Nature Communications

https://doi.org/10.1038/s41467-026-70509-w

Article in Press

TrialMatchAI: an end-to-end AI-powered clinical trial recommendation system to streamline patient-to-trial matching

Received: 26 March 2025

Accepted: 27 February 2026

Cite this article as: Abdallah, M., Nakken, S., Georges, M. *et al.* TrialMatchAI: an end-to-end AI-powered clinical trial recommendation system to streamline patient-to-trial matching. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-70509-w

Majd Abdallah, Sigve Nakken, Mikaël Georges, Mariska Bierkens, Johanna Galvis, Alexis Groppi, Slim Karkar, Lana Meiqari, Maria Alexandra Rujano, Steve Canham, Rodrigo Dienstmann, Remond Fijneman, Eivind Hovig, Gerrit Meijer & Macha Nikolski

TrialMatchAI: An End-to-End AI-powered Clinical Trial Recommendation System to Streamline Patient-to-Trial Matching

Authors:

Majd Abdallah[1,2,#], Sigve Nakken[3,4,5], Mikaël Georges[1,2], Mariska Bierkens[6], Johanna Galvis[1,2], Alexis Groppi[1,2], Slim Karkar[1,2], Lana Meiqari[6], Maria Alexandra Rujano[7], Steve Canham[7], Rodrigo Dienstmann[8,9], Remond Fijneman[6], Eivind Hovig[3,5], Gerrit Meijer[6], Macha Nikolski[1,2,#]

Affiliations:
[1] University of Bordeaux, CNRS, IBGC UMR 5095, 146 Rue Léo Saignat, 33000 Bordeaux, France.
[2] University of Bordeaux, Bordeaux Bioinformatics Center, 146 Rue Léo Saignat, 33000 Bordeaux, France.
[3] Department of Tumor Biology, Institute of Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway.
[4] Centre for Cancer Cell Reprogramming, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, 0379 Oslo, Norway.
[5] Department of Informatics, University of Oslo, 0316 Oslo, Norway.
[6] Department of Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands.
[7] European Clinical Research Infrastructure Network (ECRIN), Boulevard Saint Jacques 30, 75014, Paris, France.
[8] Oncology Data Science (ODysSey) Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain.
[9] University of Vic - Central University of Catalonia, Barcelona, Spain

 # Corresponding authors: Majd Abdallah (abdallahmajd7@gmail.com), Macha Nikolski (macha.nikolski@u-bordeaux.fr)

Abstract

Patient recruitment remains a major bottleneck in clinical trials, calling for scalable and automated solutions. We present TrialMatchAI, an AI-powered recommendation system that automates patient-to-trial matching by processing heterogeneous clinical data, including structured records and unstructured physician notes. Built on fine-tuned, open-source large language models (LLMs) within a retrieval-augmented generation framework, TrialMatchAI ensures transparency and reproducibility and maintains a lightweight deployment footprint suitable for clinical environments. The system normalizes biomedical entities, retrieves relevant trials using a hybrid search strategy combining lexical and semantic similarity, re-ranks results, and performs criterion-level eligibility assessments using medical Chain-of-Thought reasoning. This pipeline delivers explainable outputs with traceable decision rationales. In real-world validation, 92% of oncology patients had at least one relevant trial retrieved within the top 20 recommendations. Evaluation across synthetic and real clinical datasets confirmed state-of-the-art performance, with expert assessment validating over 90% accuracy in criterion-level eligibility classification - particularly excelling in biomarker-driven matches. Designed for modularity and privacy, TrialMatchAI supports Phenopackets-standardized data, enables secure local deployment, and allows seamless replacement of LLM components as more advanced models emerge. By enhancing efficiency, interpretability and offering lightweight, open-source deployment, TrialMatchAI provides a scalable solution for AI-driven clinical trial matching in precision medicine.

Introduction

The advancement of personalized medicine relies heavily on clinical trials, which rigorously evaluate the efficacy and safety of novel therapeutic strategies and validate actionable biomarkers[1,2]. Yet, a critical bottleneck persists: the timely and efficient recruitment of eligible patients. This challenge not only delays access to potentially life-saving treatments, but also leads to significant resource inefficiencies, hindering the translation of research into clinical practice[3]. Furthermore, only a small fraction of eligible patients are enrolled in clinical trials, despite the potential benefits. Addressing this problem requires scalable and efficient patient-trial matching solutions to accelerate trial completion and ensure that research findings are approved for clinical practice in a timely manner[4].

Traditionally, patient-trial matching relies on a labor-intensive manual review of patient records and trial eligibility criteria, often performed by multidisciplinary teams[5,6]. Such manual screening is a major operational bottleneck and studies report tens of minutes to multiple hours per patient, and 3.4 - 8.8 hours per enrolled patient when end-to-end screening/enrollment tasks are counted[7]. Complex criteria reviews can take 10 minutes to >2 hours per trial–per–patient. This process is inefficient, unscalable, and prone to missed enrollment opportunities, particularly in high-stakes areas like pediatric oncology, where timely trial access is critical[8]. The unstructured nature and considerable volume of trial eligibility criteria further exacerbate these challenges, making manual approaches increasingly unsustainable and underscoring the urgent need for efficient automation[9–12].

Early automation efforts relied on rule-based logic and probabilistic systems, which, while effective for structured scenarios, struggled with the semantic diversity and contextual nuances of clinical text. Deep learning approaches have addressed some of these limitations by improving feature extraction and handling sequential dependencies inherent in clinical texts text[13–18]. However, these models often rely on large, well-annotated datasets, which are scarce in the biomedical domain, limiting their scalability and generalizability[19,20]. ~~See Supplementary Materials "Literature Review" section H for additional details.~~ For example, structured rule/ontology-based systems such as MatchMiner[21] encode eligibility logic explicitly (especially genomic criteria), while neural embedding/entailment approaches such as COMPOSE[13] and DeepEnroll[15] learn patient–trial representations from text; these families respectively trade precision in structured settings against broader language coverage, but both remain limited by manual schema maintenance or the need for large labeled datasets.

Recent advances in natural language processing (NLP), particularly through Large Language Models (LLMs), have opened new avenues for processing and interpreting complex clinical texts. Pre-trained LLMs excel at capturing long-range dependencies and contextual relationships, allowing for the generation of clinically meaningful embeddings[22–27]. LLMs have even shown impressive ability to match patients to clinical trials without being explicitly trained for this task, achieving results comparable to human experts[12,22,28,29]. However, most existing LLM-based trial matching systems, such as TrialGPT[12], rely heavily on proprietary, API-driven models, creating barriers related to cost, accessibility, reproducibility, and, critically, patient data privacy and regulatory compliance (e.g., GDPR, HIPAA[30]). The dependence on black-box, closed-source solutions also hinders transparency and prevents other researchers from building upon or adapting these models for specific clinical needs. In parallel, open or hybrid LLM approaches, such as LLM-Match[31] and Panacea[32], aim to retain strong semantic matching while improving reproducibility and transparency. Yet, most LLM-based approaches remain domain-constrained

or task-specific ~~(e.g., clinical trial matching or summarization)~~ rather than fully end-to-end systems: for example, TrialGPT[12] primarily targets trial ranking, LLM-Match[31] focuses on pairwise patient–trial eligibility classification, and Panacea[32] evaluates classifier-style matching on curated pairs. This leaves local deployment, criterion-level explanations for clinical review, and reproducible benchmarking under matched candidate pools insufficiently addressed in a single workflow. ~~, and their reported results often vary due to non-standardized evaluation protocols and benchmark differences.~~ We further discuss these limitations in the Supplementary Note 1 and present a qualitative and quantitative comparisons of LLM-based methods in Supplementary Data 1, 2, and 3.

To address these limitations, we introduce TrialMatchAI, a fully open-source, locally deployable general-purpose clinical trial recommendation system designed to ensure transparency, security, and unrestricted research accessibility while eliminating reliance on proprietary LLMs. This ensures compliance with regulatory frameworks, promotes interpretability in patient-trial matching decisions, and supports continuous model improvement as new data becomes available. Unlike previous systems that target a single sub-task (e.g. ranking or classification) or require external API-based models, TrialMatchAI enables complete control over the trial-matching process, ensuring compliance with stringent data privacy regulations. Our evaluation of performance and validation includes standard benchmarks and a real-world oncology cohort with documented trial orientations.

Built on a Retrieval-Augmented Generation (RAG) framework[33] and fine-tuned for medical reasoning, TrialMatchAI combines contextual understanding and explainability in eligibility classification, overcoming the limitations of existing rules-based and machine learning-based trial matching solutions. Importantly, the system demonstrated strong applicability in oncology, where patient-trial matching is often complicated by the need to integrate and interpret diverse clinical, molecular, and genetic data, including biomarker expression profiles and genomic mutations. This is made possible by a dedicated LLM-driven data processing module that extracts, normalizes, and standardizes biomolecular information, such as genes, proteins, and mutations from both patients records and clinical trial eligibility criteria. Furthermore, TrialMatchAI is designed for interoperability, seamlessly integrating with electronic health record (EHR) systems via standardized data exchange formats such as Phenopackets[34]. Finally, the modular design enables easy adaptation to new LLM architectures and domain-specific fine-tuning, ensuring that the system remains up to date with the latest biomedical advancements.

The framework was evaluated using synthetic datasets, comprising the commonly-used benchmarks from the TREC clinical trials challenge[35,36] (years 2021 and 2022), and a custom-built “Ideal Candidates” dataset. Moreover, we leveraged a real-world cohort of 52 cancer patients from the Netherlands Cancer Institute (NKI)[37]. Although this real-world evaluation was performed for oncology trial orientation, the architecture is disease-agnostic. Expert assessments were also conducted on 1050 patient-criterion pairs, validating the system’s high accuracy in criterion-level eligibility classification. Our results on the synthetic TREC datasets demonstrate that TrialMatchAI retrieves >90% of relevant trials within the top-ranked 3% of a broad search space containing >26,000 trials in each dataset. Results on the real cancer patients dataset further support these results, showing that 92% of patients had a relevant trial retrieved within the top 20 recommendations. Finally, expert evaluation of criterion-level matching on both synthetic and real patient datasets shows that TrialMatchAI’s medical Chain-of-Thought (CoT) model[38] achieves over 90% accuracy in criterion-level classification and explanation generation. These results show that TrialMatchAI, outperforming existing AI-based tools that rely on significantly larger, proprietary GPT (Generative Pre-trained Transformer) models.

In summary, TrialMatchAI is a scalable, privacy-preserving patient-trial matching system, showing high performance in oncology settings while remaining disease-agnostic by design. It integrates

multi-modal data, supports local deployment, and leverages fine-tuned LLMs within a RAG framework to improve recruitment. The system provides clinical decision support and could potentially be integrated with existing clinical decision support systems, offering ranked recommendations for investigator review. This paper details its architecture, evaluation, and performance in real-world and synthetic benchmark settings.

Results

TrialMatchAI: A Modular AI System for Patient-Trial Matching

Clinical trial eligibility criteria are complex and often formatted as unstructured free texts, requiring a system that can efficiently process heterogeneous and semantically-rich patient data. TrialMatchAI addresses this challenge by leveraging a suite of open-source and fine-tuned, Large Language Models (LLMs) within a Retrieval-Augmented Generation (RAG) framework[33]. RAG enables the system to anchor its reasoning in retrieved trial information, improving both accuracy and transparency in patient-trial matching. TrialMatchAI's modular design allows for efficient text parsing, embedding, classification, and re-ranking, ensuring accurate and context-aware matching. Moreover, the flexible design of TrialMatchAI enables seamless integration of new models and optimization strategies, ensuring adaptability to evolving clinical needs.

~~System Overview and Core Components~~

TrialMatchAI is designed to handle heterogeneous clinical data, including structured attributes (e.g., age, sex, primary diagnosis, lab results) and unstructured sources (e.g., prior treatments, physician notes, pathology reports). To ensure interoperability with hospital systems and external research databases, the system adopts the Phenopackets exchange format[34], enabling standardized representation of patient data across diverse sources. Notice that this does not ensure full operational EHR interoperability and healthcare sites would need to use a local adapter (e.g., from FHIR[39]) to produce Phenopackets with the relevant patient information (see e.g. this implementation guide[40] for technical details). However, Phenopackets has the capacity to integrate both structured and unstructured data, which seamlessly integrates with LLM-based semantic processing of clinical narratives.

To identify relevant clinical trials for a patient, TrialMatchAI employs a hybrid search and retrieval strategy, combining lexical search[41,42] for keyword-based matching with vector search[43] using dense embeddings to capture deeper semantic relationships. Following initial retrieval, LLM-based re-ranking refines the list by prioritizing trials based on criterion-level relevance. A fine-tuned re-ranking LLM[44] assesses the relevance of candidate trials' eligibility criteria to the patient on a criterion-by-criterion basis and aggregates the criterion-level scores to trial-level scores, ultimately re-ranking the trials from the most to least applicable to the patient's profile. Finally, precise eligibility classification is performed using a fine-tuned Chain-of-Thought (CoT) reasoning model[38,45], which classifies inclusion and exclusion criteria and generates a justification for each decision. This structured pipeline (see Figure 1) optimizes patient-trial matching while maintaining explainability in decision-making, and consists of four key levels:

1. Data Ingestion and Preprocessing: Both structured trial metadata (e.g., ClinicalTrials.gov XML files) and patient records are processed, and terminology is standardized through Named Entity Recognition (NER) and entity normalization[46–50] followed by vector embedding to enable semantic search[51].
2. Candidate Trial Retrieval: For a given patient, a broad pool of relevant trials is obtained using a combination of BM25 lexical search with k-NN vector search (semantic retrieval) via Elasticsearch[41,42,52].

3. Re-Ranking for Criterion-Level Relevance: Trial prioritization is performed by a fine-tuned Gemma-2-2b[44] model based on the applicability and relevance of eligibility criteria for the patient's profile.
4. Eligibility Classification and Final Ranking: A fine-tuned Phi-4 model performs criterion-level classification using medical Chain-of-Thought (CoT)[38,45] reasoning, generating explanations for inclusion and exclusion decisions. A final scoring step prioritizes trials that best satisfy inclusion criteria while minimizing exclusion violations.
5. Table 1 provides an overview of the core components and their roles in the system; further details on model fine-tuning, candidate retrieval mechanisms, re-ranking, and criterion-level eligibility assessment are provided in the Materials and Methods section, as well as in the Supplementary Methods sections 2, 3, 4, and 5. ~~Beyond accuracy, Supplementary section I ('Eligibility Screening: Practice Overview and TrialMatchAI Runtime') describes the typical site-agnostic screening workflow and identifies where TrialMatchAI may reduce investigator effort.~~

Because TrialMatchAI is intended to support real-world trial screening, we next report its end-to-end runtime and place it in the context of the typical manual workflow. Patient–trial matching is typically an iterative, manual workflow[53,54]: (i) a clinician or coordinator flags a candidate patient; (ii) relevant EHR data are compiled (diagnosis, biomarkers, prior treatment, comorbidities); (iii) trial registries and institutional lists are searched using keywords and filters; (iv) obvious mismatches are discarded in a first-pass triage; (v) eligibility criteria are read in detail and rationales documented; and (vi) shortlists are reviewed in multidisciplinary meetings and screening documentation is completed (see Supplementary Methods for additional details).

On our test hardware, TrialMatchAI generates a ranked shortlist with criterion-level rationales in ~4.6 minutes per patient on average (mean: 2.7s first-level retrieval, 92.7s second-level retrieval + re-ranking, 180.7 s Chain-of-Thought reasoning). Manual eligibility screening typically requires 10 minutes to >2 hours per trial[8], and clinicians may review ~42 trials per patient to reproduce historical matches[9], with ~32% of coordinator time devoted to eligibility screening tasks (reported in a pediatric ED time-and-motion study[55]). By producing a ready-to-review top-K shortlist with explanations in minutes, TrialMatchAI is expected to reduce time spent on search, first-pass triage, and initial eligibility review, shifting effort toward confirming a smaller set of candidates. Exact time savings remain site-specific and should be quantified prospectively during deployment. Supplementary Methods section 8 ('Eligibility Screening: Practice Overview and TrialMatchAI Runtime') identifies where TrialMatchAI may reduce investigator effort in a typical site-agnostic workflow.

~~Benchmarking TrialMatchAI: Synthetic Patient Evaluation~~ TrialMatchAI Achieves High Recall and Accurate Ranking on Synthetic Benchmarks

To evaluate TrialMatchAI, we utilized a comprehensive dataset comprising both in-house and publicly available resources. The in-house dataset ("Ideal Candidates" dataset) consists of synthetic patient profiles generated from a randomly sampled set of cancer-related clinical trials, specifically focusing on trials with long and complex eligibility criteria. The publicly available dataset includes synthetic summaries from the Clinical Trials (CT) tracks of the Text Retrieval Conference (TREC) for the years 2021[35] and 2022[36], both widely used benchmarks for assessing patient-trial matching systems[12]. Additionally, an expert evaluation was conducted to assess TrialMatchAI's criterion-level eligibility classification and its generated explanations. Performance was measured using multiple metrics, including accuracy, recall, normalized discounted cumulative gain (nDCG) at top-k (ndcg@5, ndcg@10, ndcg@20), and precision at top-k (p@5, p@10, p@20). See Supplementary Methods section 5 for metrics definitions.

We additionally benchmarked TrialMatchAI against LLM-Match[31] and Panacea[32] under matched conditions (Supplementary Methods section 9). TrialMatchAI outperformed the non-fine-tuned LLM-Match baseline on TREC2021/2022 (e.g., TREC2022 nDCG@10 0.75 vs 0.40) and showed moderate agreement on Panacea's shared pairwise subset after label mapping (Supplementary Figure 4). ~~Furthermore, a like-for-like ranking performance comparison against LLM-Match[31] on TREC datasets and a complementary classification baseline task with Panacea[32] are provided in Supplementary section J.~~ These comparisons aim to showcase the advantages of TrialMatchAI's architecture and modelling in both retrieval accuracy and robustness across benchmark settings, highlighting its ability to generalize beyond benchmark-specific tuning and maintain superior precision–ranking balance under varying label distributions.

~~Validating AI-based Matching: The "Ideal Candidates" Dataset~~

The "Ideal Candidates" dataset (see Materials and Methods, section 4.1) serves as an initial proof of concept, demonstrating TrialMatchAI's ability to accurately align patients with their optimal clinical trial, assuming a perfect match exists. Figure 2 illustrates the distribution of ground truth trial rankings for the 100 ideal candidates. The vast majority (95%) had their assigned clinical trial ranked within the top two matches, with 92 out of 100 patients having their ground truth trial among the highest-ranked recommendations. A small subset (5%) had their ground truth trial ranked lower; however, no trial fell beyond the 9th position. This ranking discrepancy likely stems from the eligibility of some synthetic patients for multiple highly relevant trials beyond their designated ground truth trial. In such cases, TrialMatchAI identified additional trials that were equally or even more suitable based on patient characteristics, leading to slight variations in rankings. Rather than signaling a failure, this outcome highlights TrialMatchAI's potential to surface not only the expected match, but also a broader set of viable trial options: an ability that could be valuable in real-world clinical decision-making. The ground truth trial assignments and the top 10 ranked trials for each ideal candidate are provided in Supplementary Data 4.

~~High Recall and Precision with Accurate Trial Prioritization in TREC Benchmarks~~

To assess TrialMatchAI's robustness in handling diverse and complex patient cases, we evaluated its performance on the TREC 2021 and 2022 Clinical Trials (CT) datasets[35,36]. These datasets contain synthetic patient case descriptions resembling those found in admission notes. TREC2021 consists of 75 patient cases, while TREC2022 includes 50 patient cases. Each patient is associated with a list of trials labeled as "Irrelevant" (unrelated to the patient), "Excluded" / "Ineligible" (patient has the condition, but does not qualify), or "Eligible" (patient qualifies for enrollment).

TrialMatchAI's hybrid retrieval approach efficiently retrieves an initial subset of candidate trials from over 26,000 in each TREC dataset, reducing the search space for ranking and final eligibility assessment by more than 95% and achieving over 90% recall at just 3% of the total documents for both TREC2021 and TREC2022. This means that the majority of relevant trials for most patients are retrieved early, reducing the risk of missing eligible matches.

Figure 3A presents the recall performance of TrialMatchAI across varying retrieval sizes on the TREC datasets. Recall improves with an increasing number of retrieved trials, stabilizing around 90% at approximately 500 retrieved trials. This suggests that a cut-off at this retrieval size balances high recall with computational efficiency. The consistent recall performance across datasets highlights the model's robustness in generalizing across different patient cohorts.

Figures 3C and 3D illustrate TrialMatchAI's ranking performance on the TREC2021 and TREC2022 datasets. Ranking quality is evaluated using Normalized Discounted Cumulative Gain (nDCG) and Precision (P) at 5, 10, and 20 trials. We interpret recall@K as exhaustiveness

(coverage of relevant trials) and Precision@K / nDCG@K as appropriateness (prioritization quality), reflecting the clinical need to keep coverage high while surfacing the most suitable trials first. Across both TREC datasets, TrialMatchAI achieves a median nDCG@5 of 0.74 (TREC2021) and 0.82 (TREC2022) and a median nDCG@10 and nDCG@20 of 0.75 (both), meaning the system consistently ranks the most relevant trials closer to the top. Similarly, median precision scores remained relatively high across different cut-off points. The tool achieves a median p@5 of 0.8 (for both TREC2021 and 2022), median p@10 of 0.77 (TREC2021) and 0.72 (TREC2022), and median p@20 of 0.7 (TREC2021) and 0.62 (TREC2022). The consistency between TREC2021 and TREC2022 results underscores the model's robustness across different patient distributions.

While the metrics reported by other LLM-based solutions, notably TrialGPT[12], differ slightly in how they are aggregated, benchmark comparisons indicate that TrialMatchAI achieves highly competitive performance using significantly smaller, open-source models. TrialGPT's highest aggregated average nDCG@10 is 0.7275 and p@10 is 0.6688 across all its evaluated datasets. In comparison, TrialMatchAI achieves an average nDCG@10 of 0.7232 and a higher average p@10 of 0.6865 across the TREC2021 and 2022 benchmarks (with even stronger performance reflected in median values, see hereinabove). Additionally, TrialMatchAI outperforms all previously top-ranked systems from the official TREC2021 and 2022 challenges. The top ranking model in TREC2021, TD-MINER, achieved a mean nDCG@10 of 0.715 and a mean p@10 of 0.5760, whereas the top ranking model in TREC2022, h2oloo, achieved a mean nDCG@10 of 0.6125 and a mean p@10 of 0.5080.

TrialMatchAI Successfully Matches Real Cancer Patients to Biomarker-Driven Trials

Unlike general clinical trial matching, this evaluation focuses on biomarker-driven oncology trial selection, where cancer patient eligibility depends on clinical characteristics and the identified biomarkers. We incorporate real-world data from a metastatic cancer cohort derived from the Whole Genome Sequencing Implementation in Standard Diagnostics for Every Cancer Patient study (WIDE) conducted by the Netherlands Cancer Institute (NKI)[37]. See Figure 4 for baseline statistics of the selected sub-cohort of 52 patients, which had a mean age of 63.6 years (±10.3) and a sex distribution of 46.2% male and 53.8% female.

~~Trial Retrieval and Ranking Performance: High Recall and Prioritization Accuracy~~

Table 2 presents the evaluation results, demonstrating TrialMatchAI's effectiveness in retrieving and ranking relevant clinical trials for cancer patients. Overall Recall, representing the percentage of patients for whom at least one relevant trial was identified, increased from 84.6% within the top five recommendations (Top-5) to 92.3% within the top twenty (Top-20). This demonstrates TrialMatchAI's ability to consistently retrieve relevant trials while ensuring that the number of suggested trials remains clinically practical for review by oncologists.

Beyond recall, we assessed how well TrialMatchAI prioritizes relevant trials. The Mean Reciprocal Rank (MRR) measures how highly the first correct trial is ranked, with higher values indicating that relevant trials appear near the top of the recommendation list. The Mean Average Rank (MAR) reflects the typical rank position of relevant trials across all retrieved trials. TrialMatchAI's consistent MRR across different top-k values indicates that when a relevant trial is retrieved, it is accurately prioritized near the top of the recommendation list. This ensures that oncologists can identify the most promising trials quickly, reducing manual screening time and expediting patient enrollment.

Expert Evaluation of Criterion-Level Eligibility Classification

To assess TrialMatchAI's accuracy in evaluating individual eligibility criteria, we conducted an expert review of 950 randomly selected patient-criterion pairs derived from synthetic patients in the TREC 2021 and 2022 datasets. The fine-tuned Phi-4 CoT reasoning model correctly classified 88.8% of inclusion criteria as "Met" and 91.1% as "Not Met". Additionally, it accurately identified 94.2% of "Unclear" criteria and 97.4% of "Irrelevant" criteria. For exclusion criteria, the model achieved an 89.7% accuracy in detecting "Violated" criteria while correctly classifying 96.9% of "Not Violated", 98.2% of "Unclear," and 97.3% of "Irrelevant" cases.

Building on our evaluation with synthetic TREC data, we further assessed the system's performance in a real-world clinical setting, focusing on molecular biomarker-driven eligibility criteria in oncology. To this end, we conducted an expert review of 100 biomarker-related inclusion criteria classified as "Met" by the model. These criteria were drawn from the top 20 trials recommended by TrialMatchAI for patients in the WIDE cohort. To evaluate the system's ability to match molecular biomarker criteria, we constructed patient-criterion pairs. The criterion component consisted of the selected biomarker-related inclusion criteria from these top trials. The patient component was a semi-synthetic profile incorporating biomarker details derived from the WIDE cohort. Specifically, molecular biomarker data were extracted from original Molecular Tumor Board (MTB) reports, then completely rephrased into new, decontextualized statements using the Phi-4 model. This process removed all identifying information (e.g., patient ID, age, sex) while preserving evaluative utility. This approach generated patient profiles that accurately reflected real molecular characteristics while ensuring data privacy. TrialMatchAI correctly classified 91% of these biomarker-driven inclusion criteria as "Met", demonstrating its effectiveness in capturing biomolecular eligibility constraints. Full results of criterion-level expert evaluations are provided in Figure 3B and Supplementary Data 5 and 6.

Component Contributions: Hybrid Retrieval, Re-ranking, and CoT Improve Precision (Ablation Study)

To assess the individual contributions of each component within TrialMatchAI and to separate architectural effects from base-model capability, we conducted an ablation study evaluating the impact of retrieval, reranking, and reasoning modules on overall trial-matching performance. The analysis involved either modifying or completely removing one or more of the core components and re-running the matching on the TREC 2021 dataset, namely:

- Named entity recognition and normalization (NER)
- Data retrieval method (BM25 text-based or vector search instead of a hybrid search approach),
- LLM-based re-ranking,
- CoT reasoning.

Figures 5A and 5B report precision and nDCG at cutoffs 5, 10, and 20 across ablation configurations. In the precision panel (top), the Full pipeline shows its clearest gains at broader cutoffs: p@10 and p@20 improve most, with the largest drops observed when removing the LLM-based reranker or NER, underscoring their role in elevating relevant trials deeper in the ranked list. Removing CoT produces smaller but consistent declines, suggesting a refinement effect rather than a primary driver of precision. At p@5, effects are attenuated and several variants approach the full pipeline, consistent with a pattern of top-slot saturation across different variations of TrialMatchAI's pipeline.

In the nDCG panel, performance is largely stable across ablations. At nDCG@5 and nDCG@10, several variants show slightly higher central tendency than the Full configuration, but none of these differences reach statistical significance ~~(no significance bars; all p ≥ 0.05)~~. For nDCG@20, differences remain modest but trend in favor of the Full pipeline; removing the reranker or NER yields small, sometimes statistically significant declines ($p < 0.05$), whereas other ablations remain broadly comparable and within variance. Across ablations, Precision@k drops most when removing the re-ranker or NER; CoT removals show smaller but consistent decreases. By contrast, nDCG@k remains largely stable at $k=5,10$ and shows only modest declines at $k = 20$ when the re-ranker or NER are removed (some $p < 0.05$ at nDCG@20).

For nDCG@20, differences remain modest but trend in favor of the Full pipeline; removing the reranker or NER yields small, sometimes statistically significant declines (for example, nDCG@20 (No Rerank) = 0.591 ± 0.233, p = 1e-3, two-sided Mann-Whitney test, p-value adjusted by False Discovery Rate) ~~(p < 0.05)~~, whereas other ablations remain broadly comparable and within variance. Across ablations, Precision@k drops most when removing the re-ranker or NER; CoT removals show smaller but consistent decreases. By contrast, nDCG@k remains largely stable at $k = 5,10$ and shows only modest declines at $k=20$ when the re-ranker or NER are removed ~~(some p < 0.05 at nDCG@20)~~.

Discussion

Clinical trial recruitment remains a major bottleneck in the development of new therapies. Patients, particularly those with rare or complex cancers, often struggle to find suitable trials. To address this, we developed TrialMatchAI, an open-source, modular system leveraging fine-tuned open-source LLMs for precise patient-trial matching. The system's ultimate function is in clinical decision-support, providing ranked recommendations to assist investigators**.** Our evaluations demonstrate state-of-the-art performance, comparable to systems that rely on GPT-based proprietary models such as TrialGPT[12], while addressing their limitations in terms of cost, transparency, and privacy. Notably, TrialMatchAI achieves this level of performance using lightweight, open-source models, a significant advancement given the computational demands and expense typically associated with high-performance LLMs. Unlike API-dependent systems, which require external data transfer, TrialMatchAI is designed for secure operation via local deployment within hospital infrastructures. Its modular, open-source architecture makes it a strong candidate for real-world clinical integration, offering a lightweight and privacy-preserving alternative to proprietary systems for patient-trial matching.

TrialMatchAI's architecture is designed for scalability, explainability, and interoperability, and for local deployment to ensure regulatory compliance with the European Health Data Space (EHDS), GDPR, and HIPAA[30]. In production and deployment settings, TrialMatchAI would need to implement periodic corpus updates to seamlessly incorporate new trials and status updates. Notice also that details of interoperability with production EHRs (e.g., FHIR workflows) are not yet considered in the scope of the TrialMatchAI project and should be handled by hospital-specific adapters to Phenopackets. Expert validation on over 1,000 patient-criterion pairs, including molecularly-driven cases, confirmed an average accuracy exceeding 90% in eligibility assessments, demonstrating the system's robust AI-driven reasoning. The model also achieves over 90% recall within the top 3% of trials in TREC datasets, effectively prioritizing relevant options while filtering out irrelevant ones. A hybrid retrieval approach, integrating LLM-based re-ranking and medical Chain-of-Thought (CoT) reasoning, ensures accurate candidate selection, with nDCG and precision at top-k metrics indicating state-of-the-art performance. Finally, a real-world evaluation on biomarker-driven data from 52 metastatic cancer patients of the WIDE study

demonstrated that 92% of patients had relevant trials successfully retrieved within the top 20 results (Table 2), validating TrialMatchAI's practical utility. The top 20 threshold is clinically meaningful, aligning with the oncologists' workflows, where reviewing a manageable number of trials is crucial for informed decision-making.

Despite integrating a fine-tuned CoT model within a RAG framework, TrialMatchAI, like all systems based on LLMs, can be susceptible to confabulations (i.e., hallucinations) and misclassifications. However, based on the expert evaluation of criterion-level classifications and explanations, confabulation occurred in less than 1% (9 out of 950 pairs, see Supplementary Data 2) cases, where the reasoning model generated an explanation that did not correspond to the provided patient's information. To address this, future work should focus on developing robust flagging and monitoring mechanisms that allow clinicians to review AI-generated recommendations and report misclassifications, enabling continuous system improvement. Additionally, integrating an Agentic Workflows architecture[56] where AI agents dynamically collaborate to verify and refine outputs, could provide an extra layer of oversight. This approach may help reduce hallucinations, offer more contextual information to the reasoning model, and ultimately enhance trustworthiness. Another challenge is balancing computational efficiency with model size, as proprietary GPT-based models still outperform open-source alternatives in inference speed. One potential solution is knowledge distillation[57,58], where smaller, optimized models are trained to retain the accuracy of larger models while improving computational efficiency. Finally, incomplete patient records can impact accuracy. To mitigate this, collaborative filtering techniques[59], which leverage insights from patients with similar clinical and molecular profiles, could help fill gaps in missing data and improve recommendation reliability.

While general LLM capabilities have advanced since earlier TREC benchmarks, our ablation analysis shows that TrialMatchAI's gains are not simply due to a stronger backbone model. The specialized modules (hybrid retrieval, fine tuned reranking, chain of thought reasoning) consistently improved Precision@k, while nDCG@k remained largely unchanged. Taken together, the ablations indicate an architectural separation of roles: re-ranking mainly governs relative ordering (nDCG shifts when it is removed), whereas NER/CoT primarily reduce false positives, lifting Precision@k. In other words, upstream filtering cleans the candidate set, and the re-ranker preserves ordering among remaining items. Although a stronger reasoning LLM may raise position-sensitive metrics such as nDCG, the ablations isolate concrete module contributions that bring more relevant trials into view. Future work will pair these modules with larger LLMs to evaluate potential nDCG gains. Taken together, the ablations indicate that biomedical NER-powered retrieval and the LLM-based reranker are the primary drivers of precision, i.e., the pipeline design adds clinical relevance beyond the base model alone (see Figure 5).

At present, TrialMatchAI supports only the patient-to-trial workflow. The pipeline's components are, however, bidirectionally applicable. Training and validating a trial-to-patients mode would require large, labeled patient cohorts. Unlike trials (readily sourced from ClinicalTrials.gov), patient-level corpora are scarce: public benchmarks like TREC are small, and institutional data are private and governed. We therefore leave this for future work under appropriate on-site datasets and oversight. Looking ahead, the European Health Data Space, under suitable governance, may facilitate access to larger on-site or federated patient cohorts suitable for developing and validating a trial-to-patients mode.

Overall, TrialMatchAI has the potential to significantly advance AI-driven personalized medicine by streamlining patient recruitment to clinical trials. By combining high accuracy, interpretability, and privacy-preserving local deployment, TrialMatchAI sets a new standard for AI-driven clinical trial matching, enabling real-world clinical adoption within secure hospital environments and research institutions.

Methods

Ethical compliance

All analyses were performed on de-identified data and in accordance with relevant ethical regulations. The WIDE study (Netherlands Cancer Institute) was approved by the Medical Ethical Committee of the Antoni van Leeuwenhoek/Netherlands Cancer Institute, Amsterdam (approval NL68609.031.18; 10 April 2019), and all participants provided written informed consent. Sex and gender were not considered as factors in the study design or analyses; where recorded, sex was used only as part of the patient profile input/descriptive metadata and no sex- or gender-stratified analyses were performed.

Materials

~~Clinical Trials Data~~

To support patient-trial matching, we utilized publicly available metadata from ClinicalTrials.gov, the largest registry of clinical trials. This dataset includes eligibility criteria, study condition(s), study design, and intervention details. Our database included over 100,000 clinical trials spanning from 1998 to 2024, ensuring a comprehensive and up-to-date resource for matching patients with relevant trials (all trial identifiers are provided in the Supplementary Data 7).

From this corpus, we retained interventional oncology trials with non-empty eligibility sections, forming the global cancer panel used across most TrialMatchAI experiments. To identify cancer-relevant trials, we mapped the condition terms from ClinicaTtrials.gov records to a controlled vocabulary of cancer phenotypes. This vocabulary was constructed using the OncoTree ontology (v2021_02)[60] as a starting point and expanded with additional terms from the phenOncoX project (https://github.com/sigven/phenOncoX). Trials were classified as cancer-related if they included at least one condition term matching this curated list. Additionally, a small subset of the cancer-related trials (*n = 136*) was obtained solely from the Dutch public repository Centrale Commissie Mensgebonden Onderzoek (CCMO)[61], which provides information on medical research involving human subjects in the Netherlands.

For TREC-based evaluation, we utilized 26,149 trials for TREC2021 dataset and 26,581 trials for TREC2022 dataset (see Table 3). For the 'Ideal Candidates' evaluation, the complete trials corpus included 61,731 interventional oncology trials. For the real-world cohort at the Netherlands Cancer Institute (NKI), the search space consisted of 217 predefined, molecularly driven Dutch oncology trials.

Supplementary Figures 1 and 2 (In Supplementary Methods) provide descriptive statistics for the curated and parsed cancer-related clinical trials in our dataset, including the distributions of cancer types, trial phases, intervention categories, recruitment status, and geographical locations.

~~Synthetic "Ideal Candidates" Dataset~~

In order to assess the model's ability to uncover relevant trials assuming that a perfect match exists for a given trial, we generated the "Ideal Candidates" dataset. It comprises 100 synthetic cancer patient profiles, each generated using the GPT-4o-mini model[62] and manually curated for accuracy. Each patient was generated from a randomly selected clinical trial from our trial database based on specific criteria: the trial had to be cancer-related, interventional, fairly recent (starting after 2014), and include an eligibility criteria section exceeding 500 words to ensure sufficient complexity and comprehensiveness. These selected trials were then processed via an API request to GPT-4o-mini (see Supplementary Methods section 3.4 for the prompt that was used), which was instructed to generate a perfectly matching patient profile for each trial—fully satisfying all inclusion criteria while explicitly avoiding any exclusion criteria. Check the "Code and

Data Availability" section for information on accessing the supplementary data dedicated for the generated patient summaries and the patient-trial matching results.

~~Synthetic TREC Clinical Trial Benchmark Datasets~~

The second synthetic dataset consists of patient summaries from the TREC2021 and TREC2022 Clinical Trials (CT) tracks[35,36]], which include 75 and 50 cases, respectively. These summaries mimic real-world admission notes. For each patient, clinical trials in these datasets are categorized into three groups depending on the eligibility of the patient to this trial: "Irrelevant" (no connection to the trial), "Excluded" / "Ineligible" (has the targeted condition but fails exclusion criteria), and "Eligible" (meets all enrollment requirements). Summary statistics of the TREC datasets are provided in Table 3. The distribution of primary disease categories across the TREC2021 and TREC2022 topics is shown in Figure 6.

These datasets allow for rigorous evaluation using standard ranking metrics such as Normalized Discounted Cumulative Gain (nDCG@k), Precision@k, and Recall@k (see Supplementary Methods section 5 for evaluation metrics definitions). Table 3 summarizes their characteristics, including age distributions, note lengths, and trial eligibility distributions.

~~Real-World Patient Data from the Netherlands Cancer Institute (NKI)~~

Building on strong benchmark results (see Results section for synthetic patients), we further validated TrialMatchAI's practical utility using real-world clinical data. For real-world validation, we used data from 52 metastatic cancer patients selected from the Whole Genome Sequencing Implementation in Standard Diagnostics for Every Cancer Patient (WIDE) study[37] conducted at the Netherlands Cancer Institute (NKI). The WIDE study collected essential patient information, including age, biological sex, informed consent status, biopsy feasibility, cancer type, both primary and metastatic tumor sites, and whether the patient underwent chemotherapy and radiotherapy. Notably, 94% of patients had undergone prior systemic chemotherapy and/or radiotherapy. Importantly, the study recorded the clinical trials to which each patient had been assigned, as well as unstructured descriptions of actionable molecular biomarkers, extracted from each patient's Molecular Tumor Board report. In all, the WIDE study collected the following fields relevant for clinical trial matching:

- A de-identified unique patient ID assigned to each participant,
- Year and month of birth,
- Biological sex (M/F),
- Written informed consent (Yes/No),
- Biopsy feasibility - whether a histological biopsy can be safely obtained during routine diagnostic procedures (Yes/No),
- Presence or suspicion of metastatic disease from solid tumors (Yes/No),
- Prior systemic chemotherapy (Yes/No)
- Prior radiation therapy (Yes/No),
- Tumor type classification (Text),
- Primary and metastatic tumor locations (Text),
- IDs of clinical trials to which the patient has been oriented,
- The molecular tumor board conclusion report contains actionable mutations and potential beneficial treatments (Text).

While the WIDE dataset provides key attributes for 947 patients, it lacks many detailed clinical parameters essential for precise trial eligibility assessment, including functional imaging, organ function metrics, and comprehensive comorbidity records. Thus, the global dataset presented limited overlap between the fields recorded in this study and the comprehensive information in

electronic health records, which were used for the original orientation of patients to their respective trials. Moreover, some patient-trial pairs lacked any matching criteria beyond the main condition and basic demographics, limiting their utility for evaluation.

To ensure meaningful evaluation, we selected a sub-cohort of 52 patients from the WIDE study whose information in the provided study fields had a sufficient intersection with the eligibility criteria of their assigned ground-truth trials, as determined by the Molecular Tumor Board based on full EHR data. Specifically, for a patient-trial pair to be included in the selected sub-cohort, at least 75% of the available patient information in the WIDE dataset had to correspond to criteria listed in the ground-truth trials. See Figure 4 in the Results section for baseline statistics of the sub-cohort of 52 patients.

Within this selected dataset, each patient was assigned at least one ground-truth trial, with some having up to three trials they were considered for. 217 clinical trials were included in this evaluation, and trial retrieval was conducted from this predefined pool. The actionable biomarkers identified through Molecular Tumor Board (MTB) reports were a key component of the data, making this dataset particularly relevant for biomarker-driven precision oncology trials.

Data Ingestion and Preprocessing

To facilitate patient-trial matching, TrialMatchAI processes both structured and unstructured data from diverse sources, converting this heterogeneous information into a standardized format optimized for efficient retrieval and accurate ranking, ensuring compatibility across varied clinical data types. At the end of this preprocessing pipeline, both patient records and clinical trial eligibility criteria are converted into dense embeddings, ensuring semantic compatibility for downstream retrieval and ranking. TrialMatchAI assumes patient data is provided in the Phenopackets exchange format[34], a standardized schema that facilitates semantic interoperability and consistent data representation. Users are expected to convert data from their native formats (e.g., FHIR[39,40] or OMOP[63]) into Phenopackets to enable compatibility with the system. However, direct harmonization between TrialMatchAI and EHR systems is an envisioned future development. See Supplementary Methods section 7 for additional details on Phenopackets and a sample Phenopackets template (in JSON format).

~~Clinical Trial Data Preprocessing~~

Before invoking the LLM-based pipeline, we apply deterministic registry filters to attributes that do not require LLM inference (e.g., age range, sex, and country/region/site availability). Clinical trial data undergoes preprocessing before downstream language processing. Data from ClinicalTrials.gov and CCMO is originally provided in XML format, from which key fields are extracted, including:

- NCT ID, brief title, official title
- Summary and detailed descriptions
- Study start/end dates, trial locations
- Minimum/maximum age, sex
- Eligibility criteria (most critical for patient-trial matching)

Each field undergoes standard preprocessing, including lowercasing, whitespace normalization, special character removal, and punctuation cleaning. Eligibility criteria require additional processing due to their importance in clinical trial matching. These undergo text splitting to extract individual inclusion and exclusion criteria, following a methodology adapted from the European Clinical Research Infrastructure Network (ECRIN) Clinical Research Metadata Repository[64] (detailed in Supplementary Methods section 1).

~~Entity Recognition, Normalization, and Data Augmentation~~

TrialMatchAI applies a two-step process to extract and standardize biomedical entities from both patient data and clinical trial records:

1. Named Entity Recognition (NER) extracts molecular biomarkers (genes, proteins, mutations), diseases, drugs, procedures, diagnostic tests, and clinical signs. It uses fine-tuned BERT-based models: BioBERT[24], RoBERTa-large-PM-M3-Voc[48], and GliNER[49] (zero-shot NER).
2. Entity Normalization maps extracted entities to biomedical vocabularies: MeSH, OMIM, ChEBI, Cell Ontology, NCBI Gene, and UMLS[62,65–69]. It uses BioSyn[47], a BERT-based entity normalization framework optimized for synonym handling. Moreover, it ensures that correct synonyms rank among top candidates when linking, improving entity disambiguation.

Additionally, for patient data, the original Phi-4 model[45] is used for zero-shot data augmentation (i.e, query expansion) by generating rephrasings, synonyms, and keyword expansions. Notably, a fine-tuned version of this model is primarily employed for eligibility classification (see section 4.4). This augmentation enhances query versatility, enabling a broader yet precise search space for relevant clinical trials. Details on the prompt used for data augmentation are provided in the Supplementary Methods section 3.3.

~~Embedding and Text Vectorization~~

To enable semantic search, TrialMatchAI converts textual data into high-dimensional dense embeddings, facilitating similarity-based retrieval. For clinical trials, eligibility criteria are embedded using BGE-M3[51], a biomedical transformer model optimized for medical text understanding. For patient records, structured and unstructured patient descriptions in Phenopackets format are embedded using the same model to ensure alignment with trial eligibility criteria. BGE-M3 transforms clinical data into dense vector representations, enabling efficient similarity-based retrieval. The BGE-M3 model has shown remarkable performance on multiple benchmarks and supports long-form text inputs up to 8,102 tokens.

Candidate Trial Retrieval

To efficiently retrieve relevant clinical trials for a given patient, TrialMatchAI employs a hybrid search strategy that integrates both lexical and semantic retrieval. This two-pronged approach ensures that trials are matched based on both exact textual overlap and deep semantic similarity, improving the system's ability to identify the most relevant options even when eligibility criteria are expressed in different ways. The result of this stage is an initial list of k candidate clinical trials ($k = 500$ was used in our evaluations) that are passed on to the reranking stage.

~~Elasticsearch Indexing~~

Clinical trial metadata and eligibility criteria are indexed using Elasticsearch, enabling rapid retrieval and ranking of candidate trials. Each trial is indexed at two levels to optimize search performance:

1. Trial-Level Index: Includes structured fields such as NCT ID, title, summary, study design, locations, and overall trial status.
2. Criterion-Level Index: Each inclusion and exclusion criterion is stored as an independent document, linked to the corresponding trial. This fine-grained indexing strategy allows for precise ranking based on individual eligibility conditions rather than relying solely on trial-wide metadata.

To further enhance retrieval effectiveness, indexed documents are enriched with named entities and their synonyms, extracted and normalized during preprocessing. This structured representation ensures better alignment between patient profiles and trial eligibility requirements.

~~Pre-Retrieval Filtering~~

Before executing the hybrid search, an initial filtering step is applied to reduce computational overhead and eliminate obviously ineligible trials. This filtering considers:

- Demographic factors, such as age, sex, and geographic location.
- Trial recruitment status, excluding closed or withdrawn trials.
- Basic eligibility requirements, such as broad inclusion criteria that clearly rule out certain patient populations (e.g., pediatric vs. adult-only trials).

By removing non-matching trials early in the pipeline, this filtering process improves both retrieval efficiency and ranking precision, ensuring that only clinically relevant trials are considered in the subsequent ranking and eligibility assessment steps.

~~Hybrid Retrieval Strategy~~

TrialMatchAI retrieves candidate trials using a dual-ranking strategy that combines BM25 lexical search and KNN-based semantic similarity search:

- BM25 (Lexical Search)[41,42]: A probabilistic ranking function that prioritizes documents containing query terms based on their frequency and importance. Optimized with parameters $k_1 = 1.2, b = 0.75$,BM25 provides strong keyword-based matching, making it effective for retrieving trials that explicitly mention key patient attributes.

- K-Nearest Neighbor (KNN) Vector Search: A semantic retrieval mechanism based on dense embeddings generated by BGE-M3, a transformer model optimized for biomedical text. TrialMatchAI implements Hierarchical Navigable Small World (HNSW) indexing[43], which allows for efficient approximate nearest-neighbor search in high-dimensional vector space.

These two retrieval methods operate in parallel via the Elasticsearch querying mechanism, ensuring that both textually relevant and conceptually similar trials are surfaced. By integrating lexical and semantic search, TrialMatchAI is able to retrieve trials that a purely keyword-based or embedding-based system might overlook.

LLM-based Re-Ranking of Trials and Eligibility Assessment

After candidate retrieval, TrialMatchAI utilizes a two-stage process to refine and precisely rank clinical trial recommendations:

1. LLM-based Re-ranking: trials are prioritized based on the relevance of their eligibility criteria to the patient's clinical profile.
2. Criterion-level Eligibility Assessment: a detailed, criterion-by-criterion analysis assesses the exact alignment between patient characteristics and individual trial eligibility requirements.

This structured approach ensures that recommended clinical trials are not only relevant but also precisely tailored to each patient's unique clinical profile. Below is a detailed explanation of the two-stage process.

~~LLM-Based Re-Ranking of Retrieved Trials~~

To refine candidate trials, a Gemma-2-2b model, fine-tuned on an augmented medical natural language inference dataset (see section 4.6), assesses the relevance of individual eligibility criteria to the patient profile. This criterion-level re-ranking, unlike initial trial retrieval, prioritizes

trials with the most pertinent criteria for subsequent eligibility assessment. The process starts with a second-level hybrid search (BM25 + vector search) to retrieve relevant eligibility criteria for each patient query. These query-criterion pairs are then input into the Gemma-2-2b model, which determines if the patient query (Statement A) sufficiently addresses the clinical trial criterion (Statement B). The model outputs a binary relevance decision, converted into a 0–1 confidence score, where higher values indicate stronger relevance.

Trial-level relevance is computed by aggregating criterion scores using methods such as maximum score, mean, square-root normalization, logarithmic normalization, or a weighted combination (e.g., 70% sqrt-normalized sum, 30% max score), balancing trials with varying numbers of relevant criteria. These aggregated scores are combined with the initial trial-level retrieval score to produce a final ranking for eligibility assessment.

~~Criterion-level Eligibility Assessment~~

The core module of TrialMatchAI is criterion-level eligibility assessment, which serves as the final step in a Retrieval-Augmented Generation (RAG) pipeline[33]. The Phi-4 model, fine-tuned on a medical chain-of-thought (CoT) dataset[33,70], classifies each inclusion and exclusion criterion of a given clinical trial based on the retrieved relevant information and provided patient data within its context window. For inclusion criteria, the possible classifications are “Met”, “Not Met”, “Unclear”, and “Irrelevant”, while for exclusion criteria, the model assigns one of “Violated”, “Not Violated”, “Unclear”, or “Irrelevant”. Crucially, each classification must be accompanied by a detailed justification, relying exclusively on the explicitly provided patient information and the retrieved context from the RAG pipeline. The model’s final output is structured as a JSON object, where each criterion is represented as a nested field containing its original text, assigned classification, and the model’s justification, with references to the retrieved knowledge that supported its decision. The complete prompt used for instructing the Phi-4 model is provided in Supplementary Methods section 3.3.

Final Ranking Procedure

The final ranking stage in TrialMatchAI assigns a composite score to each clinical trial based on the alignment between the trial’s eligibility criteria and the patient’s information. The scoring mechanism prioritizes trials with more satisfied inclusion criteria and fewer violated exclusion criteria while treating ambiguous cases neutrally. Criteria classified as “Unclear” or “Irrelevant” are excluded from the score computation to ensure strict adherence to the available patient data. The inclusion and exclusion scores are computed as follows:

$$S_{inc} = \frac{\sum_{i=1}^{N} w_i}{N}, S_{exc} = \frac{\sum_{j=1}^{M} w_j}{M} \qquad (1)$$

where N and M are the total valid inclusion and exclusion criteria, respectively, and w_i w_j represent weighted contributions based on classification (positive for “Met” / “Not Violated”, negative for “Not Met” / “Violated”).

The final composite score for ranking is given by:

$$S = \frac{S_{inc}+S_{exc}}{2} \qquad (2)$$

where higher scores indicate better trial relevance. The trials are then sorted in descending order based on S.

Fine-tuning of LLMs for Different Tasks

TrialMatchAI's core modules utilize multiple fine-tuned LLMs primarily for biomedical entity recognition, relevance-based re-ranking, and eligibility criteria evaluation.

~~LLM Fine-Tuning for Biomedical Entity Recognition~~

For named entity recognition (NER), BioBERT and RoBERTa-large-PM-M3-Voc were fine-tuned to identify a wide range of biomedical entities using datasets such as BC2GM[71], BC4CHEMD[72], MACCROBAT 2018 and 2020[73], JNLPBA[74], BC5CDR[75], tmVar[76], and NCBI-disease[77]. The fine-tuning process was designed to support a multi-task NER framework, similar to Biomedical Entity Recognition and Normalization 2 (BERN2)[46]. Multi-task NER models enable the simultaneous recognition of multiple entity types in a single inference pass, significantly improving efficiency by reducing both computational overhead and memory consumption compared to single-task NER models, which require separate models for each entity type. Comprehensive evaluation results from fine-tuning on multiple benchmark datasets are provided in the Supplementary Methods section 4.

~~LLM Fine-Tuning for Trial Re-Ranking~~

A 4-bit quantized Gemma-2-2B model[44] was fine-tuned for trial re-ranking using QLoRA (Quantized Low-Rank Adaptation)[78]. Instruction tuning, aimed at enhancing the model's ability to interpret task-specific instructions, was performed using an augmented MedNLI dataset[79]. This dataset combined original MedNLI examples with approximately 30,000 synthetic instances generated by GPT-4o-mini. Synthetic data was created by paraphrasing and expanding randomly sampled clinical trial eligibility criteria, then generating related and unrelated patient descriptions to ensure training diversity.

QLoRA fine-tuning parameters were set as follows: rank = 16, $\alpha = 32$, warmup ratio = 0.1, learning rate = 2×10^{-4} (with scheduler), and 1 training epoch, adhering to established best practices[80]. In QLoRA, rank (r) determines adaptation capacity, and α scales LoRA updates. Given that relevance detection is a binary classification task, a moderate rank $r = 16$ was chosen to balance performance and computational efficiency, as higher ranks provide diminishing returns for this task's complexity. The fine-tuned Gemma-2-2b model achieves the following evaluation results on the MedNLI test set: precision = 0.8804, recall = 0.8594, F1-score = 0.8672.

~~Chain-of-Thought Fine-Tuning for Eligibility Criteria Evaluation~~

QLoRA was also used to fine-tune the Phi-4 model[45] with a medical chain-of-thought (CoT) training dataset for criterion-level eligibility assessment, incorporating instruction-input-output triplets. This dataset is designed to enhance an LLM's ability to engage in multi-step reasoning, allowing it to consider multiple reasoning paths before reaching a final decision[70].

Originally developed for complex Medical Question Answering (MedQA), the dataset contains over 25,000 English-language examples aimed at improving LLMs' medical reasoning capabilities. However, in this task, the focus is on criterion-level patient eligibility assessment,

where the model evaluates each eligibility criterion individually rather than answering general medical questions. The objective was to equip Phi-4 with chain-of-thought reasoning rather than simply expanding its medical knowledge base. The QLoRA fine-tuning hyper-parameters for this task are: rank = 32, $\alpha = 64$, warm-up ratio = 0.1, learning rate (with a scheduler) = 2×10^{-4}, and training epochs = 1. The higher rank (32) and alpha (64) were chosen for this task to accommodate the greater complexity of chain-of-thought reasoning, as multi-step logical inference requires a more expressive adaptation capacity than binary relevance classification. The fine-tuned Phi-4 model achieves the following evaluation results on the test set: precision = 0.8678, recall = 0.8805, F1-score = 0.8741.

Statistics and Reproducibility

No statistical method was used to predetermine sample size. Sample sizes were determined by the availability of benchmark datasets (TREC2021 *n* = 75 topics; TREC2022 *n* = 50 topics), synthetic evaluation cohorts ("Ideal Candidates", *n* = 100), and the restricted-access real-world cohort derived from WIDE (*n* = 52).

No data were excluded from the analyses except where explicitly required by the study design (e.g., selection of the WIDE sub-cohort based on evaluability of trial orientation with the available fields). Ranking and retrieval metrics were computed per patient/topic and summarized using medians (boxplots) and/or means ± s.d., as indicated. For the ablation study, statistical differences versus the full configuration were assessed using paired comparisons across patients, with false discovery rate (FDR) adjustment for multiple testing; significance was defined at $p < 0.05$ (Figure 5).

The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reproducibility is supported through public release of the source code, configurations, prompts, and evaluation outputs (Data and Code Availability).

Data Availability

~~The synthetic datasets for the TREC 2021 and 2022 clinical trial tracks can be accessed here: https://www.trec-cds.org/2021.html and http://www.trec-cds.org/2022.html. All data including the fine-tuned models (low rank adapter format), "Ideal Candidates" synthetic patients dataset and their matching results, the expert-examined patient-criteria pairs, the curated and parsed clinical trial database, the results on TREC benchmarks, and the dictionaries used for entity normalization are publicly accessible via the Zenodo repository dedicated to TrialMatchAI at https://zenodo.org/records/15045515.~~

The data generated in this study have been deposited in the Zenodo repository dedicated to TrialMatchAI under accession code 15045515: https://zenodo.org/records/15045515. These include the fine-tuned models (low rank adapter format), the "Ideal Candidates" synthetic patient dataset and matching results, the expert-examined patient–criterion pairs, the curated and parsed clinical trial database, the results on the TREC benchmarks, and the dictionaries used for entity normalization. Source data are provided as a Source File with this paper.

The synthetic datasets used in this study for the TREC 2021 and 2022 Clinical Trials tracks are available at https://www.trec-cds.org/2021.html and http://www.trec-cds.org/2022.html. The lists

of publicly accessible clinical trial identifiers used in the evaluation (ClinicalTrials.gov and CCMO) are provided in the Source Data file.

The WIDE study data from the Netherlands Cancer Institute (NKI) are available under restricted access because they contain patient-level clinical information subject to privacy and institutional governance requirements. Access can be obtained by submitting a request through the Office Manager Knowledge Transfer & Contracting (KT&C) at NKI and completing the required data transfer/use agreements; requests are reviewed on a case-by-case basis. The expected response timeframe is approximately 3–6 months, and access is typically granted for one year (extensions may be requested).

Code Availability

The TrialMatchAI source code is available under the MIT Licence on GitHub at https://github.com/cbib/TrialMatchAI. A citable archived release is available on Zenodo[81].

Acknowledgments

This work was supported by the European Union under the EOSC4Cancer project, funded by the European Research Executive Agency (REA) under the European Union's Horizon Europe program (grant agreement ID:101058427; M.A., M.G., M.A.R., S.C., S.N., M.N., R.D., R.F., G.M., M.B., L.M., E.H., J.G., A.G., S.K.). The grant supported research costs and personnel, including M.A.
This work also benefited from access to the computing resources of the "CALI 3" cluster. This cluster is operated and hosted by the University of Limoges. It is part of the HPC network in the Nouvelle-Aquitaine Region in France, funded by the French government and the Region.

Author contributions

M.A. developed the TrialMatchAI system and performed the main experiments. M.A. and M.G. conducted the benchmarking analyses. M.A.R. and S.C. designed the inclusion/exclusion criteria splitting methodology. M.A., S.N. and M.N. performed data analysis and interpretation. M.A., M.N., R.D., R.F. and G.M. contributed to the analysis and interpretation of the real patient's cohort data, with M.B. and L.M. supporting curation and clinical context. J.G. assessed the clinical plausibility of the generated eligibility rationales. A.G. and S.K. contributed to system design choices and implementation discussions. E.H. contributed to scientific discussion and project oversight. M.N. supervised and directed the project. M.A. and M.N. wrote the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Tables

Table 1. Summary of TrialMatchAI Components. Asterisk (*) denotes a fine-tuned version of the model.

Component	Role(s)	Models/Tools Used
Entity Recognition	Extract entities from unstructured text	BioBERT*, RoBERTa-large*, GLiNER
Entity Normalization	Normalize entities to knowledge bases, enrich data with synonyms	Rules-based methods, BioSyn
Text Embedding	Convert text into numerical representations	BGE-M3
Candidate Retrieval	Perform first-stage text and semantic-based searches	Elasticsearch BM25 / KNN
Neural Re-Ranking	Refines candidate trials relevance at the criterion level	Gemma-2-2B*
Recommendation Engine	Criterion-level eligibility assessment, keyword and text generation for trial recommendations	Phi-4*

Table 2. Performance metrics for top-k retrieval. Overall Recall represents the fraction of patients with at least one ground-truth trial retrieved within the top-k trials (k = 5, 10, 20). The Mean Reciprocal Rank (MRR) evaluates how well the system ranks the first ground-truth trial across all patients. MRR ranges from 0 to 1, with 1 indicating that a ground truth trial is always ranked at the top (rank 1) for every patient, and 0 indicating that no ground-truth trials were retrieved within the top-k for any patient. The Mean Average Rank is the average rank of all retrieved ground-truth trials across all patients.

Metric	Top 5	Top 10	Top 20
Overall Recall	0.8462	0.8846	0.9231
Mean Reciprocal Rank (MRR) ± Std	0.4824 ± 0.2952	0.4880 ± 0.2873	0.4904 ± 0.2835
Mean Average Rank ± Std	2.0341 ± 0.7568	2.2500 ± 1.2716	2.8438 ± 3.1493

Table 3. Summary of cohort characteristics for TREC2021 CT and TREC2022 CT. The reported baseline statistics comprise the following: total number of patients (N); average age along with standard deviation (in years); sex distribution (ratio of male to female); average length of patient notes (measured in words); mean and standard deviation of eligible trials per patient; mean and standard deviation of excluded trials per patient; average number of irrelevant trials per patient; and the total number of initial trials within the search space, as classified by TREC judges.

Metric	TREC 2021 CT	TREC 2022 CT
N (Number of patients)	75	50
Age (years, mean ± std)	41.6 ± 19.4	35.3 ± 20.2
Sex (Male/Female)	38/37	28/22
Topic length (words)	156.2 ± 45.4	109.9 ± 21.6
Eligible trials per patient	74.3 ± 49.0	78.8 ± 67.3
Excluded trials per patient	80.3 ± 60.3	60.7 ± 65.5
Irrelevant trials per patient	323.2 ± 93.2	568.4 ± 164.1
Initial trials	26,149	26,581

Figure legends

Figure 1. TrialMatchAI Workflow for Automated Patient-to-Trial Matching. (1) Clinical trial metadata, including eligibility criteria, are extracted from structured sources (XML formats) and processed through a text parsing module that includes named entity recognition, entity normalization, and synonym enrichment. The parsed and original texts are embedded into numerical representations using an embedding model (e.g., BGE-M3) and indexed using Elasticsearch for efficient retrieval. (2) Patient data, encompassing clinical and molecular information in Phenopackets exchange format (JSON), undergo a similar text parsing process, including named entity recognition, entity normalization, query expansion, and synonym enrichment, followed by embedding transformation to generate query vectors. (3) An initial list of candidate trials is retrieved using a hybrid approach that combines BM25 text-based retrieval with k-nearest neighbors (KNN) search on embedded representations. The list is further refined via criterion-level relevance assessment using a large language model (LLM) fine-tuned for biomedical text re-ranking (Gemma-2-2B). The relevant trials are subsequently assessed for criterion-by-criterion eligibility using a final LLM (Phi-4) fine-tuned for biomedical reasoning. The model generates concise and interpretable explanations for each criterion-level classification, ensuring the explainability of AI-driven decision-making. (4) Finally, criterion-level matching scores are aggregated into trial-level overall eligibility scores, producing a personalized ranked list of clinical trial recommendations for the given patient. Icons are from Bootstrap Icons (MIT License) and Font Awesome Free (CC BY 4.0), with minor adaptations in some cases.

Figure 2. Distribution of ground truth trial rankings assigned to 100 synthetic patient profiles by TrialMatchAI. (A) Synthetic patient generation workflow: GPT-4o-mini is used to

generate patients that closely match real clinical trials, using trial metadata from XML files, followed by manual curation. (B) Percentage of patients by rank position of the ground-truth trial. The x-axis shows the rank of the ground-truth clinical trial, and the y-axis shows the percentage of patients whose ground-truth trial was assigned that rank. 95% of patients had their ground-truth trial ranked in the top two, indicating high precision in patient–trial matching. Icons are from Bootstrap Icons (MIT License) and Font Awesome Free (CC BY 4.0), with minor adaptations in some cases. Source data are provided as a Source Data file.

Figure 3. Performance evaluation of TrialMatchAI on the TREC clinical trial datasets. (A) Recall vs. retrieval size for patient-trial matching on TREC2021 and TREC2022 datasets, showing increasing recall with larger retrieval sizes. (B) Expert evaluation results on 950 patient-criterion pairs, measuring the percentage of correct model decisions for inclusion and exclusion criteria. (C) Evaluation metrics (nDCG and precision at 5, 10, and 20) across TREC2021 patients, illustrating ranking performance. (D) Evaluation metrics across TREC2022 patients, demonstrating consistent model performance. These results validate the effectiveness of the system in retrieving relevant trials and correctly assessing eligibility criteria. In each box plot, the solid horizontal line denotes the median, while the dotted horizontal line denotes the mean. The lower and upper bounds of the box correspond to the 25th and 75th percentiles (interquartile range, IQR); whiskers extend to the minimum and maximum values within 1.5×IQR from the box; points beyond the whiskers represent individual observations. n = number of independent benchmark topics (synthetic patient cases): TREC2021 n=75; TREC2022 n=50. Source data are provided as a Source Data file.

Figure 4. Summary Statistics of the selected subset of 52 cancer patients enrolled in the WIDE study at the NKI. (A) The Primary Cancer Type Distribution illustrates the distribution of primary cancer types in the cohort, with colorectal, lung, and breast cancer being the most prevalent. (B) The Treatment and Mutation Presence displays the proportions of patients who have undergone systemic chemotherapy and radiotherapy, along with the percentage of patients whose Molecular Tumor Board (MTB) report discusses actionable mutations. (C) The Actionable Biomarkers Prevalence presents the distribution of actionable genes identified in the MTB reports, with percentages indicating the proportion of patients harboring each mutated gene. Note: 100% of patients have metastatic cancer, provided written informed consent, and have been allocated to clinical trials.

Figure 5. Ablation analysis of TrialMatchAI pipeline components across ranking and precision metrics. Boxplots show per-patient scores (n = 75) for each ablation configuration; the Full pipeline (leftmost) includes NER, medical chain-of-thought (CoT), multi-stage retrieval (BM25 + vector fusion), and LLM reranking. The top row reports nDCG@k and the bottom row reports Precision@k at k = 5, 10, and 20: (A) nDCG@5, (B) nDCG@10, (C) nDCG@20, (D) P@5, (E) P@10, (F) P@20. Configurations (x-axis) include Full; BM25+vector with second-stage retrieval, CoT, and NER removed; BM25 Full; single-component removals (No 2nd Level, No CoT, No NER, No Rerank); and VECTOR Full. Horizontal brackets indicate pairwise comparisons versus Full (two-sided Mann–Whitney U tests) with FDR correction across Full-vs-ablation tests within each panel (*P < 0.05, **P < 0.01, ***P < 0.001). Boxes show the interquartile range (25th–75th percentiles); the median is the solid line and the mean the dashed line. Whiskers extend to 1.5×IQR, with outliers plotted as points; overlaid points represent individual patients. The long dashed horizontal line marks the Full-pipeline median in each panel. Source data are provided in the Source Data file.

Figure 6. Distribution of primary disease categories across TREC-2021 and TREC-2022. Proportional representation (%) of primary diagnostic domains among patient topics in the TREC

2021 (n = 75) and TREC 2022 (n = 50) Clinical Trials tracks. The taxonomy spans fifteen clinically interpretable categories; the Complex/Multimorbid group includes patients with multiple coexisting main chronic conditions (e.g., concurrent cardiovascular, metabolic, and renal disease) lacking a single dominant pathology. Source data are provided as a Source Data file.

References

1. Garraway, L. A., Verweij, J. & Ballman, K. V. Precision Oncology: An Overview. *J. Clin. Oncol.* **31**, 1803–1805 (2013).
2. Dienstmann, R., Jang, I. S., Bot, B., Friend, S. & Guinney, J. Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer Discov.* **5**, 118–123 (2015).
3. Briel, M. *et al.* A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable. *J. Clin. Epidemiol.* **80**, 8–15 (2016).
4. Fogel, D. B. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemp. Clin. Trials Commun.* **11**, 156–164 (2018).
5. Luchini, C., Lawlor, R. T., Milella, M. & Scarpa, A. Molecular Tumor Boards in Clinical Practice. *Trends Cancer* **6**, 738–744 (2020).
6. Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L. & Siu, L. L. Molecular profiling for precision cancer therapies. *Genome Med.* **12**, 8 (2020).
7. Penberthy, L. T., Dahman, B. A., Petkov, V. I. & DeShazo, J. P. Effort required in eligibility screening for clinical trials. *J. Oncol. Pract.* **8**, 365–370 (2012).
8. Meystre, S. M. *et al.* Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models. *BMC Med. Res. Methodol.* **23**, 88 (2023).
9. Ni, Y. *et al.* Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. *BMC Med. Inform. Decis. Mak.* **15**, 28 (2015).

10. Hassanzadeh, H., Karimi, S. & Nguyen, A. Matching patients to clinical trials using semantically enriched document representation. *J. Biomed. Inform.* **105**, 103406 (2020).

11. Weng, C. *et al.* An Integrated Model for Patient Care and Clinical Trials (IMPACT) to support clinical research visit scheduling workflow for future learning health systems. *J. Biomed. Inform.* **46**, 642–652 (2013).

12. Jin, Q. *et al.* Matching patients to clinical trials with large language models. *Nat. Commun.* **15**, 9074 (2024).

13. Gao, J., Xiao, C., Glass, L. M. & Sun, J. COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching. Preprint at https://doi.org/10.48550/arXiv.2006.08765 (2020).

14. Thadani, S. R., Weng, C., Bigger, J. T., Ennever, J. F. & Wajngurt, D. Electronic Screening Improves Efficiency in Clinical Trial Recruitment. *J. Am. Med. Inform. Assoc. JAMIA* **16**, 869–873 (2009).

15. Zhang, X., Xiao, C., Glass, L. M. & Sun, J. DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction. Preprint at https://doi.org/10.48550/arXiv.2001.08179 (2020).

16. Weng, C. *et al.* EliXR: an approach to eligibility criteria extraction and representation. *J. Am. Med. Inform. Assoc. JAMIA* **18**, i116–i124 (2011).

17. Yuan, C. *et al.* Criteria2Query: a natural language interface to clinical databases for cohort definition. *J. Am. Med. Inform. Assoc. JAMIA* **26**, 294–305 (2019).

18. Shi, J., Graves, K. & Hurdle, J. F. A generic rule-based system for clinical trial patient selection. Preprint at https://doi.org/10.48550/arXiv.1907.06860 (2019).

19. O'Regan, P. *et al.* Digital ECMT Cancer Trial Matching Tool: an Open Source Research Application to Support Oncologists in the Identification of Precision Medicine Clinical Trials. *JCO Clin. Cancer Inform.* **7**, e2200137 (2023).

20. Alzubaidi, L. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **8**, 53 (2021).

21. Klein, H. *et al.* MatchMiner: an open-source platform for cancer precision medicine. *Npj Precis. Oncol.* **6**, 69 (2022).

22. Rybinski, M., Kusa, W., Karimi, S. & Hanbury, A. Learning to match patients to clinical trials using large language models. *J. Biomed. Inform.* **159**, 104734 (2024).

23. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).

24. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

25. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.1810.04805 (2019).

26. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).

27. Meng, X. *et al.* The application of large language models in medicine: A scoping review. *iScience* **27**, 109713 (2024).

28. Kaskovich, S. *et al.* Automated Matching of Patients to Clinical Trials: A Patient-Centric Natural Language Processing Approach for Pediatric Leukemia. *JCO Clin. Cancer Inform.* **7**, e2300009 (2023).

29. Gueguen, L. *et al.* A prospective pragmatic evaluation of automatic trial matching tools in a molecular tumor board. *Npj Precis. Oncol.* **9**, 28 (2025).

30. de Kok, J. W. T. M. *et al.* A guide to sharing open healthcare data under the General Data Protection Regulation. *Sci. Data* **10**, 404 (2023).

31. Li, X. *et al.* LLM-Match: An Open-Sourced Patient Matching Model Based on Large Language Models and Retrieval-Augmented Generation. Preprint at https://doi.org/10.48550/arXiv.2503.13281 (2025).

32. Lin, J., Xu, H., Wang, Z., Wang, S. & Sun, J. Panacea: A foundation model for clinical trial search, summarization, design, and recruitment. Preprint at https://doi.org/10.48550/arXiv.2407.11007 (2024).

33. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Preprint at https://doi.org/10.48550/arXiv.2005.11401 (2021).

34. Jacobsen, J. O. B. *et al.* The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol.* **40**, 817–820 (2022).

35. Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S. & Hersh, W. R. Overview of the TREC 2021 Clinical Trials Track. in (2021).

36. Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S. & Hersh, W. R. Overview of the TREC 2022 Clinical Trials Track. in (2022).

37. Samsom, K. G. *et al.* Study protocol: Whole genome sequencing Implementation in standard Diagnostics for Every cancer patient (WIDE). *BMC Med. Genomics* **13**, 169 (2020).

38. Wei, J. *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2201.11903 (2023).

39. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R. & Stiawan, D. The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR Med. Inform.* **9**, e21929 (2021).

40. Jacobsen, J. O. B. *et al.* The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol.* 40, 817–820 (2022).

41. Robertson, S. *et al.* Okapi at TREC-3. *NIST Spec. Publ.* 109 (1995).

42. Robertson, S. & Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found Trends Inf Retr* **3**, 333–389 (2009).

43. Malkov, Y. A. & Yashunin, D. A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans Pattern Anal Mach Intell* **42**, 824–836 (2020).

44. Team, G. *et al.* Gemma 2: Improving Open Language Models at a Practical Size. Preprint at https://doi.org/10.48550/arXiv.2408.00118 (2024).

45. Abdin, M. *et al.* Phi-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2412.08905 (2024).

46. Sung, M. *et al.* BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* **38**, 4837–4839 (2022).

47. Sung, M., Jeon, H., Lee, J. & Kang, J. Biomedical Entity Representations with Synonym Marginalization. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J.) 3641–3650 (Association for Computational Linguistics, Online, 2020). doi:10.18653/v1/2020.acl-main.335.

48. Lewis, P., Ott, M., Du, J. & Stoyanov, V. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. in *Proceedings of the 3rd Clinical Natural Language Processing Workshop* (eds Rumshisky, A., Roberts, K., Bethard, S. & Naumann, T.) 146–157 (Association for Computational Linguistics, Online, 2020). doi:10.18653/v1/2020.clinicalnlp-1.17.

49. Zaratiana, U., Tomeh, N., Holat, P. & Charnois, T. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. Preprint at https://doi.org/10.48550/arXiv.2311.08526 (2023).

50. Peng, H. *et al.* Biomedical named entity normalization via interaction-based synonym marginalization. *J. Biomed. Inform.* **136**, 104238 (2022).

51. Chen, J. *et al.* M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. Preprint at https://doi.org/10.48550/arXiv.2402.03216 (2025).

52. Gormley, C. & Tong, Z. *Elasticsearch: The Definitive Guide*. (O'Reilly Media, Inc., 2015).

53. Durden, K. *et al.* Provider motivations and barriers to cancer clinical trial screening, referral, and operations: Findings from a survey. *Cancer* **130**, 68–76 (2024).

54. Davis, S. ICH GCP E6 has arrived: Are we ready? *Perspect. Clin. Res.* **16**, 171–172 (2025).

55. Dexheimer, J. W. *et al.* A Time-and-Motion Study of Clinical Trial Eligibility Screening in a Pediatric Emergency Department. *Pediatr. Emerg. Care* **35**, 868–873 (2019).

56. Yu, C. *et al.* A Survey on Agent Workflow -- Status and Future. in *2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD)* 770–781 (2025). doi:10.1109/ICAIBD64986.2025.11082076.

57. Xu, X. *et al.* A Survey on Knowledge Distillation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2402.13116 (2024).

58. Nievas, M., Basu, A., Wang, Y. & Singh, H. Distilling large language models for matching patients to clinical trials. *J. Am. Med. Inform. Assoc. JAMIA* **31**, 1953–1963 (2024).

59. Schafer, J. B., Frankowski, D., Herlocker, J. & Sen, S. Collaborative Filtering Recommender Systems. in *The Adaptive Web: Methods and Strategies of Web Personalization* (eds Brusilovsky, P., Kobsa, A. & Nejdl, W.) 291–324 (Springer, Berlin, Heidelberg, 2007). doi:10.1007/978-3-540-72079-9_9.

60. Kundra, R. *et al.* OncoTree: A Cancer Classification System for Precision Oncology. *JCO Clin. Cancer Inform.* **5**, 221–230 (2021).

61. Huiskens, J. *et al.* From registration to publication: A study on Dutch academic randomized controlled trials. *Res. Synth. Methods* **11**, 218–226 (2020).

62. Gallifant J. *et al.* Peer review of GPT-4 technical report and systems card. *PLOS Digital Health* **3**, e0000417 (2024).

63. Biedermann, P. *et al.* Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Med. Res. Methodol.* **21**, 238 (2021).

64. Canham, S. & Ohmann, C. A metadata schema for data objects in clinical research. *Trials* **17**, 557 (2016).

65. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* **15**, 57–61 (2000).

66. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344–D350 (2008).

67. Meehan, T. F. *et al.* Logical Development of the Cell Ontology. *BMC Bioinformatics* **12**, 6 (2011).

68. Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, D36-42 (2015).

69. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-270 (2004).

70. Chen, J. *et al.* HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. Preprint at https://doi.org/10.48550/arXiv.2412.18925 (2024).

71. Smith, L. *et al.* Overview of BioCreative II gene mention recognition. *Genome Biol.* **9 Suppl 2**, S2 (2008).

72. Krallinger, M. *et al.* The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminformatics* **7**, S2 (2015).

73. Caufield, J. H. *et al.* A reference set of curated biomedical data and metadata from clinical case reports. *Sci. Data* **5**, 180258 (2018).

74. Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y. & Collier, N. Introduction to the bio-entity recognition task at JNLPBA. in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04* 70 (Association for Computational Linguistics, Geneva, Switzerland, 2004). doi:10.3115/1567594.1567610.

75. Li, J. *et al.* BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation* **2016**, baw068 (2016).

76. Wei, C.-H. *et al.* tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinforma. Oxf. Engl.* **34**, 80–87 (2018).

77. Doğan, R. I., Leaman, R. & Lu, Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **47**, 1–10 (2014).

78. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. Preprint at https://doi.org/10.48550/arXiv.2305.14314 (2023).

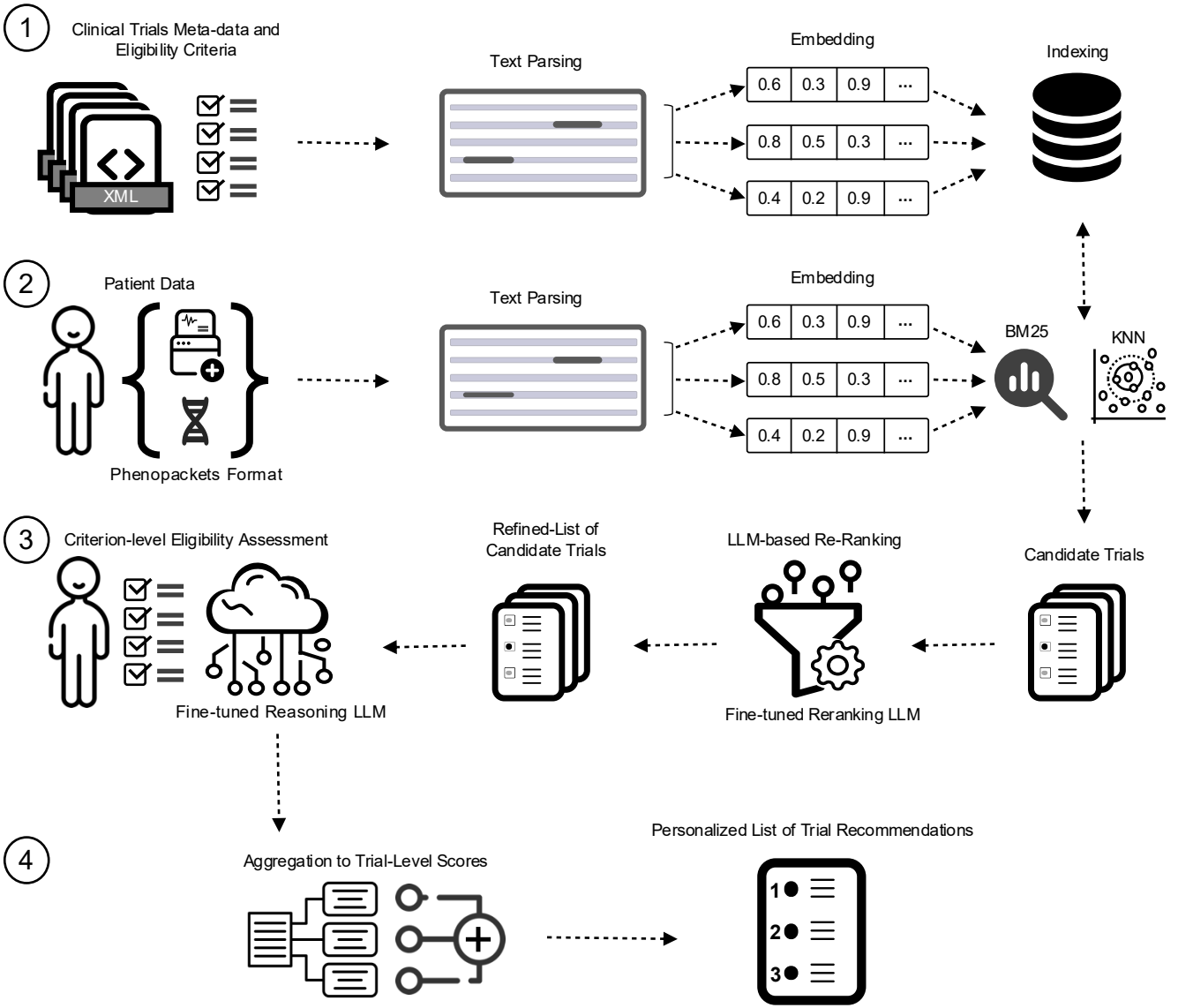
79. Romanov, A. & Shivade, C. Lessons from Natural Language Inference in the Clinical Domain. Preprint at https://doi.org/10.48550/arXiv.1808.06752 (2018).

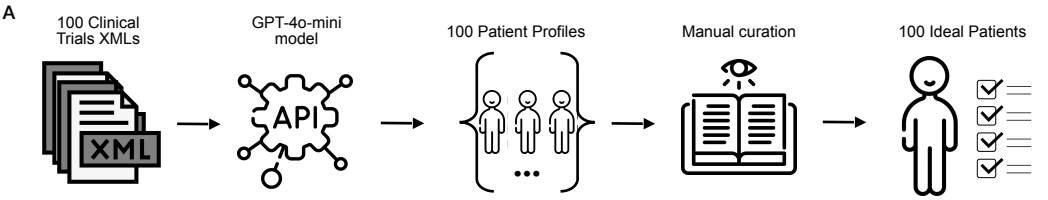
80. Parthasarathy, V. B., Zafar, A., Khan, A. & Shahid, A. The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. Preprint at https://doi.org/10.48550/arXiv.2408.13296 (2024).

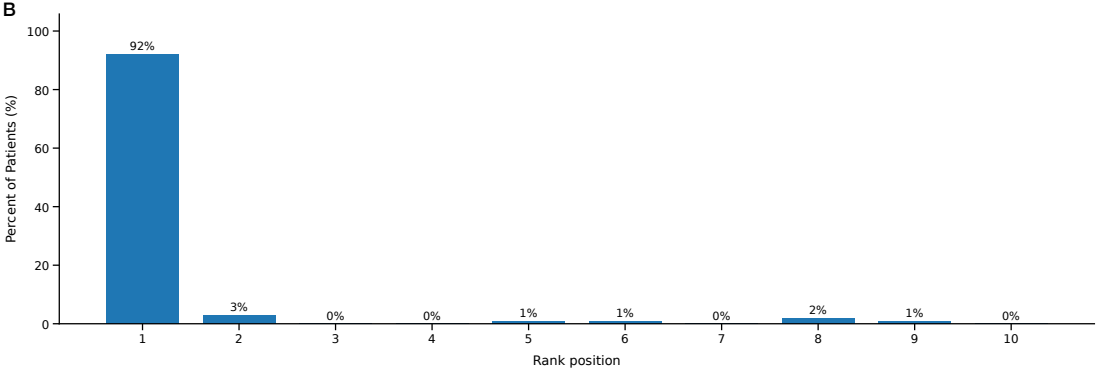
81. Abdallah, M., Georges, M., Nikolski, M. TrialMatchAI: An End-to-End AI-powered Clinical Trial Recommendation System to Streamline Patient-to-Trial Matching. cbib/TrialMatchAI. doi:10.5281/zenodo.18329084 (2026).

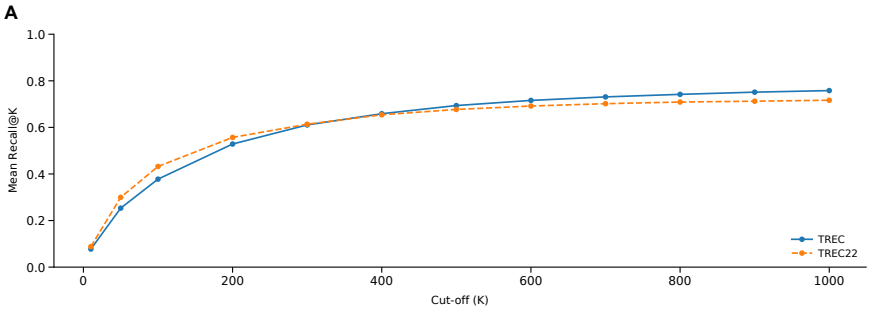
Editorial Summary
Patient recruitment remains a major bottleneck in clinical trials. Here, the authors present TrialMatchAI, an open-source AI-powered recommendation system that automates patient-to-trial matching from heterogeneous clinical data with state-of-the-art performance, high efficiency and interpretability, and secure local deployment.

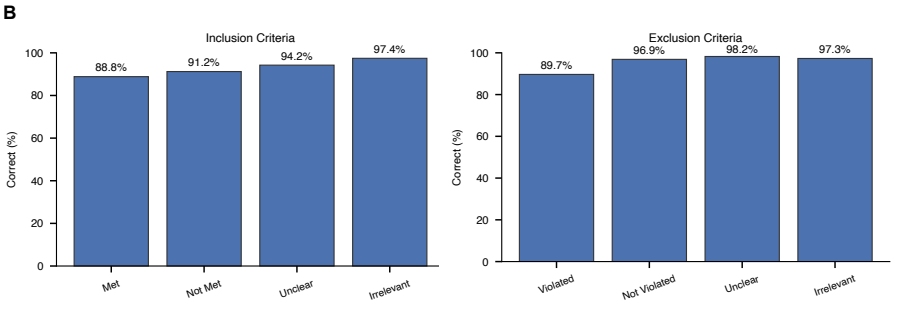
Peer review information: *Nature Communications* thanks Kirk Roberts, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

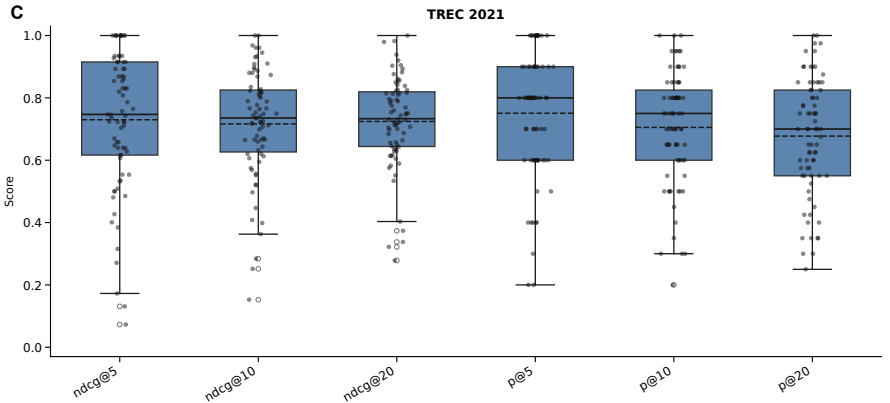

1
Clinical Trials Meta-data and Eligibility Criteria
XML
Text Parsing
Embedding
0.6 0.3 0.9 ...
0.8 0.5 0.3 ...
0.4 0.2 0.9 ...
Indexing
2
Patient Data
Phenopackets Format
Text Parsing
Embedding
0.6 0.3 0.9 ...
0.8 0.5 0.3 ...
0.4 0.2 0.9 ...
BM25
KNN
3
Criterion-level Eligibility Assessment
Fine-tuned Reasoning LLM
Refined-List of Candidate Trials
LLM-based Re-Ranking
Fine-tuned Reranking LLM
Candidate Trials
4
Aggregation to Trial-Level Scores
Personalized List of Trial Recommendations
1
2
3

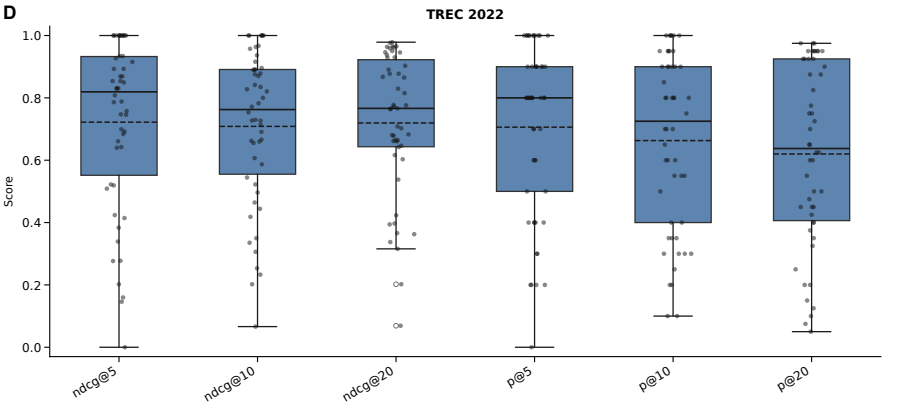


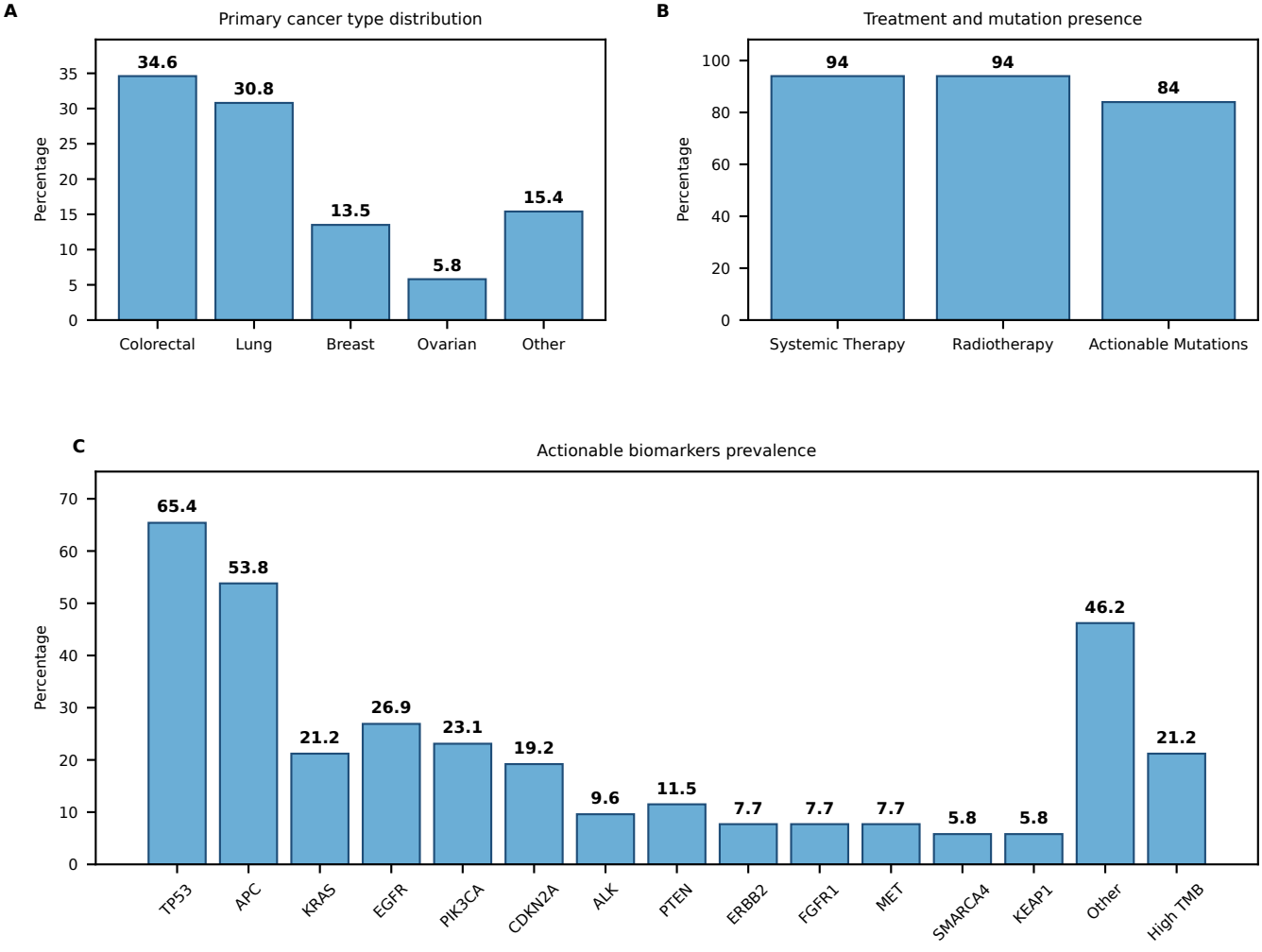



A
Mean Recall@K
Cut-off (K)
TREC
TREC22


B
Inclusion Criteria
Correct (%)
88.8%
91.2%
94.2%
97.4%
Met
Not Met
Unclear
Irrelevant
Exclusion Criteria
89.7%
96.9%
98.2%
97.3%
Violated
Not Violated
Unclear
Irrelevant


C
TREC 2021
Score
ndcg@5
ndcg@10
ndcg@20
p@5
p@10
p@20


D
TREC 2022
Score
ndcg@5
ndcg@10
ndcg@20
p@5
p@10
p@20


A
Primary cancer type distribution
Percentage
34.6
30.8
13.5
5.8
15.4
Colorectal
Lung
Breast
Ovarian
Other
B
Treatment and mutation presence
Percentage
94
94
84
Systemic Therapy
Radiotherapy
Actionable Mutations
C
Actionable biomarkers prevalence
Percentage
65.4
53.8
21.2
26.9
23.1
19.2
9.6
11.5
7.7
7.7
7.7
5.8
5.8
46.2
21.2
TP53
APC
KRAS
EGFR
PIK3CA
CDKN2A
ALK
PTEN
ERBB2
FGFR1
MET
SMARCA4
KEAP1
Other
High TMB