

RECITE: Revising Eligibility Criteria Incorporating Textual Evidence

Russell Ro^{1,2}, Reza Abbasi-Asl^{1*}

¹University of California, San Francisco ²University of California, Berkeley

ABSTRACT

Overly restrictive eligibility criteria contribute to widespread enrollment failure in clinical trials, yet published recommendations for how specific criteria should be broadened remain disconnected from trial design at scale. We introduce RECITE (Revising Eligibility Criteria Incorporating Textual Evidence), a framework for automated discovery, extraction, and implementation of evidence-based eligibility criteria revisions. RECITE uses protocol amendment documents from ClinicalTrials.gov to construct a dataset of 11,913 curated instances across 5,735 trials, each pairing original criteria with documentary evidence and target revisions. We benchmark fifteen LLM configurations under truncated-to-context-window and retrieval-augmented generation conditions, on a 3,116-instance sample under an LLM-as-judge protocol validated against ten domain experts (judge vs. human quadratic-weighted Cohen’s $\kappa_{qw} = 0.694 \pm 0.096$). The best open-weight model, Gemma 2 9B, achieves κ_{qw} -aligned ordinal score 3.0 ± 0.9 on a 0–4 scale (normalized 0.742) and the best closed-weight model, GPT-4o, reaches 3.0 ± 0.8 (normalized 0.745). Using an agentic literature discovery system, we then identified 44 paper–trial matches across five publications providing trial-design and EC-modification evidence. Our findings suggest that existing clinical knowledge can be operationalized into actionable trial design recommendations.

Keywords: Clinical trials, Eligibility criteria, Large language models, Retrieval-augmented generation

An earlier version of this work appears in the Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’26), <https://doi.org/10.1145/3770855.3819027>.

Resource Availability: The source code and benchmark data are publicly available at <https://doi.org/10.5281/zenodo.20487342>. The source repository is hosted at <https://github.com/abbasilab/RECITE>.

1 Introduction

Clinical trials are the primary mechanism for evaluating medical interventions, yet enrollment remains a persistent bottleneck. Only 31% of UK Medical Research Council trials met their original enrollment goals [4], one-third of publicly funded trials required time extensions due to recruitment shortfalls [5], and 25% of cancer trials failed to enroll a sufficient number of participants [10]. Insufficient accrual is the single largest cause of trial termination, accounting for 36.7% of all stoppages in a recent analysis of 28,561 discontinued trials [40]. When enrollment shortfalls trigger protocol amendments, the costs compound: 76% of protocols undergo at least one amendment, and each Phase III amendment costs a median of \$535,000 [15, 16]. Overly restrictive eligibility criteria are widely recognized as a contributing factor [32].

A growing body of work already articulates explicit recommendations for how specific eligibility criteria should be loosened. Governing body directives from the American Society of Clinical Oncology and Friends of Cancer Research (ASCO-Friends) recommend inclusion of patients with controlled brain metastases, prior malignancies, and reduced creatinine clearance thresholds [25]. The U.S. Food and Drug Administration has issued cross-cutting guidance encouraging sponsors to broaden criteria, justify each exclusion, and include medically complex participants [41].

*Corresponding author: reza.abbasiasl@ucsf.edu

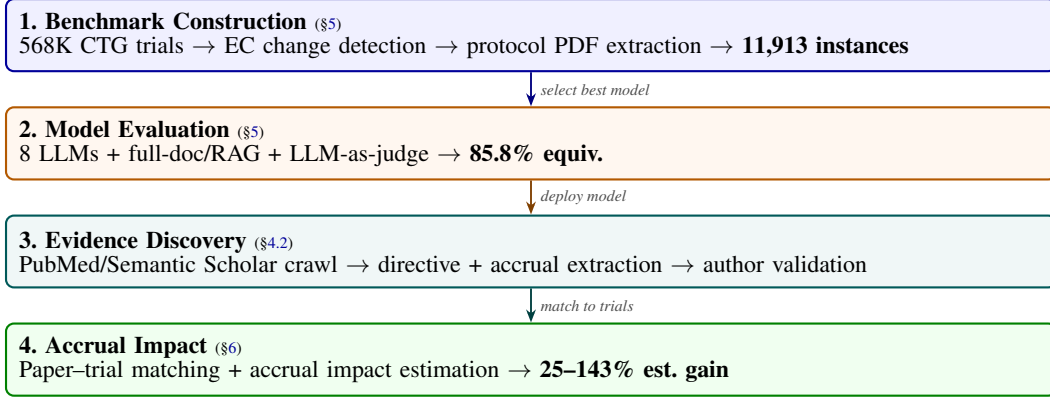


Figure 1: RECITE framework overview. Four stages: benchmark construction from ClinicalTrials.gov protocols (§5), model evaluation via LLM-as-judge, evidence discovery from published literature (§4.2), and estimated accrual impact (§6).

Published research provides disease-specific evidence across therapeutic areas including oncology [24, 32], cardiology [47], neurology [28], psychiatry [49], rheumatology [9], and nephrology [6], supported by consensus directives from KDIGO [1] and the Global Cardiovascular Clinical Trialists Forum [11].

Despite this wealth of guidance, no existing system automates the discovery, extraction, and implementation of such recommendations across trials at scale. Individual trialists may consult specific publications when designing or amending protocols, but systematically connecting the large body of published recommendations to active trial designs remains an unsolved technical problem. The gap is not one of clinical knowledge (the recommendations exist) but of infrastructure: the retrieval and matching between two large bodies of text, published evidence and registered trials, exceeds what manual processes can achieve.

Recent computational work has addressed adjacent problems, including amendment prediction [7] and criteria generation from trial descriptions [30, 43], but none automates evidence-grounded revision (Section 2).

To address this gap, we introduce RECITE (**R**evising **E**ligibility **C**riteria **I**ncorporating **T**extual **E**vidence), a framework for evidence-guided eligibility criteria revision. Given original eligibility criteria and supporting documentary evidence, whether from protocol amendment documents, published research papers, or clinical guidelines, RECITE extracts explicit modification directives and generates revised criteria with cited provenance. The framework operates across diverse evidence sources and handles the variation in structure, length, and style typical of real-world clinical documents.

Our contributions are as follows:

1. We formulate the task of *evidence-guided eligibility criteria revision*, defining it over diverse documentary sources including protocol amendment documents, published literature, and clinical guidelines.
2. We develop an agentic literature discovery system that crawls PubMed, matches papers to active trials, and extracts actionable modification directives.
3. We construct the RECITE benchmark from 567,815 trials on ClinicalTrials.gov, filtering through version histories and protocol evidence to yield 11,913 instances across 5,735 unique trials. We evaluate on a sample of 3,116 instances with original criteria, grounding evidence from protocol amendment documents, and target revisions.
4. We demonstrate real-world applicability: 44 paper-trial matches across five publications surface evidence-grounded EC-broadening directives. The results show that existing clinical knowledge can be operationalized into evidence-based trial design (with directional, not numeric, accrual claims).

2 Background and Related Work

2.1 Eligibility Criteria Design and the Case for Liberalization

Eligibility criteria balance participant safety with scientific validity in clinical trials. Evidence across therapeutic areas suggests that criteria more restrictive than necessary can reduce enrollment, limit demographic diversity, and

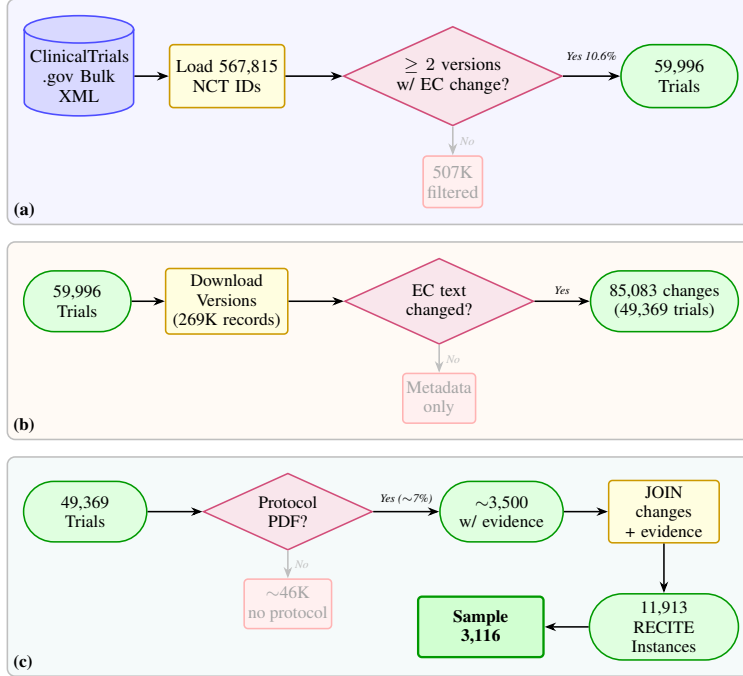


Figure 2: RECITE benchmark construction pipeline. **(a)** Discovery and version filtering. **(b)** EC change detection across consecutive protocol versions. **(c)** Evidence extraction from protocol PDFs and benchmark assembly.

narrow the generalizability of findings. Oncology has produced the largest evidence base for reform. The ASCO-Friends consensus statement provides concrete recommendations for broadening criteria related to brain metastases, prior malignancies, and organ function thresholds [25], while Trial Pathfinder demonstrated that data-driven relaxation of common exclusions does not compromise hazard ratios [32]. Real-world data analyses have validated these recommendations, showing that applying ASCO-Friends expanded criteria can increase the eligible pool by 78% or more [24]. The FDA’s 2020 guidance provides cross-cutting regulatory encouragement for sponsors to broaden criteria and justify each exclusion [41].

This pattern extends beyond oncology: restrictive criteria exclude 63% of lupus patients [9], 86% of depressed outpatients from antidepressant trials [49], and up to 88% of candidates in preclinical Alzheimer’s trials [28]. Franzen et al. [12] show that common Alzheimer’s exclusions reduce racial and ethnic diversity in enrolled populations, while eligibility complexity in dementia trials has increased over time despite calls for simplification [34]. The proportion of cardiovascular trials excluding chronic kidney disease patients has risen from 66% to 79% over twenty years [6]. Governing bodies have responded with directives from KDIGO [1], the Global Cardiovascular Clinical Trialists Forum [11], and ICH E17 [21]. MDS-specific guidance has called for reducing eligibility barriers in hematologic malignancy trials [2]. Taken together, these are explicit, actionable recommendations distributed across governing bodies, academic literature, and institutional protocols.

2.2 Predicting and Implementing Criteria Changes

AMEND++ [7] introduces eligibility criteria amendment prediction as a supervised task, using Change-Aware Masked Language Modeling on historical version diffs spanning 161K trials. Their benchmark provides a resource for studying EC instability over time. RECITE shares the ClinicalTrials.gov data source but addresses a different goal: rather than predicting *whether* amendments will occur, RECITE discovers evidence specifying *what* changes should be made and generates the revised criteria accordingly. The two approaches are complementary: AMEND++ identifies trials at high amendment risk, while RECITE supplies evidence-grounded revisions. TrialPanorama [44] provides a large-scale trial database with benchmark tasks but does not address evidence-guided revision. Kim et al. [26] demonstrate data-driven eligibility criteria optimization for COVID-19 trials using real-world data, an approach complementary to our evidence-guided revision from published literature.

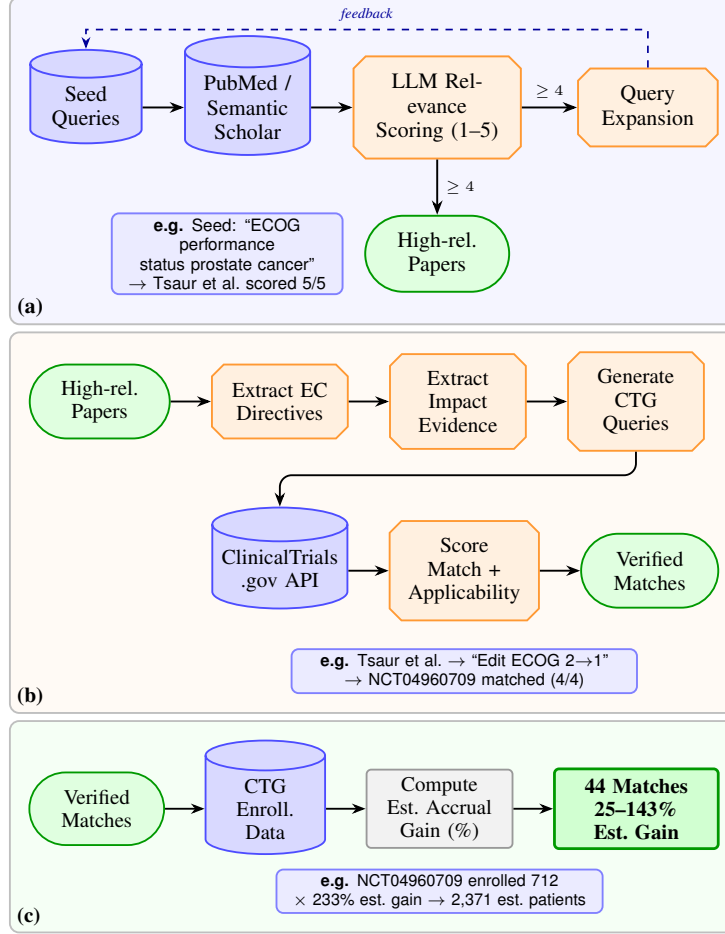


Figure 3: Agentic literature discovery system. (a) Iterative crawl: seed queries retrieve papers from PubMed and Semantic Scholar; an LLM scores relevance (1–5) and high-scoring papers expand the query set. (b) Screening and matching: directives and impact evidence are extracted from each paper, CTG API queries are generated, and candidate (paper, trial) pairs are scored. (c) Accrual estimation: verified matches are joined with CTG enrollment data to compute estimated gains (25–143%).

2.3 AI for Clinical Trial Design and Literature Mining

Generative approaches produce eligibility criteria from trial-level metadata. EC-RAFT [30] uses retrieval-augmented fine-tuning to generate criteria from trial descriptions, while AutoTrial [43] prompts language models for protocol generation. Embedding-based methods have also addressed trial-level representation; for example, Trial2Vec [42] learns self-supervised trial document embeddings for similarity search, and Bornet et al. [3] cluster eligibility criteria using sentence-level LLM embeddings, achieving 95% of generative-LLM ROUGE scores with compressed cluster representations. These approaches capture distributional patterns across existing criteria. Downstream tasks such as patient–trial matching [22] and criteria-to-query translation [37] would benefit from improved criteria. RECITE builds on retrieval-augmented generation [31] and dense passage retrieval [23] to connect published recommendations to trial-level eligibility criteria.

3 Problem Formulation

The task of *evidence-guided eligibility criteria revision* is defined as follows. Given a set of original eligibility criteria E_{orig} for a clinical trial and a documentary evidence source D (which may be a protocol amendment document, a published research paper, a clinical guideline, or any other text containing recommendations for criteria modification), the goal is to produce revised eligibility criteria E_{rev} with provenance. Specifically, a provenance mapping P links each revision to specific passages in D . Formally, the task requires a function $f : (E_{\text{orig}}, D) \rightarrow (E_{\text{rev}}, P)$ that performs two

subtasks: (1) extracting explicit modification directives from the documentary text, and (2) faithfully implementing those directives into revised criteria while preserving unmodified elements.

Protocol amendment documents provide naturally occurring ground-truth instances: each yields a triplet of original criteria, evidence document, and revised criteria. These documents vary in length, structure, and detail across institutions, phases, and therapeutic areas, exhibiting the same variation encountered across published papers, guidelines, and institutional reports. Validating on protocol documents therefore tests the core capability required for any evidence source. This framing is supported by prevalence: 76% of protocols undergo at least one amendment, and 53% modify eligibility criteria [15, 16].

Revised criteria are evaluated using an LLM-as-judge protocol. GPT-4o has demonstrated strong agreement with human evaluators on text equivalence and quality assessment tasks [48]. The judge (GPT-4o, accessed through UCSF Versa) scores each generated revision on a 0–4 ordinal scale: 0 indicates no overlap with the target revision, 1 indicates minimal overlap, 2 indicates partial overlap with notable gaps, 3 indicates the revision is largely equivalent to the target, and 4 indicates the revision is fully equivalent. We treat the ordinal score as the *primary* metric because validation against ten domain experts (Section 5.4) produces quadratic-weighted Cohen’s $\kappa_{qw} = 0.694 \pm 0.096$ (“substantial” per Landis & Koch [27]), within the inter-human range [0.531, 0.823]; on the direct binary acceptability scale, $\kappa = 0.292 \pm 0.286$ (“fair”), with roughly triple the standard deviation. We report direct binary acceptability, the proportion scoring ≥ 3 , and normalized ordinal score (judge score divided by 4) alongside the primary ordinal metric. Full prompt templates and scoring rubrics are provided in Appendix B.

4 RECITE System

4.1 Framework Overview

RECITE operates through a two-stage pipeline. In Stage 1 (Discovery), the system identifies relevant documentary evidence for a given trial’s eligibility criteria by searching published literature, matching papers to trials, and extracting modification directives. In Stage 2 (Revision), given original eligibility criteria and discovered (or provided) evidence, the system generates revised criteria with provenance links to specific evidence passages. For the benchmark evaluation in Section 5, Stage 1 is bypassed because the evidence documents are provided directly. For the real-world applications in Section 6, Stage 1 employs the agentic literature discovery system described below.

The revision stage supports two modes. In truncated-to-context-window mode, the entire evidence document is provided as context alongside the original eligibility criteria. In RAG mode, the evidence document is first chunked and embedded using BAAI/bge-small-en-v1.5 [46] via LlamaIndex, and relevant passages are retrieved via dense passage retrieval with a top- k of 10. Token budgets are enforced using tiktoken’s cl100k_base encoding: the query is truncated to 8,192 tokens before embedding, and retrieved passages are fitted within the LLM context window (16,384 tokens for locally hosted models) via LlamaIndex’s compact response mode. Both modes produce revised eligibility criteria that cite specific passages as provenance for each change. This dual-mode design allows full-context processing of shorter evidence documents (such as targeted amendment rationale sections), while RAG mode supports aggregation across longer, multi-source evidence collections such as research papers and clinical guidelines.

4.2 Agentic Literature Discovery

The literature discovery system uses orchestrator-driven iterative query expansion. A Python control loop alternates between LLM-generated search query formulation, adapter-based API search, and LLM-based relevance scoring. The LLM handles only the generative steps; all orchestration logic, rate limiting, and state management reside in the Python controller.

The system interfaces with two external APIs through dedicated adapters. The PubMed adapter issues boolean queries via NCBI E-utilities (esearch for identifier retrieval, efetch for metadata) with exponential backoff on transient failures. The Semantic Scholar adapter queries the Graph API, performing citation graph traversal to retrieve both the references and citing papers of each seed (up to 20 in each direction).

In each iteration, the system draws from a seed queue of high-relevance papers managed as a FIFO. If a seed is available, the LLM generates targeted queries informed by the seed’s content; otherwise, queries are generated from scratch, avoiding duplicates tracked in a seen-identifier set. Queries are dispatched to PubMed and Semantic Scholar adapters. Each candidate paper is scored on relevance (1–5), extraction confidence (1–5), and accrual impact ease (1–5); papers above the relevance threshold enter the seed queue for progressive scope refinement. The orchestrator terminates after 5 consecutive rounds in which zero new relevant papers are discovered and the seed queue is empty.

For paper-trial matching, each paper passing directive extraction (relevance ≥ 4) yields extracted change directives and accrual impact estimates. The LLM generates ClinicalTrials.gov search queries targeting the public API (v2), and each (paper, trial) pair is scored for match quality and applicability (both 1–5). High-quality matches (both ≥ 4) seed additional queries to expand coverage iteratively. Full scoring prompts and rubrics are provided in Appendix B.

4.3 Evidence-Grounded Revision

The revision pipeline accepts original criteria and evidence and produces revised criteria with passage-level provenance. In RAG mode, evidence documents are chunked, embedded via dense passage retrieval [23], and the top- k passages are provided as context. In truncated-to-context-window mode, the entire document serves as context. Both modes instruct the model to extract explicit directives, apply them, and cite supporting passages.

Directive extraction is validated by fuzzy matching against source text to prevent hallucination. The matching procedure uses a sliding-window approach over the evidence document, comparing each window to the extracted directive with SequenceMatcher-based similarity. Directives with low similarity scores are flagged as potentially unsupported; final validation was performed by manual review against source text.

5 RECITE Benchmark

Protocol amendment documents anchor the RECITE benchmark because each amendment produces a naturally occurring triplet—original eligibility criteria, documentary evidence, and revised criteria, requiring no manual annotation. These documents vary considerably in length, structure, and detail across sponsoring institutions, clinical phases, and therapeutic areas, similar to the diversity encountered in published papers, clinical guidelines, and institutional reports. Validating directive extraction on this heterogeneous evidence base therefore tests the core capability required for any documentary source type.

5.1 Data Collection and Curation

The dataset construction begins with 567,815 registered trials from the ClinicalTrials.gov bulk XML download. To identify trials with eligibility criteria modifications, a two-stage heuristic is applied. First, the ClinicalTrials.gov history API reports which modules changed in each version via `moduleLabels` metadata; only versions flagged as modifying the “Eligibility” module are retained. This reduces the candidate set from 567K to 59,996 trials without requiring full text comparison across all version pairs. Second, within this subset, both versions of the eligibility criteria text are normalized (whitespace collapsed, empty lines removed, HTML tags stripped) and compared using a `diff`lib-based similarity check. Pairs with similarity above 0.80 or with fewer than 5% of lines changed are discarded as metadata-only or formatting-only differences. Comparing consecutive version pairs across the retained trials produced 85,083 confirmed eligibility criteria changes across 49,369 unique trials.

To include grounding evidence, trials with downloadable protocol amendment PDFs were identified. The ClinicalTrials.gov API was queried for each of the 49,369 trials with EC changes, checking the `largeDocumentModule` metadata for protocol availability. Roughly 3,500 trials (approximately 7%) had such documents. Evidence text was extracted from these PDFs using PyMuPDF, then joined with the corresponding eligibility criteria changes to produce complete triplets: original criteria, evidence document, and revised criteria. Amendment-to-version matching uses a date-based heuristic with a 7-day tolerance window; where dates are unavailable, a fallback assumes amendment N corresponds to version N . The resulting dataset contains 11,913 instances across 5,735 trials. A stratified sample of 3,116 instances was selected for benchmarking to manage computational costs: evaluating all fifteen model configurations under both truncated-to-context-window and RAG conditions requires substantial resources given the mean evidence document length of approximately 134,000 characters. Because evidence documents frequently exceed model context windows, we report *evidence retention* (Ret%) for each context size: 8K models (Gemma 2) retain 23.5% of the evidence text (12.5% full-EC coverage); 32K models (Qwen 2.5 base) retain 81.1% (53.5%); and 128K models (GPT-4o, GPT-4o-mini, Qwen 2.5-7B-128K, Llama-3.1-70B, Qwen3-32B) retain 99.4% (98.6%). Each instance provides three components: (1) the original eligibility criteria prior to amendment, (2) grounding evidence extracted from the protocol amendment document, and (3) the target revised eligibility criteria as implemented in the subsequent protocol version.

5.2 Dataset Characteristics

Table 1 summarizes the benchmark dataset. The 5,735 unique trials span a broad range of conditions, with breast cancer (98 trials), prostate cancer (57), COVID-19 (56), obesity (53), and heart failure (50) as the most represented.

Table 1: RECITE benchmark dataset summary.

Statistic	Value
Total instances	11,913
Unique trials	5,735
Sampled for benchmark	3,116
Avg pre-amendment EC len. (chars)	3,331
Avg amended EC len. (chars)	3,473
Avg evidence len. (chars)	134,180
<i>Trial phases</i>	
Phase 1 / 1–2 / 2	433 / 439 / 1,663
Phase 3 / 4	1,059 / 316
Other / N/A	451 / 1,374
Interventional / observational	5,462 / 272
Trial start date range	2008–2023
<i>Top conditions (unique trials)</i>	
Breast cancer	98
Prostate cancer	57
COVID-19	56
Obesity / Heart failure	53 / 50

Phase 2 trials account for the largest share (1,663 trials), followed by trials with no reported phase (1,374), Phase 3 (1,059), and Phase 1/2 (439). Trial start dates range from 2008 to 2023. The average evidence document length is approximately 134,000 characters, with wide variation (up to 3.1M characters), reflecting the diversity of protocol documents across institutions and trial phases.

Protocol amendment documents differ considerably in length, structure, and level of detail across sponsoring institutions, clinical phases, and therapeutic areas. Some amendments provide tabular summaries of changes with explicit rationale; others embed modification directives within narrative justification spanning multiple pages. This variation is similar to the diversity of evidence sources the framework must handle, since published papers differ in structure from clinical guidelines, which differ from institutional reports.

5.3 Model Evaluation

Fifteen model configurations were evaluated on the 3,116-instance sample under truncated-to-context-window conditions. The suite spans (i) base open-weight models (Qwen 2.5 0.5B/3B/7B; Qwen 2.5-7B-128K; Gemma 2 2B/9B); (ii) commercial API models accessed through UCSF Versa, a PHI-safe HIPAA-compliant Azure deployment (GPT-4o-mini, GPT-4o [35], GPT-4.1 [36]); (iii) scaling open-weights (Gemma 2-27B [14], Qwen 2.5-72B [38], Llama-3.1-70B [18], Qwen3-32B [39]); and (iv) two finetuned variants (Qwen 2.5-3B and Qwen 2.5-0.5B with QLoRA [8] on 720 stratified samples drawn from 8,838 non-benchmark instances).

Reporting strategy. We present the ordinal 0–4 score as the primary metric and report direct binary acceptability and the proportion scoring ≥ 3 alongside it. This choice is based on judge-against-human inter-rater reliability (Section 5.4): on the ordinal scale, quadratic-weighted Cohen’s $\kappa_{qw} = 0.694 \pm 0.096$ (“substantial” per Landis & Koch [27]); on the direct binary acceptability scale, $\kappa = 0.292 \pm 0.286$ (“fair”), with standard deviation roughly triple that of the ordinal scale.

Substantial-change subset. Many EC version pairs in the benchmark differ only by trivial edits (whitespace, single-token rewordings) that any model accepts. To test directive extraction more rigorously, we report a *substantial-change* subset using Gestalt pattern-matching similarity (Python’s `SequenceMatcher`) at threshold $\text{sim} \leq 0.95$, which retains $n = 1,351$ pairs (43.4% of the benchmark).

Table 2 presents results for all fifteen configurations, with both full-benchmark and substantial-change splits.

Results. On the full benchmark, GPT-4o achieves the highest ordinal score (3.0 ± 0.8 , normalized 0.745, 75.3% scoring ≥ 3); Gemma 2-9B and Qwen 2.5-72B match it within margin (3.0 ± 0.9 and 3.0 ± 0.8 respectively). Qwen 2.5-7B-128K, despite its long-context capacity (Ret% 99.4), achieves only 0.2 ± 0.6 ordinal; context length alone does not substitute for capability. On the substantial-change subset, all models show lower performance (e.g., GPT-4o drops

Table 2: RECITE benchmark across 15 model configurations on the full sample ($n = 3,116$) and the substantial-change subset ($\text{sim} \leq 0.95$, $n = 1,351$, of full). Bin Acc = direct binary acceptability from LLM judge (0/1). Ord = ordinal 0–4 score (mean $\pm \sigma$). NormOrd = ordinal/4. $\% \geq 3$ / $\% \geq 4$ = proportion of ordinal ratings at or above threshold. Ret% = evidence retained after truncation to model context window. Reliability against ten domain experts: $\kappa_{\text{binary}} = .292 \pm .286$; $\kappa_{qw} = .694 \pm .096$ (Landis & Koch [27]: “fair” vs. “substantial”). Finetuned: QLoRA [8] on 720 stratified samples drawn from 8,838 non-benchmark instances.

Model	Aggregate			Full ($n=3,116$)		sim $\leq .95$ ($n=1,351$)		Ret%
	Bin Acc $\pm \sigma$	Ord $\pm \sigma$	NormOrd $\pm \sigma$	% ≥ 3	% ≥ 4	% ≥ 3	% ≥ 4	
<i>Base Models</i>								
Qwen2.5-0.5B	.01 $\pm$.11	0.7 $\pm$ 0.9	.179 $\pm$.22	3.8	1.4	3.2	1.0	81.1
Qwen2.5-3B	.09 $\pm$.29	2.5 $\pm$ 0.9	.618 $\pm$.23	49.5	12.6	42.8	8.0	81.1
Qwen2.5-7B	.18 $\pm$.39	2.8 $\pm$ 0.9	.700 $\pm$.22	63.1	22.7	56.3	16.9	81.1
Qwen2.5-7B-128K	.01 $\pm$.08	0.2 $\pm$ 0.6	.050 $\pm$.15	1.3	0.7	0.9	0.4	99.4
Gemma2-2B	.04 $\pm$.20	1.9 $\pm$ 0.9	.483 $\pm$.21	19.9	4.8	17.5	2.8	23.5
Gemma2-9B	.25 $\pm$.44	3.0$\pm$0.9	.742 $\pm$.22	70.1	30.7	58.4	18.8	23.5
<i>API Models</i>								
GPT-4o-mini	.18 $\pm$.38	2.9 $\pm$ 0.8	.726 $\pm$.19	69.7	22.4	63.4	17.4	99.4
GPT-4o	.20 $\pm$.40	3.0$\pm$0.8	.745 $\pm$.19	75.3	25.0	68.4	19.0	99.4
GPT-4.1	.04 $\pm$.19	2.7 $\pm$ 0.6	.673 $\pm$.16	66.5	5.8	54.4	4.1	99.4
<i>Scaling Models</i>								
Gemma2-27B	.23 $\pm$.42	2.8 $\pm$ 1.0	.699 $\pm$.25	63.7	27.1	54.0	16.7	23.5
Qwen2.5-72B	.23 $\pm$.42	3.0$\pm$0.8	.754 $\pm$.20	77.0	27.8	67.6	18.1	97.3
Llama-3.1-70B	.38 $\pm$.49	2.8 $\pm$ 1.2	.691 $\pm$.31	61.6	37.2	52.6	28.6	99.4
Qwen3-32B	.35 $\pm$.48	2.6 $\pm$ 1.4	.642 $\pm$.36	60.8	34.6	54.1	25.9	99.4
<i>Finetuned Models (QLoRA, 720 stratified samples)</i>								
Qwen2.5-3B (QLoRA)	.13 $\pm$.34	2.6 $\pm$ 0.9	.645 $\pm$.23	51.7	17.5	40.1	9.4	81.1
Qwen2.5-0.5B (QLoRA)	.00 $\pm$.06	0.5 $\pm$ 0.8	.130 $\pm$.19	1.5	0.6	1.3	0.2	81.1

from 75.3% to 68.4% scoring ≥ 3), indicating that precise directive implementation remains an open problem even for the strongest models. Scaling within families is consistent: Qwen 2.5 (0.5B $\rightarrow$ 3B $\rightarrow$ 7B) shows ordinal 0.7 $\rightarrow$ 2.5 $\rightarrow$ 2.8; Gemma 2 (2B $\rightarrow$ 9B $\rightarrow$ 27B) shows 1.9 $\rightarrow$ 3.0 $\rightarrow$ 2.8. QLoRA finetuning of small models produces modest gains for the 3B (ordinal 2.5 $\rightarrow$ 2.6) and minimal change for the 0.5B (0.7 $\rightarrow$ 0.5).

RAG provides modest and variable gains across the long-context models tested in both conditions. For capable 128K-window models, the truncated-to-context-window setting already saturates available evidence (Ret% 99.4); RAG gains primarily benefit small-context models on long evidence documents.

Figure 4 shows that evidence document length has negligible effect on revision quality (Spearman $\rho = -0.04$). The RAG and no-RAG curves are highly correlated (Pearson $R > 0.96$) and differ primarily in level rather than trend. In the single-document setting, even documents exceeding 10^6 characters remain within capable models’ processing range; retrieval-based methods will become necessary when extending to multi-document evidence sources.

5.4 Human Evaluation of the Automated Judge

To validate the GPT-4o-as-judge protocol, we conducted a structured human evaluation in which ten domain expert judges rated 160 held-out instances, producing 200 paired ratings against the automated judge. Inter-human reliability spans a moderate-to-substantial range, $\kappa_{qw} \in [0.531, 0.823]$. EC equivalence judgments retain irreducible expert subjectivity even with a fixed rubric.

Comparing the automated judge to the human ratings yields:

- **Ordinal (0–4):** $\kappa_{qw} = 0.694 \pm 0.096$ — “substantial” per Landis & Koch [27].
- **Binary:** $\kappa = 0.292 \pm 0.286$ — “fair”. The standard deviation is roughly triple that of the ordinal scale; binary judgment discards information about partial agreement.

The ordinal κ_{qw} falls within the inter-human range, so the automated judge is not measurably worse than a representative human rater on the 0–4 scale. We therefore retain ordinal as the primary metric throughout the paper and report direct binary acceptability for continuity with prior work.

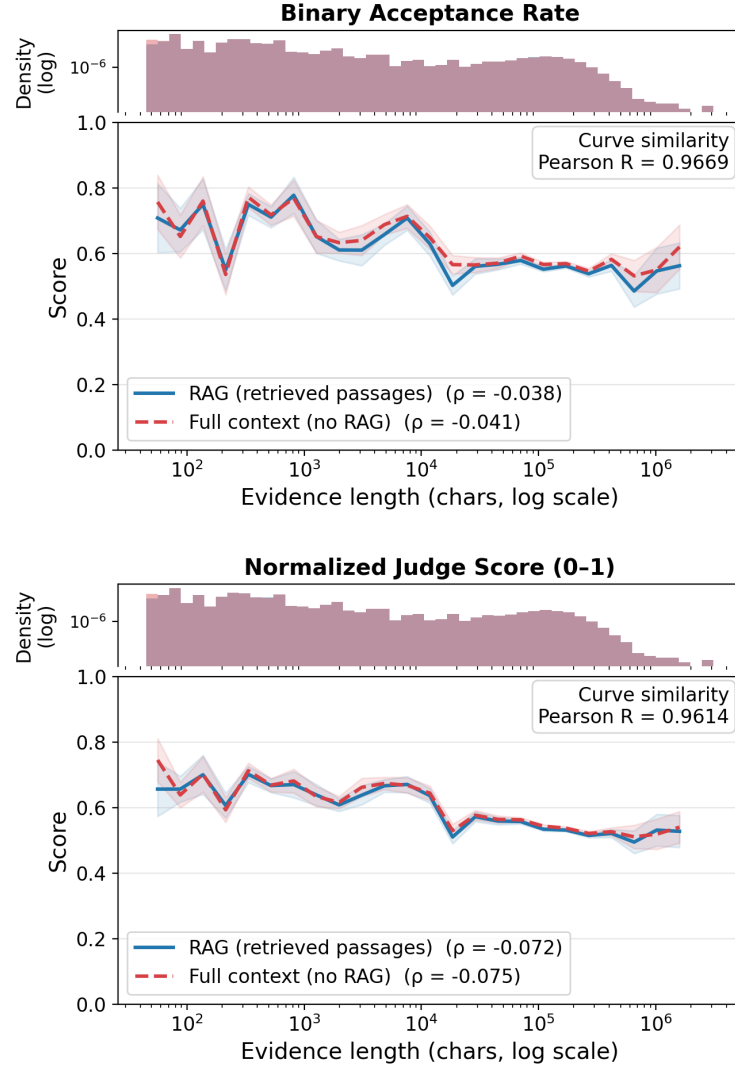


Figure 4: LLM judge performance as a function of evidence document length, pooled across the 128K-context models evaluated under both conditions. (a) Direct binary acceptability rate. (b) Normalized judge score (0–1). Each panel shows RAG and full-context (no-RAG) curves over 25 log-spaced length bins with 95% CI bands; a marginal density histogram (log-scaled) along the top axis shows the evidence-length distribution.

6 Real-World Applications

Table 3: All paper-trial matches with extracted directives and accrual impact quotes. **Des/EC**: source paper provides trial-design recommendations / explicit EC modification directives (T/F). ECOG GU and ECOG all reclassified F/F (document scoring practices, not protocol-level EC changes). Enrollment columns show current and directional estimate under broadened criteria; estimates are omitted (—) for F/F papers whose reported gains reflect scoring-practice changes, not protocol-level EC modifications. M/A: LLM match/applicability scores (1–5).

Source	Des/EC	EC Change Directive	Accrual Impact Quote	Trial (NCT ID)	Enrollment		M / A
					Current →	Est.	
Harvey [20]	T/T	Allow brain mets, other malignancies, CrCl ≥ 30 mL/min	Potential to double eligible patient population	NCT03101995	18 →	36	4/5
				NCT04163432	43 →	86	4/5
Kim [19]	T/T	Allow brain mets, other malignancies, lower CrCl to >30 mL/min	Nearly 2× aNSCLC pts eligible (4851, 46.2%)	NCT02792582	160 →	320	3/4
				NCT03367819	44 →	88	3/4
				NCT04114136	72 →	144	4/5
				NCT04488874	50 →	100	3/4
				NCT04739761	506 →	1,012	5/5
				NCT05588453	30 →	60	4/5
				NCT06157541	58 →	116	4/5
KRAS G12C [33]	T/T	Liberalize renal dysfunction and brain mets criteria	+25% enrollment without decreasing median OS	NCT06481813	730 →	1,460	4/5
				NCT01509937	433 →	541	3/3
				NCT04013620	48 →	60	4/4
				NCT04467021	61 →	76	4/4
				NCT04704258	100 →	125	3/4
				NCT05534620	36 →	45	4/4
Fung [13]	T/T	Include ECOG 2 or Child-Pugh B7 patients	$>2\times$ patients treated if MEC applied	NCT06531824	11,000 →	13,750	4/4
				NCT00485303	58 →	141	4/5
				NCT01273155	72 →	175	5/5
				NCT01967576	14 →	34	4/5
				NCT02314143	48 →	116	3/4
				NCT03494816	24 →	58	4/4
				NCT06889675	36 →	87	4/5
				NCT06926673	62 →	151	4/4
				NCT06984497	32 →	78	4/4
				NCT06990659	206 →	500	4/4
ECOG GU [17]	F/F	Edit ECOG 2→1 (prostate), 3→2 (bladder)	+233% (prostate), +300% (bladder) matched trials/pt	NCT07037394	252 →	612	4/4
				NCT00428714	73 →	—	4/4
				NCT00485303	58 →	—	4/5
				NCT00699751	921 →	—	4/5
				NCT01756638	48 →	—	4/5
				NCT03768570	82 →	—	4/4
				NCT04960709	712 →	—	4/4
				NCT05171387	78 →	—	4/5
				NCT05348876	50 →	—	4/5
				NCT06524544	320 →	—	4/4
ECOG all [29]	F/F	Edit ECOG 3→2 or 2→1 across cancer types	+246% and +189% matched CTs ($p<0.001$)	NCT06661122	100 →	—	4/4
				NCT00022672	208 →	—	3/3
				NCT00485303	58 →	—	4/5
				NCT01967576	14 →	—	4/5
				NCT03022032	102 →	—	3/4
				NCT03288987	126 →	—	4/5
				NCT03355079	2 →	—	4/5
				NCT05721755	290 →	—	4/4
Total:	6 source papers, 44 paper-trial matches;	Des=T/EC Mod=T for 4 papers; reclassified F/F for 2 (ECOG GU, ECOG all).		NCT06677255	149 →	—	3/4

To evaluate RECITE beyond the protocol-based benchmark, we applied the full two-stage pipeline to published research literature. The agentic crawling system (Section 4.2) retrieved papers from PubMed that recommend specific eligibility criteria changes, matched them to registered trials on ClinicalTrials.gov, and extracted modification directives with associated accrual impact estimates.

6.1 Literature–Trial Matching

The system identified 44 paper–trial matches across 6 source publications and 41 unique trials, spanning ASCO-Friends criteria broadening [19, 20], organ function threshold liberalization [33], expanded hepatocellular carcinoma criteria [13], and ECOG performance status adjustments [17, 29]. Each match is a pairing in which the paper’s recommended eligibility change is applicable to the trial’s current criteria, as scored by the LLM pipeline described in Section 4.2.

Table 3 presents all matches grouped by source paper with current and estimated enrollment. Directives range from targeted threshold changes to broader performance status adjustments, all validated by fuzzy matching against source publications. Match quality scores and rubrics are in Appendix A and Appendix B.

6.2 Estimated Accrual Impact

Here, we present the underlying quoted evidence for the estimated accrual impact of each study, when possible. We also provide a binary classification of each source for whether it (i) provides trial-design recommendations (Des) and (ii) issues explicit directives to modify eligibility criteria (EC Mod):

- **Harvey et al. [20]** [Des: T, EC Mod: T] — “Broadening [brain metastases, prior malignancies, CrCl $\geq$ 30 mL/min] has the potential to *double* the eligible patient population.”
- **Harvey et al. [19]** [Des: T, EC Mod: T] — “Patients with treated brain metastases should be routinely included”; “strive for inclusiveness” (ASCO-Friends consensus).
- **KRAS G12C [33]** [Des: T, EC Mod: T] — “Liberalizing renal dysfunction and brain metastases criteria yielded +25% enrollment without decreasing median OS.”
- **Fung et al. [13]** [Des: T, EC Mod: T] — “Expanding to [ECOG PS 2 or Child-Pugh B7] increased eligibility to 31.7%” (from 13.1%).

All four source papers issue explicit directives to broaden eligibility. The directional signal is consistent: liberalization of brain-metastasis exclusions, prior-malignancy exclusions, organ-function thresholds, and ECOG cutoffs is associated with substantially expanded eligibility in the source-paper analyses. We do not assert specific percentage gains for the matched trials; downstream verification against actual enrollment outcomes remains future work.

6.3 Scope and Interpretation

The current application focuses on oncology, given the maturity of EC liberalization research in that area; extension to other therapeutic areas is architecturally straightforward but not yet empirically validated. All matches were among the highest-scoring pairs (see Appendix B).

6.4 Illustrative Example

To illustrate how a paper’s directive applies to a specific trial, consider Harvey et al. [20] matched to NCT04163432 (Durvalumab chemo-immunotherapy for metastatic non-squamous NSCLC; match 4/5, applicability 5/5). The extracted directive recommends allowing patients with brain metastases, other malignancies, and a creatinine clearance of $\geq$ 30 mL/min. The original trial already included a renal criterion of eGFR $\geq$ 30 mL/min, consistent with the paper’s CrCl recommendation. However, the trial excluded patients with “active central nervous system metastases and/or carcinomatous meningitis” and “other active malignancy within 2 years except for basal cell carcinoma of the skin or carcinoma in situ of the cervix.” Applying the directive broadens both exclusions to permit controlled brain metastases and prior malignancies beyond the narrow exceptions. The source paper reports that broadening these three criteria has “the potential to double the eligible patient population” [20]. Additional before-and-after criteria examples are included in Appendix C.

7 Limitations and Ethical Considerations

Data and benchmark. The RECITE benchmark draws from a specific subset of the trial population: of 567,815 registered trials, only those with publicly downloadable protocol amendment PDFs contribute evidence documents, producing roughly 3,500 trials (0.6% of the registry). Table 1 shows that Phase 2 and Phase 3 trials account for a disproportionate share, and top conditions are weighted toward oncology and well-funded disease areas. The benchmark construction pipeline relies on two key heuristics that trade completeness for efficiency. First, eligibility criteria changes are identified via ClinicalTrials.gov `moduleLabels` metadata rather than exhaustive version-pair comparison across all 567K trials, potentially missing changes where registry metadata is incomplete. Second, text-level change detection uses similarity thresholds that filter out very small substantive changes alongside formatting-only differences. The benchmark validates the framework on protocol documents, and the applications in Section 6 demonstrate transfer to published papers; systematic evaluation across the full range of document types (e.g., clinical guidelines, institutional reports) remains future work. The LLM-as-judge evaluation was validated through manual inspection of scored instances. The equivalence-judging task (determining whether two eligibility criteria texts are functionally identical) is well-suited to LLM evaluation given its reliance on structured text comparison. However, LLM judges may exhibit systematic biases related to hallucination or scoring precision, and inter-rater reliability against human expert judgment has not been measured at scale.

Applications. The 44 paper-trial matches represent the highest-scoring pairs within the scoring system (see Appendix B), not a random sample. Reported accrual signals are inherited directly from the source publications and treated as directional rather than as independent numeric predictions for the matched trials (Section 6.2). RECITE surfaces and operationalizes existing research conclusions rather than generating enrollment forecasts. Mechanically scaling source-paper percentages onto matched trials would assume patient-population homogeneity that the source papers do not establish. These estimates were not independently verified against actual enrollment data. The current applications focus on oncology, and directive extraction operates on abstracts rather than full text; more detailed recommendations in methods or results sections may therefore be missed.

Ethics. RECITE is a decision-support tool, not a replacement for clinical judgment. All proposed revisions require review by clinical investigators, IRB approval, and regulatory review where applicable [41]. The system provides provenance linking each revision to specific source passages to support expert evaluation. The published literature from which RECITE draws recommendations may itself reflect biases in what has been studied. Although broadening eligibility criteria generally improves demographic diversity [24], RECITE does not explicitly optimize for representativeness.

Future work. Several directions remain. Extending directive extraction to full paper text (beyond abstracts) would capture more detailed recommendations in methods and results sections. Constructing a literature-linked benchmark from protocol document references would complement our protocol-based benchmark with a paper-based evaluation setting. Matching modified criteria to actual enrollment outcomes using longitudinal registry data would provide ground-truth validation of the accrual impact estimation pipeline. Temporal meta-analysis of how eligibility criteria evolve in response to published evidence would further clarify the relationship between evidence publication and protocol modification. Finally, assessing and improving the robustness of LLM-inferred directives through tests specifically tailored to trial design specifications [45] would strengthen the trustworthiness of extracted directives and represents a natural continuation of this line of work.

8 Conclusion

RECITE provides infrastructure to discover, extract, and implement evidence-based eligibility criteria revisions with full provenance. On a benchmark of 3,116 instances derived from protocol amendments on ClinicalTrials.gov, the best configurations (GPT-4o, Gemma 2-9B, Qwen 2.5-72B) achieve ordinal scores of approximately 3.0/4 under an LLM-as-judge protocol validated against ten domain experts ($\kappa_{qw} = 0.694$, “substantial”). Applied to published literature through an agentic discovery system, RECITE identifies 44 paper-trial matches across five publications providing evidence-grounded directives for EC liberalization. Both evidence types (protocol amendments and published papers) share the property that makes this task tractable: explicit modification directives embedded in varied text. The underlying approach of extracting directives and implementing them in design documents could generalize to other trial elements such as endpoint selection or dosing regimens.

GenAI Disclosure

Research methodology. Large language models are the primary subject of this work. Fifteen LLM configurations were evaluated across base, scaling, API, and finetuned (QLoRA) categories. GPT-4o served as the LLM-as-judge evaluator (Section 3), validated against ten domain experts. LLMs were also used within the agentic literature discovery pipeline for search query formulation, relevance scoring, and directive extraction (Section 4.2).

Manuscript preparation. Claude (Anthropic) was used to assist with editing, restructuring, and refining manuscript prose. All generated or modified text was reviewed, verified, and approved by the authors.

Code development. Claude (Anthropic) was used for code completion and debugging support. All code was reviewed and tested by the authors.

Acknowledgments

This work was supported in part by the National Science Foundation Graduate Research Fellowship Program. Computational resources were provided by UCSF Versa. The RECITE framework is available at <https://github.com/abbasilab/RECITE>.

References

- [1] Colin Baigent, William G Herrington, Josef Coresh, Martin J Landray, Adeera Levin, Vlado Perkovic, et al. Challenges in conducting clinical trials in nephrology: conclusions from a Kidney Disease—Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney International*, 92(2):297–305, 2017. doi: 10.1016/j.kint.2017.04.019. “Eligibility criteria must be practical and broad”.
- [2] Uma Borate, Kelly Pugh, Allyson Waller, Rina Li Welkie, Ying Huang, Jan Philipp Bewersdorf, Maximilian Stahl, Amy E DeZern, Uwe Platzbecker, Mikkael A Sekeres, et al. Reducing clinical trial eligibility barriers for patients with myelodysplastic syndromes. *Blood*, 145(13):1369, 2025. URL <https://ashpublications.org/blood/article/145/13/1369/534836/Reducing-clinical-trial-eligibility-barriers-for>. Non-oncology EC liberalization example in hematology/MDS.
- [3] Alban Bornet, Philipp Khlebnikov, Florian Meer, Quentin Haas, Anthony Yazdani, Boya Zhang, Poorya Amini, and Douglas Teodoro. Analysis of eligibility criteria clusters based on large language models for clinical trial design. *Journal of the American Medical Informatics Association*, 32(3):447–458, 2025. doi: 10.1093/jamia/ocae311. LLM-embedding clustering of EC for protocol generation.
- [4] Peter Bower, Paul Wallace, Elaine Ward, Jonathan Graffy, Joe Miller, Brendan Delaney, and Ann Louise Kinmonth. Improving recruitment to health research in primary care. *Family Practice*, 26(5):391–397, 2009. doi: 10.1093/fampra/cmp037. Study of 114 UK MRC trials: only 31% met enrollment goals.
- [5] Marion K Campbell, Claire Snowdon, Diane Francis, Diana Elbourne, Alison M McDonald, Rosemary Knight, and Adrian Grant. Recruitment to randomised trials: strategies for trial enrollment and participation study: the STEPS study. *Health Technology Assessment*, 11(48):iii–ix, 105, 2007. One-third of publicly funded trials required time extensions for recruitment shortfalls.
- [6] Julia M T Colombijn, Demy L Idema, Sanne van Beem, et al. Representation of patients with chronic kidney disease in clinical trials of cardiovascular disease medications: A systematic review. *JAMA Network Open*, 7(3):e240427, 2024. doi: 10.1001/jamanetworkopen.2024.0427. CKD exclusion rose from 66% to 79% over two decades.
- [7] Trisha Das, Mandis Beigi, Jacob Aptekar, and Jimeng Sun. AMEND++: Benchmarking eligibility criteria amendments in clinical trials, 2025. Binary classification: predicts WHETHER amendments occur.
- [8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 2023.
- [9] Sarah Dyball, Sophie Collinson, Emily Sutton, Eoghan M McCarthy, Ian N Bruce, and Ben Parker. Lupus clinical trial eligibility in a real-world setting: results from the British Isles Lupus Assessment Group-Biologics Register (BILAG-BR). *Lupus Science & Medicine*, 8:e000513, 2021. doi: 10.1136/lupus-2021-000513. 63% of SLE patients ineligible for non-renal Phase III trials.
- [10] Stephen Feller. One in four cancer trials fails to enroll enough participants. UPI Health News, 2015. URL https://www.upi.com/Health_News/2015/12/30/

- [One-in-four-cancer-trials-fails-to-enroll-enough-participants/2611451485504/](#). 25% of cancer trials failed to enroll sufficient patients; 18% closed with less than half target.
- [11] Laura Filbey, Harriette G C Van Spall, et al. Improving representativeness in trials: a call to action from the Global Cardiovascular Clinical Trialists Forum. *European Heart Journal*, 44(11):921–930, 2023. doi: 10.1093/eurheartj/ehac810.
 - [12] Sanne Franzen, Jade Emily Smith, Esther van den Berg, Monica Rivera Mindt, et al. Diversity in Alzheimer’s disease drug trials: The importance of eligibility criteria. *Alzheimer’s & Dementia*, 18(4):810–823, 2022. doi: 10.1002/alz.12433.
 - [13] Alan S Fung, Vincent C Tam, Daniel E Meyers, Hao-Wen Sim, Jennifer J Knox, et al. Real world eligibility for cabozantinib (C), regorafenib (Reg), and ramucirumab (Ram) in hepatocellular carcinoma (HCC) patients after sorafenib (S). In *Journal of Clinical Oncology*, volume 37, page 422, 2019. doi: 10.1200/JCO.2019.37.4_SUPPL.422.
 - [14] Gemma Team. Gemma 2: Improving open language models at a practical size. arXiv:2408.00118, 2024.
 - [15] Kenneth Getz, Zachary Smith, Emily Botto, Elisabeth Murphy, and Arnaud Dauchy. New benchmarks on protocol amendment practices, trends and their impact on clinical trial performance. *Therapeutic Innovation & Regulatory Science*, 58(3):539–548, 2024. doi: 10.1007/s43441-024-00622-9. 75.8% prevalence; mean 3.3 amendments; 260-day implementation.
 - [16] Kenneth A Getz, Stella Stergiopoulos, Mary Short, Leon Surgeon, Randy Krauss, Sybrand Pretorius, Julian Desmond, and Derek Dunn. The impact of protocol amendments on clinical trial performance and cost. *Therapeutic Innovation & Regulatory Science*, 50(4):436–441, 2016. doi: 10.1177/2168479016632271. 57% of protocols amended; 53% modify EC; \$535K per Phase III amendment.
 - [17] Liat Gortzak Uzan, Liran Lapidot, Tal Bader, and Anat Gaziel. ECOG score as a barrier for clinical trial eligibility in the patient population with genitourinary cancer: Is it time to adapt this criterion to real-world data? In *Journal of Clinical Oncology*, volume 41, page 78, 2023. doi: 10.1200/jco.2023.41.6_suppl.78.
 - [18] Aaron Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.
 - [19] R. Donald Harvey, Wendy S Rubinstein, Gwynn Ison, Sean Khozin, Li Chen, Elizabeth Garrett-Mayer, et al. Impact of broadening clinical trial eligibility criteria for advanced non-small cell lung cancer patients: Real-world analysis. In *Journal of Clinical Oncology*, volume 37, page LBA108, 2019. doi: 10.1200/JCO.2019.37.18_SUPPL.LBA108.
 - [20] R. Donald Harvey, Suanna S Bruinooge, Li Chen, Elizabeth Garrett-Mayer, Whitney Rhodes, Edward Stepan-ski, Thomas S Uldrick, Gwynn Ison, Sean Khozin, William S Rubinstein, Charles Schenkel, Richard S Miller, George A Komatsoulis, Richard L Schilsky, and Edward S Kim. Impact of broadening trial eligibility criteria for patients with advanced non–small cell lung cancer: Real-World analysis of select ASCO-Friends recommendations. *Clinical Cancer Research*, 27(9):2430–2434, 2021. doi: 10.1158/1078-0432.CCR-20-3857.
 - [21] International Council for Harmonisation. General principles for planning and design of multi-regional clinical trials (E17). <https://www.fda.gov/media/99974/download>, 2017. Addresses eligibility harmonization across regions.
 - [22] Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetha Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. TrialGPT: Matching patients to clinical trials with large language models. *Nature Communications*, 15(1):9074, 2024. doi: 10.1038/s41467-024-53081-z. Patient-trial matching downstream of EC.
 - [23] Vladimir Karpukhin, Barlas Öğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.
 - [24] Maneet Kaur, Filip Frahm, Yichen Lu, Mustafa S Ascha, Jenny S Guadamuz, Efrat Dotan, Adam S Gottesman, et al. Broadening eligibility criteria and diversity among patients for cancer clinical trials. *NEJM Evidence*, 3(4): EVIDoa2300236, 2024. doi: 10.1056/EVIDoa2300236. 78% increase in eligible pool with broadened criteria.
 - [25] Edward S Kim, Suanna S Bruinooge, Sarah Roberts, Gladys Berénos-Riley, Robert Lorence, Thomas Mack, Richard L Schilsky, Sigal A Wagner, James R Rigas, Patrick J Loehrer, et al. Broadening eligibility criteria to make clinical trials more representative: American Society of Clinical Oncology and Friends of Cancer Research joint research statement. *Journal of Clinical Oncology*, 35(33):3737–3744, 2017. doi: 10.1200/JCO.2017.73.7916. ASCO-Friends recommendations for brain mets, prior malignancies, organ function.

- [26] Jae Hyun Kim, Casey N Ta, Cong Liu, Cynthia Sung, Alex M Butler, Latoya A Stewart, Lyudmila Ena, James R Rogers, Junghwan Lee, Anna Ostropolets, Patrick B Ryan, Hao Liu, Shing M Lee, Mitchell S V Elkind, and Chunhua Weng. Towards clinical data-driven eligibility criteria optimization for interventional COVID-19 clinical trials. *Journal of the American Medical Informatics Association*, 28(1):14–22, 2021. doi: 10.1093/jamia/ocaa276. Data-driven EC optimization for COVID-19 trials using real-world data.
- [27] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- [28] Jessica B Langbaum et al. Recommendations to address key recruitment challenges of Alzheimer’s disease clinical trials. *Alzheimer’s & Dementia*, 19(2):696–707, 2023. doi: 10.1002/alz.12737. Screen-failure rates reach 88% in preclinical AD trials.
- [29] Yael Lapidot, Merav Safran, Tal Bader, and Anat Gaziel. Profound impact of ECOG performance status on clinical trial eligibility and access: Diverging effects by cancer type. In *JCO Global Oncology*, volume 9, page 97, 2023. doi: 10.1200/go.2023.9.supplement.1.97.
- [30] Nopporn Lekuthai, Nattawit Pewngam, Supitcha Sokrai, and Titipat Achakulvisut. EC-RAFT: Automated generation of clinical trial eligibility criteria through retrieval-augmented fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9432–9444, 2025. doi: 10.18653/v1/2025.findings-acl.491. Retrieval-augmented fine-tuning for EC generation from trial descriptions.
- [31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.
- [32] Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ryan Copping, Marius Garmhausen, and James Zou. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*, 592(7855):629–633, 2021. doi: 10.1038/s41586-021-03430-5. Trial Pathfinder: data-driven EC optimization.
- [33] Meaghan S May, Mark Wooster, Prashasti Agrawal, Jae Won Lee, Brandon May, Xiaomei Ma, et al. Multicenter study of the impact of trial eligibility criteria on enrollment to KRAS G12C inhibitor trials in patients with non-small cell lung cancer. In *Journal of Clinical Oncology*, volume 43, page 11171, 2025. doi: 10.1200/jco.2025.43.16_suppl.11171.
- [34] Ashlee K Mitchell et al. Analysis of eligibility criteria in Alzheimer’s and related dementias clinical trials. *Scientific Reports*, 14:15234, 2024. doi: 10.1038/s41598-024-65767-x.
- [35] OpenAI. GPT-4o system card. arXiv:2410.21276, 2024.
- [36] OpenAI. GPT-4.1. OpenAI technical announcement, 2025.
- [37] Jimyung Park, Yilu Fang, Casey Ta, Gongbo Zhang, Betina Idnay, Fangyi Chen, David Feng, Rebecca Shyu, Emily R Schriver, and Chunhua Weng. Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *Journal of Biomedical Informatics*, 154:104649, 2024. doi: 10.1016/j.jbi.2024.104649. Maps EC to executable database queries.
- [38] Qwen Team. Qwen2.5 technical report. arXiv:2412.15115, 2024.
- [39] Qwen Team. Qwen3 technical report. arXiv:2505.09388, 2025.
- [40] Olesya Razuvayevskaya, Irene Lopez, Ian Dunham, and David Ochoa. Genetic factors associated with reasons for clinical trial stoppage. *Nature Genetics*, 56:1862–1867, 2024. doi: 10.1038/s41588-024-01854-z. 28,561 stopped trials; insufficient enrollment is #1 reason at 36.67%.
- [41] U.S. Food and Drug Administration. Enhancing the diversity of clinical trial populations: Eligibility criteria, enrollment practices, and trial designs: Guidance for industry. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enhancing-participation-clinical-trials-eligibility-criteria-enrollment-practices-and-2020>. Cross-cutting guidance: broaden EC, justify each exclusion.
- [42] Zifeng Wang and Jimeng Sun. Trial2Vec: Zero-shot clinical trial document similarity search using self-supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6377–6390. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.476. Self-supervised trial embeddings for similarity search.
- [43] Zifeng Wang et al. AutoTrial: Prompting language models for automated clinical trial design. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. LLM-based protocol generation without external evidence grounding.

- [44] Zifeng Wang et al. TrialPanorama: A large-scale, diverse, and fine-grained clinical trial database and benchmark, 2025. Large-scale trial database and benchmark tasks.
- [45] R. Patrick Xian, Noah R. Baker, Tom David, Qiming Cui, A. Jay Holmgren, Stefan Bauer, Madhumita Sushil, and Reza Abbasi-Asl. Robustness tests for biomedical foundation models should tailor to specifications. *npj Digital Medicine*, 8(1):557, 2025. doi: 10.1038/s41746-025-01926-2.
- [46] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packaged resources to advance general chinese embedding. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–650. ACM, 2024. doi: 10.1145/3626772.3657878. Introduces BGE embedding model family including bge-small-en-v1.5.
- [47] Faiez Zannad, Harriette G C Van Spall, et al. How to make cardiology clinical trials more inclusive. *Nature Medicine*, 30(10):2745–2755, 2024. doi: 10.1038/s41591-024-03273-3.
- [48] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623, 2024.
- [49] Mark Zimmerman, Caroline Balling, Iwona Chelminski, and Kristy Dalrymple. Have treatment studies of depression become even less generalizable? Applying the inclusion and exclusion criteria in placebo-controlled antidepressant efficacy trials published over 20 years to a clinical sample. *Psychotherapy and Psychosomatics*, 88(3):165–170, 2019. doi: 10.1159/000499917. 86.1% mean exclusion rate; worsening over time (91.4% for 2010–2014).

A Full Paper–Trial Matches

All 44 paper–trial matches with enrollment impact estimates, source papers, extracted directives, and LLM evaluation scores are presented in Table 3. For each match, the Match score (M, 1–5) assesses topical alignment between the paper’s recommendations and the trial’s eligibility criteria, and the Applicability score (A, 1–5) measures how directly the paper’s directive can be implemented in the trial. Scoring rubrics and prompt templates are provided in Appendix B.

B Scoring Prompts and Rubrics

This appendix documents the LLM scoring prompts, rubrics, thresholds, and filtering logic used throughout the RECITE pipeline. All prompts are stored in JSON configuration files and loaded at runtime. Full prompt templates are available in the supplementary materials.

B.1 Paper Relevance Scoring

Each retrieved paper is scored on three dimensions (1–5 each). Table 4 shows the relevance rubric (used for initial filtering) alongside the match and applicability rubrics (used for paper–trial pairing). Extraction confidence measures how actionable the paper’s recommendations are (1 = no actionable information; 5 = directly applicable threshold changes). Accrual ease measures how quantifiable the enrollment impact is (1 = vague or subjective; 5 = easily quantifiable biomarkers). Papers with relevance ≥ 3 proceed to directive extraction.

The scorer also extracts verbatim **change directive quotes** and **change rationale quotes**. High-quality matches (both scores ≥ 4) seed additional query generation. Default filtering: match ≥ 3 , sorted by match then applicability descending, maximum 10 matches per paper.

B.2 LLM Judge (Benchmark Evaluation)

The benchmark judge scores each (target, predicted) eligibility criteria pair on two scales: a **binary score** (0 = incorrect, 1 = correct) and an **ordinal score** on a 0–4 scale where 0 indicates no correspondence with the target, 1 indicates poor correspondence with major omissions, 2 indicates partial correspondence with notable gaps, 3 indicates good correspondence with minor differences, and 4 indicates the prediction is essentially identical to the target. The judge returns both scores independently; Bin Acc in the main table reflects the direct binary judgment, while $\% \geq 3$ reports the proportion of ordinal scores at or above 3. For efficiency, predictions are evaluated in batches, with the judge returning a JSON dictionary mapping instance indices to score pairs.

Table 4: Scoring rubrics for paper relevance (left), paper-trial match alignment (center), and applicability (right). Relevance is scored per-paper; match and applicability are scored per (paper, trial) pair.

Score	Relevance	Match	Applicability
1	Not relevant to EC or enrollment	No alignment; different conditions	Change implied, requires deep digging
2	Tangentially relevant; mentions trials broadly	Minimal; broad topic overlap only	Vague hints; may require full paper
3	Discusses EC but not central focus	Some alignment; similar condition, partial EC overlap	Mentioned but not overt; some interpretation needed
4	Focuses on EC, enrollment barriers, or participation	Good alignment; same/similar condition, substantial overlap	Directive present, minor interpretation needed
5	Directly about EC modification or broadening	Strong alignment across all dimensions	Overt in abstract; immediately applicable

B.3 Directive and Impact Extraction

The directive extraction prompt classifies each paper as containing exact directives (extractable from abstract), directives requiring full text, or no directives. Only papers with extractable directives proceed. The impact extraction prompt identifies numeric accrual estimates (percent increase or absolute patient counts) with supporting verbatim evidence. Key computed quantities are: `paper_pct_gain` (percent increase in eligible patients, e.g., 100 = doubling) and `scalar_gain` (estimated additional patients for a given trial = `enrollment` $\times$ `paper_pct_gain` / 100).

B.4 Query Generation

Three query generation prompts are used: (1) literature search queries for PubMed and Semantic Scholar, (2) trial matching queries for ClinicalTrials.gov derived from paper content, and (3) seed expansion queries generated from high-quality matches. All prompts receive previously issued queries to avoid repetition and return JSON arrays of query strings.

C Representative Directive Applications

Six examples illustrate how RECITE applies paper directives to matched trial eligibility criteria. Examples 1, 3, 4, 5 are actionable broadening directives (T/T in Table 3); Examples 2 and 6 (both Tsaur et al.) are F/F scoring-practice observations. The full set of 44 applications is available in the supplementary materials.

C.1 Example 1: ASCO-Friends Broadening (Durvalumab Trial)

Source: Harvey et al. [20] (Match 4/5, Applicability 5/5) **Trial:** NCT04163432 (Durvalumab, metastatic non-squamous NSCLC)

Directive: Allow patients with brain metastases, other malignancies, and CrCl ≥ 30 mL/min. **Accrual gain:** +100% (43 patients)

The trial’s renal criterion (eGFR ≥ 30) already satisfies the CrCl recommendation. The primary broadening removes exclusions for controlled brain metastases and prior non-lung malignancies.

C.2 Example 2: ECOG Scoring-Practice Observation (F/F, Prostate Cancer)

Source: Tsaur et al. [17] (Match 4/5, Applicability 5/5) **Trial:** NCT00485303 (Abiraterone, mCRPC)

Directive (extracted): Edit ECOG from 2 to 1 for prostate cancer patients. **Classification:** F/F

The source documents inter-rater ECOG scoring variability, not a protocol-level EC modification. Re-scoring ECOG 2 $\rightarrow$ 1 would expand matched trials (+233%) but does not constitute an actionable EC change.

C.3 Example 3: ASCO-Friends Broadening (T-DXd HER2 Trial)

Source: Harvey et al. [19] (Match 5/5, Applicability 5/5) **Trial:** NCT04739761 (Trastuzumab deruxtecan, HER2+ breast cancer with brain metastases)

Directive: Allow brain metastases, other malignancies, and CrCl >30 mL/min. **Accrual gain:** +100% (506 patients)

This trial already permits brain metastases (5/5 match). The primary broadening applies to the “other malignancies” exclusion and creatinine clearance threshold.

C.4 Example 4: Renal and Brain Metastasis Liberalization (KRAS G12C Trial)

Source: KRAS G12C multicenter study [33] (Match 3/5, Applicability 3/5) **Trial:** NCT01509937 (Phase I/II, renal dysfunction)

Directive: Liberalize criteria for renal dysfunction and brain metastases. **Accrual gain:** +25% (108 patients)

The moderate match (3/5) reflects a transplant rather than NSCLC context, but renal function threshold liberalization remains applicable. This illustrates cross-specialty applicability of broadening recommendations.

C.5 Example 5: HCC Performance Status Expansion (Belinostat Trial)

Source: Fung et al. [13] (Match 5/5, Applicability 5/5) **Trial:** NCT01273155 (Belinostat, solid tumors/lymphomas, Phase I/II)

Directive: Include ECOG 2 patients and limited liver dysfunction (Child-Pugh B7). **Accrual gain:** +143% (103 patients)

The trial already allows ECOG ≤ 2 . The primary broadening is expanding hepatic function criteria to include Child-Pugh B7 patients with limited liver dysfunction.

C.6 Example 6: ECOG Scoring-Practice Observation (F/F, Pan-Cancer)

Source: Tsauro et al. [29] (Match 4/5, Applicability 5/5) **Trial:** NCT03288987 (Stivant, metastatic colorectal cancer)

Directive (extracted): Edit ECOG 3 $\rightarrow$ 2 or 2 $\rightarrow$ 1 to increase matched trials. **Classification:** F/F

Pan-cancer companion to Example 2. Documents systematic ECOG overestimation (+246% matched trials, $p < 0.001$) rather than recommending protocol-level EC changes.