 6: **Distribution of candidate-side full-text availability (FT ratio) across tiers.** For each topic, we compute $r_m = (\#\text{candidates with PMCID})/|D_m|$. The histogram shows the normalized proportion of topics falling into each FT-ratio bin, reported separately for Tier 1 and Tier 2. The substantial overlap indicates that FT availability is not a split artifact, while ratios well below 1 highlight incomplete and OA-skewed full-text access at PubMed scale.

using the PMCID identifier. This procedure is deterministic and auditable from the released PMCID metadata.

**Input formatting and truncation.** Full texts can exceed model context windows; thus we use a deterministic truncation policy to bound compute. In our FT setting, the screener receives:

- the topic query view (`picos-lite2`, PIOS), and
- the candidate's title/abstract, plus a truncated full-text snippet (when available).

When section structure is available (e.g., PMC XML), we prioritize eligibility-relevant sections (e.g., METHODS/PARTICIPANTS/INTERVENTIONS/OUTCOMES) before truncation; otherwise we fall back to a fixed-prefix token budget. We release the preprocessing script and the exact truncation parameters to ensure reproducibility.

### J.5 Training and evaluation protocol (paired TA vs. FT)

**Goal and controlled comparison.** This auxiliary track is designed to measure *information gain* from full text while controlling confounds. We therefore compare:

**TA train → TA test** vs. **FT train → FT test**

on the same PMCID-eligible subset and under the same negative sampling strategy.

**Training instances.** We construct supervised training instances by pairing each FT-positive with sampled FT-negatives from the same topic. To control compute, we adopt a lightweight negative ratio (e.g., `n1`: one sampled FT-negative per FT-positive) consistent with the main-track negative sampling design. All FT-positives in the subset are retained; FT-negatives are sampled from the PMCID-negative pool.

**Evaluation instances.** On Tier 1 topics of the PMCID-eligible subset, we evaluate paired TA vs. FT on the same set of candidates (positives and negatives available in the subset). We report standard screening metrics (Sensitivity, Specificity, Precision, F1/F2, Accuracy). Unless otherwise noted, we compute *topic-level* metrics and macro-average across topics (consistent with the main Task 2 reporting) to avoid domination by topics with larger pools.

**Distribution-mismatch reference.** To illustrate that this OA-constrained slice can differ from the main training distribution, we additionally report a reference where a TA model trained on the full corpus is evaluated on the PMCID-eligible subset. This provides context for calibration shifts under distribution changes and should not be interpreted as a main-track performance comparison.

### J.6 Auxiliary full-text results

Table 31 reports the auxiliary FT-track results on the PMCID-eligible Tier 1 subset. FT yields consistent improvements over TA under the same training strategy, with notably higher specificity and accuracy, while maintaining high sensitivity. We reiterate that these results are OA-constrained and therefore reported as auxiliary evidence rather than headline benchmark conclusions.

**Interpretation and limitations.** The FT gains observed here should be interpreted as *information gain under OA constraints* on a PMCID-eligible slice. They do not imply that full-text screening is universally feasible at PubMed scale, nor that OA-only evaluation would be unbiased. Accordingly, we (i) keep TA as the main-track representation, (ii) expose PMCID metadata for transparency and reproducibility, and (iii) report the FT track as auxiliary analysis with explicit OA-skew caveats.

## K Venue/Citation-based Stratification Analysis

To further examine whether *source-review quality* can be approximated by simple bibliometric metadata, we analyze two optional topic-level proxies derived from the source meta-analysis records: **journal impact factor (IF)** and **citation count**. These signals are easy to obtain from venue metadata and were suggested as potential indicators of source-review reliability. However, unlike our coverage-based tiering—which is designed to control replay fidelity and benchmark interpretability—venue/citation filters may also alter the *difficulty distribution* of the benchmark by changing topic scale, maturity, and retrieval breadth.

**Why analyze citation counts both overall and by year?** Citation counts accumulate over time and are therefore strongly year-dependent. A 2023 meta-analysis has had substantially more time to accrue citations than a 2025 meta-analysis. For this reason, besides an **overall** citation stratification, we also perform **year-specific** citation stratification for 2023, 2024, and 2025, so that topics are compared against contemporaries with similar exposure time.

**Evaluation protocol.** We keep the **main training setup unchanged** and only stratify **Tier 1** topics at evaluation time. For Task 1, we report the performance of the best reranker, MEDEMBED_LITE2, on each subset. For Task 2, we report the performance of the best screener configuration, QWEN3-8B (N3+S1), at $N{=}500$. This design isolates the effect of topic stratification from training-size changes.

**Stratification bins.** For IF, we divide topics into three bins using the empirical quartile-like cutpoints observed in the data: **low** ($\leq 2.4$), **medium** ($(2.4, 4.5]$), and **high** ($> 4.5$). For citation count, we use: **overall**: low ($\leq 1$), medium ($(1, 11]$), high ($> 11$); **2023**: low ($\leq 5.5$), medium ($(5.5, 19]$), high ($> 19$); **2024**: low ($\leq 2$), medium ($(2, 10]$), high ($> 10$); **2025**: low ($= 0$), medium ($(0, 3.75]$), high ($> 3.75$).

**Main takeaway.** Contrary to our initial expectation, venue/citation stratification does *not* simply yield an "easier" or uniformly "higher-quality" subset. Instead, higher-IF and especially higher-citation topics tend to contain *more included studies* and *larger replayed candidate pools*, which makes both reranking and fixed-budget screening harder. This suggests that IF/citation act as **confounded proxies**: they may correlate with review prestige or maturity, but they also strongly encode topic scale and retrieval difficulty. Therefore, we do **not** use them as hard default filters for the main benchmark. Instead, we provide these fields as metadata for optional user-defined subgroup analysis.

### K.1 Distribution of IF and Citation Metadata

Table 32 summarizes the overall distribution of journal IF and citation counts. Both are strongly right-skewed, especially citation count. The year-wise statistics confirm the temporal accumulation effect for citations: the overall mean citation count is 8.07, but the mean is much higher for 2023 topics (15.30) than for 2025 topics (2.55).

Figure 7 compares the normalized IF and citation distributions between Tier 1 and Tier 2. The overall shapes are similar, indicating that coverage-based tiering does not induce a dramatic shift along these metadata dimensions.

### K.2 Topic Scale under IF/Citation Stratification

To understand whether venue/citation differences may reflect topic difficulty rather than source-review quality alone, Table 33 summarizes the number of included studies and replayed query results across bins.

| Setting (subset) | Sens. | Spec. | Prec. | F1 | F2 | Acc. |
|---|---|---|---|---|---|---|
| TA train → TA test (n1) | 0.9399 | 0.7393 | 0.6439 | 0.7642 | 0.8608 | 0.8063 |
| FT train → FT test (n1) | 0.9537 | 0.7627 | 0.6504 | 0.7733 | 0.8723 | 0.8231 |
| TA (full-train) → subset test (n1) | 0.9593 | 0.7125 | 0.6069 | 0.7435 | 0.8595 | 0.7905 |

Table 31: **Auxiliary full-text track (PMCID-eligible subset; Tier 1 evaluation).** We compare paired TA vs. full-text (FT) screening under the same training strategy (n1). The bottom row reports a distribution-mismatch reference: a TA model trained on the full corpus evaluated on the same subset. This track is OA-constrained and is not used as the primary benchmark result.

| Metric | Mean | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|
| Journal IF (overall) | 3.7964 | 0.6 | 2.4 | 3.0 | 4.4 | 88.5 |
| Citation count (overall) | 8.0684 | 0.0 | 1.0 | 4.0 | 11.0 | 99.0 |
| Citation count (2023) | 15.2993 | 1.0 | 5.5 | 11.0 | 19.0 | 99.0 |
| Citation count (2024) | 7.4830 | 0.0 | 2.0 | 5.0 | 10.0 | 49.0 |
| Citation count (2025) | 2.5517 | 0.0 | 0.0 | 1.0 | 3.75 | 24.0 |

Table 32: **Distribution of optional venue/citation metadata.** Journal impact factor (IF) and citation count are right-skewed. Because citation counts accumulate over time, we report both overall and year-specific statistics for 2023–2025.

Two patterns are consistent. First, higher-IF and higher-citation subsets tend to have **more included studies**, meaning that the model must recover a larger set of positives. Second, they also tend to have **larger replayed candidate pools**, especially for citation count. For example, in the overall citation stratification, the median query result size increases from 2117 in the low-citation subset to 5308 in the high-citation subset. This makes high-citation topics intrinsically harder under both top-$K$ retrieval and fixed-budget screening.

### K.3 Task 1: Reranking under IF/Citation Stratification

Table 34 reports Task 1 performance of MEDEMBED_LITE2 on Tier 1 subsets defined by IF/citation strata. For reference, we also include AUTHORPUBMED, RANDOM, and the overall Tier 1 result.

The main pattern is that higher-IF and especially higher-citation subsets generally exhibit **lower raw Recall/nDCG and higher absolute effort**, which is consistent with the larger candidate pools and larger positive sets shown in Table 33. For example, in the overall citation stratification, moving from low to high citation reduces Recall@100 from 0.8774 to 0.8433 and increases median Effort@0.95 from 39.5 to 70.0. However, after size normalization, the pattern becomes **weaker and less monotonic**, particularly for IF: the median NormEffort@0.95 is 0.2036 for high IF, 0.2548 for medium IF, and 0.1993 for low IF. Thus, venue/citation primarily affect the *scale* of the retrieval problem; they do not behave like a simple monotonic "quality" axis.

### K.4 Task 2: Fixed-Budget Screening under IF/Citation Stratification

Table 35 reports Task 2 results for QWEN3-8B (N3+S1) at $N$=500. The same general trend persists. Higher-IF and higher-citation subsets often show lower F1/F2 than lower-value subsets, again consistent with the fact that they correspond to broader, larger, and more demanding retrieval pools rather than simply "cleaner" source reviews.

For IF, the differences are modest but consistent: F2 decreases from 0.7084 in the low-IF subset to 0.6788 in the high-IF subset. For citation, the effect is stronger. Under overall citation stratification, F2 drops from 0.7129 in the low-citation subset to 0.6687 in the high-citation subset. The year-wise stratification shows a similar pattern in 2023 and 2024, while 2025 is somewhat weaker due to the compressed citation range in more recent topics.

Overall, these results indicate that venue/citation metadata should be interpreted as **difficulty-related and confounded descriptors**, not as clean stand-alone guarantees of source-review quality. Using them as hard default filters would therefore risk changing the benchmark itself, rather than merely removing noisy source reviews.

**Implication for benchmark design.** Taken together, the IF/citation analyses show that these metadata dimensions are **not benign hard filters**. Applying them as default inclusion rules would substantially alter topic scale and task difficulty, thereby trading the representativeness of the PubMed-universe benchmark for a smaller and systematically shifted subset. For this reason, we keep

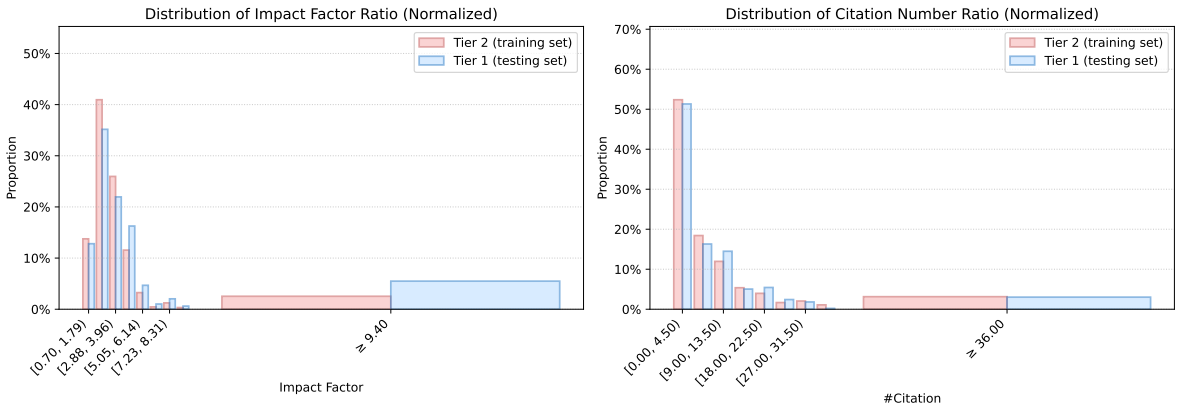



Figure 7: **Normalized distributions of journal IF and citation count for Tier 1 and Tier 2.** Left: IF distribution of the source meta-analysis journals. Right: citation count distribution of the source meta-analyses. The two tiers remain broadly comparable on these metadata dimensions.

the main benchmark setup unchanged and release venue/citation metadata for optional downstream filtering and subgroup analysis.

## L Experimental Setup and Hyperparameters

To improve reproducibility, we report the main training, inference, and hardware settings used for Task 1 reranker fine-tuning (Table 36) and Task 2 screener fine-tuning (Table 37). Unless otherwise stated, all reported models are trained on Tier 2 and evaluated on Tier 1.

## M PubMed MeSH Tree-Number Distribution

To characterize the topical composition of the dataset, we mapped each topic to the first-letter prefixes of its associated MeSH tree numbers, which correspond to the top-level MeSH branches. For each topic, duplicate prefixes were removed so that a topic contributed at most once to each branch, even if multiple MeSH terms (or multiple tree numbers of the same term) fell under the same top-level category. We then aggregated these topic-level branch assignments within each split (Tier 1 vs. Tier 2). Thus, if a topic was associated with both branches `B` and `C`, it contributed one count to `B` and one count to `C` in the global tally. The plotted proportions were obtained by normalizing these aggregated counts within each split. The distribution comparison of Tier 1 and Tier 2 on MeSH tree numbers is shown in Figure 14.

In MeSH, a tree number locates a descriptor within a hierarchical subject tree: the initial letter indicates the top-level branch, and the following numeric segments indicate increasingly specific sub-branches. In our data, the observed initials correspond to the following MeSH branches:

- `A`: Anatomy
- `B`: Organisms
- `C`: Diseases
- `D`: Chemicals and Drugs
- `E`: Analytical, Diagnostic and Therapeutic Techniques and Equipment
- `F`: Psychiatry and Psychology
- `G`: Phenomena and Processes
- `H`: Disciplines and Occupations
- `I`: Anthropology, Education, Sociology and Social Phenomena
- `J`: Technology, Industry and Agriculture
- `K`: Humanities
- `L`: Information Science
- `M`: Named Groups
- `N`: Health Care
- `Z`: Geographic Locations

| Group | Stat | Included studies size | | | Query result size | | |
|---|---|---|---|---|---|---|---|
| | | High | Medium | Low | High | Medium | Low |
| **IF** | Mean | 15.56 | 13.30 | 12.15 | 9936.98 | 9495.50 | 11532.97 |
| | 25% | 8.00 | 7.00 | 6.00 | 1539.75 | 806.75 | 1009.75 |
| | 50% | 12.00 | 10.00 | 10.50 | 4737.50 | 2667.00 | 2421.00 |
| | 75% | 18.00 | 16.00 | 15.75 | 10284.75 | 7409.25 | 9496.25 |
| **Citation (overall)** | Mean | 15.78 | 12.85 | 12.91 | 14709.76 | 9444.31 | 7190.98 |
| | 25% | 9.00 | 7.00 | 6.00 | 1632.00 | 1000.75 | 828.00 |
| | 50% | 13.00 | 10.00 | 10.00 | 5308.00 | 2659.50 | 2117.00 |
| | 75% | 19.00 | 16.00 | 16.25 | 13001.00 | 8983.25 | 6375.25 |
| **Citation (2023)** | Mean | 15.23 | 13.88 | 10.60 | 15657.32 | 12313.92 | 6294.89 |
| | 25% | 8.00 | 8.00 | 6.00 | 1674.00 | 1092.00 | 737.00 |
| | 50% | 13.00 | 11.00 | 8.00 | 6345.00 | 4692.00 | 1464.00 |
| | 75% | 20.00 | 17.00 | 14.00 | 12692.00 | 12369.00 | 5583.00 |
| **Citation (2024)** | Mean | 16.85 | 13.25 | 10.28 | 15201.83 | 10108.27 | 6332.72 |
| | 25% | 8.00 | 7.00 | 5.25 | 2511.00 | 1180.50 | 445.50 |
| | 50% | 13.50 | 11.00 | 8.00 | 5395.00 | 2631.00 | 983.50 |
| | 75% | 20.25 | 17.00 | 11.75 | 12301.00 | 9567.50 | 3598.25 |
| **Citation (2025)** | Mean | 16.16 | 14.28 | 12.09 | 9794.86 | 10117.99 | 6194.29 |
| | 25% | 8.50 | 7.00 | 7.00 | 2000.25 | 948.00 | 918.00 |
| | 50% | 12.00 | 10.00 | 10.00 | 4662.00 | 3000.00 | 2827.00 |
| | 75% | 17.00 | 17.00 | 16.75 | 8659.50 | 6332.00 | 6966.75 |

Table 33: **Topic scale statistics under IF/citation stratification.** For each IF/citation grouping, we report summary statistics of the number of included studies and the replayed query result size. Higher-value IF/citation subsets tend to contain more included studies and, especially for citation, substantially larger candidate pools, indicating that venue/citation stratification changes topic difficulty and retrieval scale rather than acting as a clean quality-only filter. For IF, High / Medium / Low correspond to $> 4.5$, $(2.4, 4.5]$, and $\leq 2.4$, respectively. For citation count, the cutpoints are defined separately for overall / 2023 / 2024 / 2025 as described in the text.

| Model / Subset | Recall ↑ | | | nDCG ↑ | Effort@0.95 ↓ | | NormEffort@0.95 ↓ | |
|---|---|---|---|---|---|---|---|---|
| | @100 | @200 | @500 | @200 | median [IQR] | p95 | median [IQR] | p95 |
| Author-PubMed | 0.5048 | 0.6438 | 0.7977 | 0.3069 | 194.0 [60.5, 536.5] | 2859.0 | 0.80 [0.63, 0.92] | 0.9919 |
| Random | 0.4717 | 0.6081 | 0.7683 | 0.2945 | 219.0 [76.0, 693.5] | 4147.5 | 0.94 [0.88, 0.97] | 1.0000 |
| ME_LITE2 (overall) | 0.8730 | 0.9148 | 0.9371 | 0.7404 | 43.0 [18.0, 114.0] | 462.0 | 0.23 [0.10, 0.40] | 0.7113 |
| High IF | 0.8669 | 0.9115 | 0.9399 | 0.7362 | 47.0 [22.0, 132.75] | 513.5 | 0.20 [0.09, 0.36] | 0.6837 |
| Medium IF | 0.8687 | 0.9098 | 0.9311 | 0.7350 | 41.5 [18.0, 105.0] | 525.4 | 0.25 [0.11, 0.45] | 0.7202 |
| Low IF | 0.8874 | 0.9289 | 0.9460 | 0.7561 | 51.5 [15.25, 112.75] | 315.75 | 0.20 [0.09, 0.39] | 0.7143 |
| High citation, overall | 0.8433 | 0.9045 | 0.9361 | 0.7142 | 70.0 [29.0, 151.0] | 552.2 | 0.23 [0.07, 0.43] | 0.7206 |
| Medium citation, overall | 0.8834 | 0.9197 | 0.9375 | 0.7417 | 37.0 [17.25, 105.25] | 420.25 | 0.22 [0.10, 0.38] | 0.6466 |
| Low citation, overall | 0.8774 | 0.9133 | 0.9365 | 0.7603 | 39.5 [12.75, 89.25] | 381.9 | 0.23 [0.11, 0.48] | 0.7877 |
| High citation, 2023 | 0.8287 | 0.8923 | 0.9264 | 0.6868 | 81.0 [29.0, 170.0] | 1435.0 | 0.23 [0.09, 0.47] | 0.7828 |
| Medium citation, 2023 | 0.8680 | 0.9272 | 0.9499 | 0.7264 | 41.0 [20.0, 137.0] | 387.6 | 0.20 [0.07, 0.39] | 0.6583 |
| Low citation, 2023 | 0.9116 | 0.9359 | 0.9462 | 0.7666 | 31.0 [11.0, 60.5] | 259.3 | 0.22 [0.10, 0.37] | 0.5043 |
| High citation, 2024 | 0.8539 | 0.9090 | 0.9400 | 0.7156 | 71.5 [30.25, 155.75] | 544.8 | 0.19 [0.05, 0.38] | 0.6737 |
| Medium citation, 2024 | 0.8746 | 0.9150 | 0.9321 | 0.7305 | 42.0 [24.0, 115.5] | 651.1 | 0.25 [0.12, 0.44] | 0.7005 |
| Low citation, 2024 | 0.8995 | 0.9185 | 0.9289 | 0.8031 | 20.0 [8.25, 49.75] | 203.85 | 0.23 [0.13, 0.44] | 0.7116 |
| High citation, 2025 | 0.8567 | 0.9062 | 0.9367 | 0.7278 | 57.5 [23.75, 118.5] | 508.3 | 0.22 [0.11, 0.36] | 0.6020 |
| Medium citation, 2025 | 0.8799 | 0.9080 | 0.9339 | 0.7577 | 32.0 [14.0, 85.0] | 581.6 | 0.22 [0.07, 0.40] | 0.8418 |
| Low citation, 2025 | 0.8681 | 0.9103 | 0.9353 | 0.7364 | 51.0 [16.25, 97.5] | 338.9 | 0.25 [0.15, 0.48] | 0.7471 |

Table 34: **Task 1 performance under IF/citation stratification.** All subset results are evaluated on Tier 1 using the same globally trained reranker (MEDEMBED_LITE2). Higher-IF and especially higher-citation subsets tend to exhibit lower raw retrieval performance and higher absolute effort, largely because they contain more positives and larger replayed candidate pools.

| Training Strategy / Subset | Sensitivity | Sensitivity (e2e) | Specificity | Precision | F1 | F2 |
|---|---|---|---|---|---|---|
| Qwen3-8B (base) | 0.3628 | 0.3608 | 0.8994 | 0.2714 | 0.2640 | 0.2962 |
| Qwen3-8B (n3+s1, overall) | 0.8564 | 0.8504 | 0.8966 | 0.4902 | 0.5823 | 0.6932 |
| High IF ($> 4.5$) | 0.8298 | 0.8220 | 0.9139 | 0.4861 | 0.5727 | 0.6788 |
| Medium IF ($(2.4, 4.5]$) | 0.8598 | 0.8530 | 0.8796 | 0.4874 | 0.5796 | 0.6910 |
| Low IF ($\leq 2.4$) | 0.8714 | 0.8696 | 0.9111 | 0.4983 | 0.5945 | 0.7084 |
| High citation ($> 11$), overall | 0.8470 | 0.8334 | 0.8879 | 0.4485 | 0.5499 | 0.6687 |
| Medium citation ($(1, 11]$), overall | 0.8636 | 0.8586 | 0.8993 | 0.4912 | 0.5829 | 0.6947 |
| Low citation ($\leq 1$), overall | 0.8515 | 0.8496 | 0.9022 | 0.5266 | 0.6112 | 0.7129 |
| High citation ($> 19$), 2023 | 0.8630 | 0.8318 | 0.8682 | 0.4205 | 0.5261 | 0.6516 |
| Medium citation ($(5.5, 19]$), 2023 | 0.8762 | 0.8733 | 0.9021 | 0.4694 | 0.5798 | 0.7043 |
| Low citation ($\leq 5.5$), 2023 | 0.8411 | 0.8375 | 0.9216 | 0.5397 | 0.5993 | 0.6855 |
| High citation ($> 10$), 2024 | 0.8229 | 0.8146 | 0.9105 | 0.4599 | 0.5504 | 0.6625 |
| Medium citation ($(2, 10]$), 2024 | 0.8403 | 0.8336 | 0.8910 | 0.4727 | 0.5637 | 0.6750 |
| Low citation ($\leq 2$), 2024 | 0.8567 | 0.8560 | 0.8959 | 0.5905 | 0.6651 | 0.7505 |
| High citation ($> 3.75$), 2025 | 0.8672 | 0.8614 | 0.8902 | 0.4850 | 0.5686 | 0.6758 |
| Medium citation ($(0, 3.75]$), 2025 | 0.8674 | 0.8628 | 0.8933 | 0.4858 | 0.5852 | 0.7066 |
| Low citation ($= 0$), 2025 | 0.8617 | 0.8586 | 0.9038 | 0.4970 | 0.5926 | 0.7024 |

Table 35: **Task 2 performance under IF/citation stratification ($N$=500).** All subset results are evaluated on Tier 1 using the same globally trained screening model (QWEN3-8B (N3+S1)). Higher-value IF/citation bins are not uniformly easier; instead they often correspond to lower F1/F2 because these topics are associated with larger candidate pools and more positives to recover.

| Item | Setting |
|---|---|
| Fine-tuning type | Full dense-encoder fine-tuning |
| Input | Query view + candidate title/abstract |
| Output | Dense similarity score |
| Loss | MultipleNegativesRankingLoss |
| Epochs | 1 |
| Batch size | 8 |
| Optimizer | AdamW, SentenceTransformers default |
| Learning rate | $2 \times 10^{-5}$, SentenceTransformers default |
| Weight decay | 0.01, SentenceTransformers default |
| Scheduler / warmup | WarmupLinear; warmup steps = 10% of dataloader steps |
| Max sequence length | Not manually overridden; each backbone uses its default max sequence length |
| Mixed precision | AMP enabled |
| Seed | 42 |
| Hardware | vGPU, 32GB memory |

Table 36: **Task 1 reranker training setup.** Dense rerankers are fine-tuned on Tier 2 and evaluated on Tier 1 under the candidate-pool protocol described in Section 5.

| Item | Setting |
|---|---|
| Fine-tuning type | LoRA SFT |
| Input | `picos-lite2` + candidate title/abstract |
| Output | JSON include/exclude decision |
| LoRA rank / alpha / dropout | 16 / 32 / 0.05 |
| LoRA target modules | q, k, v, o projections; gate/up/down projections |
| QLoRA / 4-bit | No |
| Epochs | 1 |
| Train batch size | 2 |
| Gradient accumulation | 8 |
| Effective batch size | 16 |
| Optimizer | AdamW |
| Learning rate | $2 \times 10^{-4}$ |
| Scheduler / warmup | Linear warmup/decay; warmup ratio 0.03 |
| Max input length | 2048 tokens |
| Precision | bfloat16 when CUDA is available |
| Gradient checkpointing | Enabled |
| Seed | 42 |
| Inference | Greedy decoding; max 32 new tokens |
| Hardware | RTX PRO 6000-class GPU, 98GB memory |

Table 37: **Task 2 screener fine-tuning and inference setup.** LLM screeners are fine-tuned on Tier 2 and evaluated on reranker-produced Tier 1 Top-$N$ pools.

Query Reconstruction Prompt Template

Part 1

```
You are an expert medical librarian and information specialist with extensive experience
designing high-quality PubMed search strategies for systematic reviews and meta-analyses.

I will give you:
- The TITLE of a meta-analysis
- The ABSTRACT of the meta-analysis
- The QUERY SECTION: any descriptive text about the original literature search
  (which may include search concepts or sample strategies for various databases
   such as MEDLINE, Embase, CENTRAL, Web of Science, trial registries, etc.).

Your task is to design a single, well-structured **PubMed core search query** that captures
the main clinical question of the meta-analysis.

IMPORTANT SCOPE:
- You are ONLY asked to produce the **core topical query** for PubMed:
  - Population / participants
  - Intervention or exposure
  - Comparator (if relevant)
  - Outcomes (if they are central to the concept)
  - Study design filters ONLY if they are part of the meta-analysis' s inclusion criteria
    (e.g., randomized controlled trials), and only if they are clearly indicated.
- You MUST NOT include:
  - Date limits (no [dp], [edat], or explicit year/date ranges).
  - Publication type limits (no Review[pt], Meta-Analysis[pt], Clinical Trial[pt], etc.).
  - Generic article-type exclusion blocks (e.g., NOT (Review[pt] OR Editorial[pt] ...)).

--------------------
INPUT
--------------------
META-ANALYSIS TITLE:
{title}

META-ANALYSIS ABSTRACT:
{abstract}

QUERY SECTION (may contain search descriptions for multiple databases and noisy formatting):
{query_section}

--------------------
DESIGN RULES FOR THE PUBMED CORE QUERY
--------------------
1. Focus on the *intended* meta-analysis question:
   - Use the title and abstract to understand the main PICOS elements.
   - Use the query section only as auxiliary information (e.g., synonyms, key MeSH terms,
     population definitions), but do NOT blindly copy Ovid/Embase/Web of Science syntax.

...
```

Part 2

```
...

2. Use proper PubMed syntax:
   - Combine controlled vocabulary (MeSH terms) and free-text terms (Title/Abstract).
   - Typical patterns:
     - "Disease"[Mesh] OR disease*[tiab]
     - "Treatment Name"[Mesh] OR treatment*[tiab]
   - Use field tags like [Mesh], [MeSH Terms], [tiab], [Title/Abstract], [mh], etc.
   - Use AND to combine different concept blocks (e.g., population AND intervention).
   - Use OR to combine synonyms within the same concept.
   - Use parentheses to group concepts correctly.

3. You MAY:
   - Add obvious synonyms or variants that are strongly supported by the title/abstract/query section.
   - Use MeSH headings that clearly match the described concepts.
   - Include an RCT / clinical trial filter ONLY IF the meta-analysis explicitly focuses on randomized
     or interventional trials. If not clearly required, do NOT add a design filter.
   - Add humans/animals filters unless explicitly demanded by the meta-analysis question.

4. You MUST NOT:
   - Add any date restrictions (no date ranges, no [dp] or [edat] filters).
   - Add any publication type filters (no [pt], [ptyp] blocks like Review[pt], Meta-Analysis[pt], etc.).

5. Formatting and style:
   - The final query must be a **single-line** string that is syntactically valid in PubMed.
   - Do NOT include newline characters in the query.
   - Use straight double quotes "..." only (no curly quotes).
   - Parentheses must be balanced.
   - Keep wildcards * where appropriate (e.g., pregnan*, diabet*, trial*), as long as they
     are valid for PubMed.
   - Keep the query reasonably compact and focused: do not explode into hundreds of terms
     unless clearly implied by the input.

--------------------
OUTPUT FORMAT
--------------------
Return ONLY a single JSON object with the following fields:
- "pubmed_query_core":
  - A single-line, syntactically correct PubMed core query string, WITHOUT any date limits,
    publication type limits.
- "confidence":
  - Your confidence that "pubmed_query_core" correctly reflects the intended meta-analysis
question.

Do NOT wrap the JSON in markdown.
Do NOT add any extra text before or after the JSON.
```

Figure 8: Query Reconstruction Prompt Template

**Task 2 (Inclusion Decision) Prompt Template**

You are assisting with study selection for a systematic review.

Screening target (meta-analysis question):
{meta_block}

{study_block}

Task: Based on the meta-analysis question and PICOS criteria above, decide whether this candidate study should be INCLUDED or EXCLUDED from the review.

Output strictly in JSON format, with one of:
{"decision": "include"}
{"decision": "exclude"}

Figure 9: Task 2 Prompt Template

## PICOS-based Hard Negatives Generation Prompt Template

You are an expert clinical epidemiologist and trialist.

Goal:
Create ONE synthetic *hard negative* study in English for a given meta-analysis. The study must look realistic and stylistically similar to the real included study, but it MUST VIOLATE EXACTLY ONE PICOS ELEMENT of the original meta-analysis.

META-ANALYSIS PICOS & ELIGIBILITY
{META_PICOS}

STYLE TEMPLATE (REAL INCLUDED STUDY)
TITLE:
{TEMPLATE_TITLE}

ABSTRACT:
{TEMPLATE_ABSTRACT}

MUTATION TARGET
You must modify ONLY this PICOS element:
- Target element: {MUTATION_TARGET_LABEL} (code: {MUTATION_TARGET_CODE})

Apply a *small but important* change so that the new study would NOT satisfy the original meta-analysis inclusion criteria for this element, while all other PICOS elements remain compatible with the meta-analysis.

Mutation guidance for this element:
{MUTATION_HINTS}

GLOBAL CONSTRAINTS
1. Change EXACTLY ONE PICOS element (the target above). All other PICOS parts must remain compatible with the original meta-analysis.
2. The new study must remain medically realistic and clinically plausible.
  - Do NOT switch from humans to animals or in vitro models.
  - Keep overall disease area, clinical context, and general topic consistent.
3. Do NOT reuse sentences from the real study's title or abstract. Write new text in a similar style and level of detail.
4. Do NOT mention any real PMID, trial registration number, hospital name, or real institution.
5. The abstract should follow an IMRaD-like structure with explicit or implicit sections:
  Background, Methods, Results, Conclusion.
  - Include realistic sample size and numeric results (effect sizes, confidence intervals, p-values, follow-up duration, etc.).
6. The final synthetic study MUST be clearly EXCLUDED by a careful human screener
  when checking against the original meta-analysis PICOS (because of the single mutated element).

Figure 10: PICOS-based Hard Negatives Generation Prompt Template

## Guidelines & Criteria that used to filter Meta-Analyses

1. is_meta_analysis: one of ***yes/no/uncertain***
 - "yes" if the title/abstract clearly indicates a meta-analysis or pooled quantitative synthesis of primary studies.
 - "no" if it is clearly not a meta-analysis (narrative review, scoping review, single trial, cohort study, etc).
 - "uncertain" if not sure.

2. is_umbrella_or_meta_meta: one of ***yes/no/uncertain***
 - "yes" if it is an umbrella review, overview of systematic reviews, meta-analysis of meta-analyses, or similar.
 - "no" if it is a meta-analysis of primary studies.
 - "uncertain" if unclear.

3. is_human_clinical: one of ***yes/no/uncertain***
 - "yes" if the main evidence clearly involves human patients/participants.
 - "no" if it is mainly animal, in vitro, or purely non-clinical research.
 - "uncertain" if unclear.

4. primary_study_design: one of ***trials/observational/mixed/non_clinical/uncertain***
 - "trials": mainly randomized controlled trials or other interventional clinical trials.
 - "observational": mainly cohort, case-control, cross-sectional, registry-based, or other non-interventional observational designs.
 - "mixed": clearly both trials and observational studies are included.
 - "non_clinical": animal or preclinical studies only.
 - "uncertain": if unclear.

5. is_english: one of ***yes/no/uncertain***
 - "yes" if the abstract is in English.
 - "no" if the abstract is clearly not in English.
 - "uncertain" if unclear.

6. main_focus: one of ***intervention/diagnostic/prognostic/risk_factor/other/uncertain***
 - "intervention": effect of a treatment or intervention.
 - "diagnostic": diagnostic test accuracy.
 - "prognostic": prediction of prognosis or outcomes.
 - "risk_factor": association between exposures and outcomes.
 - "other": anything else (e.g., prevalence, incidence, methodology).
 - "uncertain": if unclear.

7. is_eligible_for_meta_benchmark: one of ***yes/no/uncertain***
 - Set to "yes" ONLY IF ALL of the following are true:
  * is_meta_analysis == "yes"
  * is_umbrella_or_meta_meta == "no"
  * is_human_clinical == "yes"
  * primary_study_design is "trials" or "observational" or "mixed"
  * is_english == "yes"
 - Set to "no" if it clearly does NOT meet these conditions.
 - Otherwise set to "uncertain".

Figure 11: Meta-Analyses' inclusion guideline and criteria

PICOS Card Annotation Guideline (for Medical Annotators)

Part 1

**Goal:** From the **full text** of a systematic review/meta-analysis and its **section headings**, produce a standardized **PICOS card** for downstream **automated retrieval** and **study screening**, with **zero leakage of results**.
**Deliverable:** PICOS-card fields only: design_type / study_type / population / intervention / exposure / comparator / outcomes / setting / timeframe / additional_criteria.

## 1. Core Principles & Source Priority

### 1.1 Absolute rule: No results leakage

You **must not** use or paraphrase content from these sections to construct PICOS (even if visible in the full text):
**Results / Findings / Synthesis / Meta-analysis results**

- **Discussion / Conclusion / Implications**
- **Abstract** results/conclusion sentences (you may use objective/methods-style lines, but avoid outcome statements)
- **Forbidden content includes:**

Any numbers (effect sizes, CIs, p-values, sensitivity/specificity values, counts that imply findings)

- Any conclusion language ( “effective” , “improved” , “high accuracy” , “significant” , “beneficial” )
- Any description of direction/magnitude of findings
- Allowed: only **methods-level intended eligibility**, as described in Methods/Eligibility sections.

### 1.2 Source priority (highest to lowest)

**Eligibility criteria / Inclusion & Exclusion / Study selection** (primary)

1. **Methods** (search strategy, study design, outcomes, participants, interventions/exposures)
2. **Introduction/Objectives** (only to clarify the research question/population; no inference)
3. Other sections only as secondary support (e.g., Supplementary Methods)
4. If a field is not explicitly supported by Methods/Eligibility, leave it as "" rather than guessing.

## 2. PICOS Card Fields: Definitions & Writing Rules

You must fill the following fields (aligned with your LLM prompt schema):

### 2.1 design_type (required; choose one)

"trials"

- : randomized or non-randomized interventional studies (RCTs, cluster RCTs, crossover trials, etc.)
- "observational"
- : cohort/case-control/cross-sectional/diagnostic accuracy, etc.
- "mixed"
- : explicitly includes both interventional and observational designs
- "uncertain"
- : cannot be determined from Methods/Eligibility (use only when truly unavoidable)
- **Heuristics:**

Mentions of “randomized” , “trial” , “sham” , “placebo” , allocation → typically "trials"

- “cohort/case-control/cross-sectional/observational/diagnostic accuracy” → typically "observational"
- Diagnostic accuracy meta-analyses are usually "observational" unless a trial design is explicit.

### 2.2 study_type (required; short phrase)

A concise description of included study types, e.g.:
“randomized controlled trials”

- “observational diagnostic accuracy studies”
- “prospective cohort and case-control studies”
- “mixed randomized and observational studies”
- Keep it short and method-level. No results language.

### 2.3 population (required)

Describe the target population using Methods/Eligibility language:
Condition/disease/state (e.g., AD/MCI; menopausal women with insomnia)

- Age/sex if specified
- Diagnostic criteria if specified
- Explicit population exclusions if stated
- If unclear, be conservative. Do not add assumptions.

Part 2

### 2.4 intervention vs exposure (mutually primary)

- If design_type="trials": populate intervention, set exposure to ""
- If design_type="observational": populate exposure, set intervention to ""

**Diagnostic accuracy meta-analyses:**

- The “index test” (measurement method/biomarker assessment) generally belongs under **exposure**
- Use outcomes to specify the type of diagnostic accuracy data needed (e.g., TP/FP/TN/FN constructable)

### 2.5 comparator (required, but may be "")

Report the comparator as stated:

- Trials: sham/placebo/usual care/active control
- Observational: unexposed group, different exposure levels, healthy controls, non-cases, alternative diagnostic categories

If no comparator is specified, use "" (avoid inventing).

### 2.6 outcomes (required)

Only outcomes required for inclusion, in methods terms:

- Trials: validated scales (PSQI/ISI) or “validated sleep quality measures”
- Diagnostic: data enabling a 2×2 table (TP/FP/TN/FN) or “diagnostic accuracy outcomes”
- Observational: incidence, mortality, recurrence, etc. (as specified)

Never include outcome direction or effect.

### 2.7 setting (optional; default "")

Only if explicitly stated:

- inpatient/outpatient/community, ICU/primary care, etc.

### 2.8 timeframe (optional; default "")

Only if explicitly stated:

- minimum follow-up, assessment window, perioperative period, etc.

### 2.9 additional_criteria (optional but commonly needed)

Capture other inclusion/exclusion restrictions stated in Methods/Eligibility:

- language limits (e.g., English/Persian)
- publication type (full-text only; exclude reviews/case reports)
- human-only restriction
- data availability requirements (extractable data; diagnostic 2×2)
- date restrictions (or “no date restriction” if explicitly stated)

Do **not** include databases/software unless they directly impose an inclusion restriction.

## 3. Annotation Workflow (SOP)

### Step 1 — Locate Methods-level eligibility content

Use headings to jump to:

- “Eligibility criteria / Inclusion and exclusion criteria”
- “Study selection”
- “Methods”
- “Search strategy”
- “Outcomes”
- “Participants / Interventions”

### Step 2 — Extract “intended eligibility”, not results

PICOS describes what studies were *eligible*, not what the authors ultimately found.

### Step 3 — Fill PICOS in a fixed order

1. design_type
2. + study_type
3. population
4. intervention
5. or exposure
6. comparator
7. outcomes
8. setting
9. , timeframe, additional_criteria

### Step 4 — Consistency & leakage check

Verify:

- No results wording or numeric findings
- design_type
- matches intervention/exposure usage
- Comparator/outcomes are supported by Methods/Eligibility text, not inferred

Figure 12: PICOS card Annotation Guideline Part1/2.

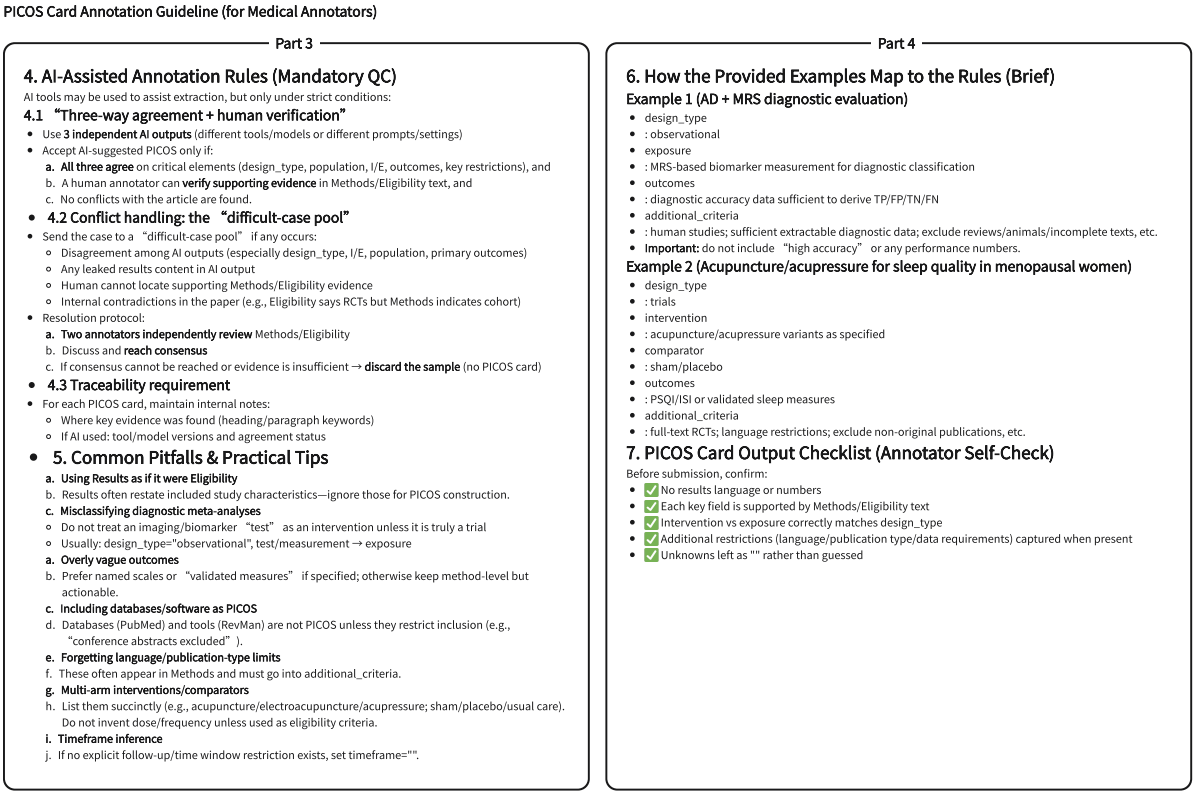



Figure 13: PICOS card Annotation Guideline Part3/4.

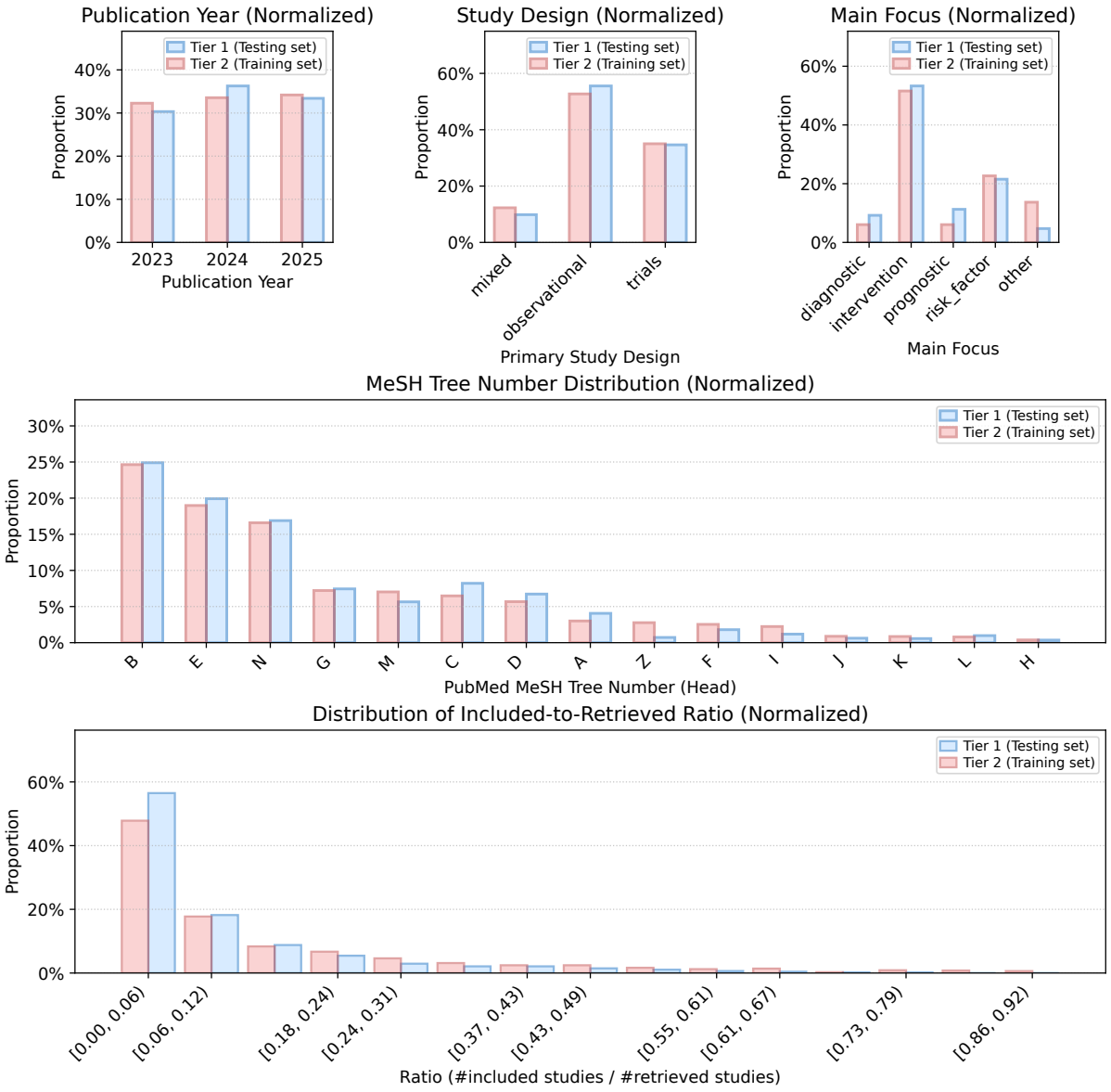



Figure 14: Dataset Statistics Comparison of Tier 1/2 Across Five Dimensions