FACTRIAL: Factorized Clinical Contrastive Training for Scalable Patient-Trial Retrieval

Xuanren Chen[♣], **Chongyang Tao**[♣*], **Tao Shen**[△], **Shuai Ma**[♣]
[♣]SKLCCSE Lab, Beihang University [△] University of Technology Sydney
{cxrbuaa,chongyang,mashuai}@buaa.edu.cn {tao.shen}@uts.edu.au

Abstract

Patient–trial retrieval is a challenging problem that requires nuanced clinical reasoning beyond surface-level semantic similarity. However, scarce and costly relevance annotations force existing approaches to rely on very limited supervision or zero-shot transfer, reducing the task to generic semantic matching and failing to capture multi-factor eligibility reasoning. To this end, we propose FACTRIAL, a factorized contrastive training framework that leverages LLMs to synthesize diagnosis-aware supervision for scalable patient–trial retrieval. FACTRIAL decomposes each patient note into a primary diagnosis and a set of concomitant, eligibility-triggering conditions, and constructs complementary contrastive signals through structured trial augmentation. Specifically, we generate primary-target and concomitant-target positives, together with clinically confusable near-miss negatives, to enforce diagnostic specificity under contrastive learning. Two specialized bi-encoder experts are trained to balance primary-diagnosis prioritization and concomitant-driven recall, and fused into a single deployable retriever. Experiments on three public benchmarks demonstrate that FACTRIAL achieves state-of-the-art performance, improving both top-ranked quality and high-recall coverage.

1 Introduction

Clinical trials (National Institutes of Health, 2024) are carefully designed studies in which volunteer patients receive medical interventions so researchers can evaluate their safety and effectiveness. They are fundamental to the healthcare system, translating scientific discoveries into validated treatments, informing clinical guidelines, enabling regulatory approval, and ultimately improving patient outcomes through evidence-based treatments. As the scale and diversity of clinical trial repositories continue to grow (National Library of Medicine, 2025), effectively matching the right patients to the right trials has become increasingly critical for improving the cost-efficiency and success rate of clinical research (Brøgger-Mikkelsen et al., 2020). Yet patient-trial matching remains a major operational bottleneck: roughly one third of trials fail due to insufficient enrollment, recruitment costs consume about 32% of a trial's total budget (Insights, 2020), and 94% of patients are never informed about trials for which they may be eligible because screening is still largely manual, fragmented, and time-consuming (Nuttall, 2012; Kadam et al., 2016).

This drives the study of scalable patient-trial retrieval (Roberts et al., 2022a; Koopman and Zuccon, 2016), where given a patient characterized by an unstructured clinical note, the goal is to efficiently retrieve a small, high-quality set of relevant trials from a large-scale repository. This retrieval stage is a foundational component of downstream trial matching pipelines (Gupta et al., 2024; Jin et al., 2024; Wornow et al., 2025), supporting subsequent eligibility reasoning, fine-grained reranking, and clinician-in-the-loop review; it must therefore achieve both high recall, to avoid prematurely excluding suitable trials and compounding downstream errors, and high efficiency, to ensure scalability to continuously expanding trial corpora and real-world clinical workflows (Jin et al., 2024, 2021; Kusa et al., 2023).

To address this scalable retrieval problem, the common practice is to adopt a bi-encoder architecture (Karpukhin et al., 2020; Reimers and Gurevych, 2019), where all clinical trials are encoded and indexed offline based on their natural language descriptions, and a patient is represented as a single embedding derived from the clinical note that is used to query this index via efficient nearest-neighbor search. However, a lack of patient-trial relevance labels leads existing approaches to fall

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19127–19145

into two main paradigms in a zero-shot transferring manner (Kusa et al., 2023; Datta et al., 2025) i.e., off-the-shelf encoders or retrievers (Lee et al., 2020; Gu et al., 2021; Chen et al., 2024) and domain-specific query rewriting (e.g., TrialGPT-retrieval (Jin et al., 2024)).

The lack of relevance labels stems from the fact that patient-trial matching is not simple semantic relatedness or keyword overlap, but multi-factor eligibility matching that depends on nuanced clinical judgment (Alexander et al., 2020). Accurately annotating such relevance requires domain experts to jointly reason over diagnoses, etiology, treatments, and exclusion criteria, making large-scale labeling extremely time-consuming and resource-intensive.

To circumvent this challenge, we propose FACTRIAL, **Fa**ctorized **C**linical contrastive training via large language models-based data synthesis for scalable patient-trial retrieval. Specially, as shown in Figure 1, FACTRIAL first factorizes a patient note into a primary diagnosis and a set of concomitant, eligibility-triggering diagnoses, then uses these factors to synthesize complementary trial views: primary-target positives with primary-avoiding negatives, and concomitant-target positives with clinically confusable near-miss negatives, all verified for relevance and eligibility. We then train two contrastive bi-encoder experts specialized for primary-first ranking and concomitant-driven coverage, and fuse them into a single deployable retriever at the parameter level.

We evaluate FACTRIAL on three publicly available cohorts comprising 183 synthetic patients and over 75,000 trial-eligibility annotations. Across all cohorts, FACTRIAL achieves state-of-the-art patient–trial retrieval performance, improving both top-ranked quality (e.g., nDCG@10) and high-recall coverage (e.g., Recall@500). These results show that FACTRIAL effectively prioritizes trials aligned with a patient's primary diagnosis while preserving strong recall for trials triggered by concomitant conditions and avoiding clinically confusable near-miss matches. In summary, our main contributions are:

- We propose an LLM-based data synthesis framework that factorizes patient notes into primary and concomitant diagnoses and generates verified contrastive supervision for patient–trial retrieval.
- We introduce a dual-expert contrastive training and parameter fusion approach that integrates primary-diagnosis prioritization with concomitant-driven coverage in a single bi-encoder.
- We release an open-source, task-specific embedding model that achieves superior performance for patient-trial retrieval while reducing reliance on closed-source medical LLMs.

2 Preliminary: Patient–Trial Retrieval

Given a patient $p \in \mathcal{P}$ with note x_p and a trial corpus $\mathcal{T}$, patient-trial retrieval ranks trials $t \in \mathcal{T}$ so that trials for which p is both clinically relevant and eligible are ranked higher. Each trial t includes a title, summary, target diseases, intervention drugs, and inclusion and exclusion criteria ($\mathcal{I}_t$ and $\mathcal{E}_t$), as illustrated in Appendix A.

The retrieval objective is to learn a scoring function $\mathcal{R}(p, t)$ between x_p and t using a latency-efficient dual encoder that maps a text sequence x (representing either patient note p or trial t) into a d-dimensional embedding space:

$$\boldsymbol{x} = \text{Dual-Enc}(x; \theta) \in \mathbb{R}^d, \tag{1}$$

where θ parameterizes either a dense retriever or a lexicon-aware retriever based on LLM-based encoders. Match score is computed by a dot product:

$$\mathcal{R}(p, t; \theta) = \boldsymbol{x}_p^T \boldsymbol{t}. \tag{2}$$

The retriever need to align multi-faceted clinical factors in x_p with the eligibility constraints $(\mathcal{I}_t, \mathcal{E}_t)$. Trial embeddings in $\mathcal{T}$ are pre-computed and indexed offline, reducing inference to query encoding followed by Maximum Inner Product Search.

3 FACTRIAL

3.1 Model Overview

As shown in Figure 1, FACTRIAL constructs clinically aligned patient-trial retrieval via a two-stage pipeline. We first perform *factorized clinical trial synthesis* (§3.2), which decomposes each patient note into a primary diagnosis and concomitant diagnoses, and then synthesizes primary-target, concomitant-target, and near-miss trials with relevance/eligibility verification to produce structured contrastive supervision. We then conduct *factorized clinical contrastive training* (§3.3) by training two bi-encoder experts: a primary-diagnosis prioritization expert to enforce primary-first ranking, and a concomitant-diagnosis coverage expert to improve recall while rejecting diagnostically confusable near-miss matches. Finally, *parametric*

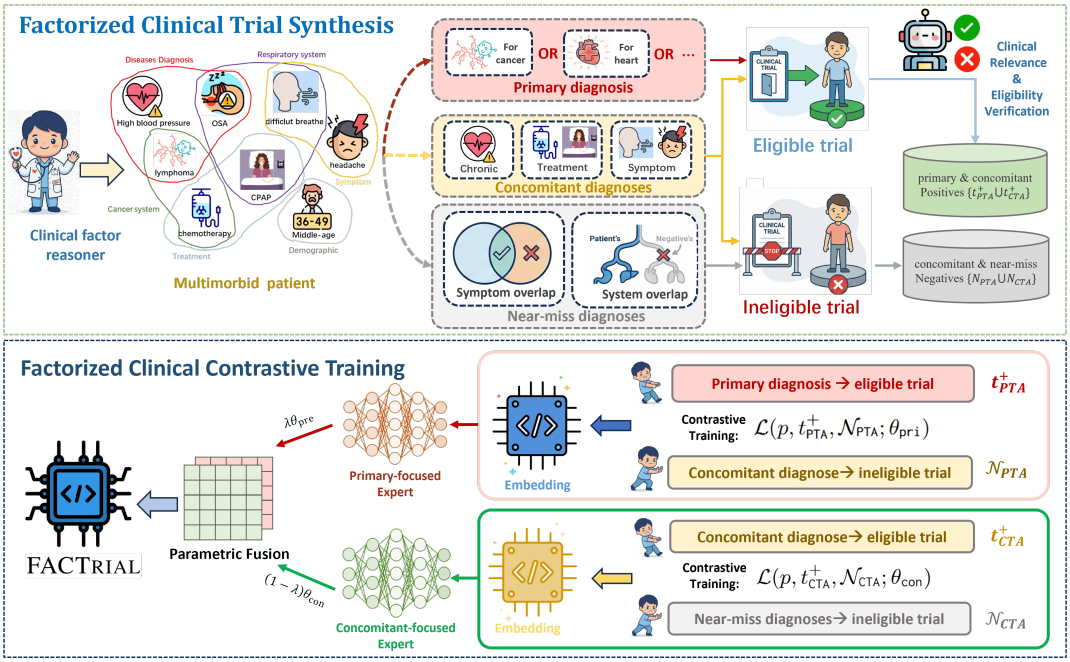

Figure 1: Overview of FACTRIAL. We first perform *factorized clinical trial synthesis* based on patient clinical factor decomposition process. We then conduct *factorized clinical contrastive training* by training two bi-encoder experts, and merge them into a single deployable retriever.

fusion interpolates the two experts into a single deployable retriever whose embedding space captures both prioritization and coverage.

3.2 Factorized Clinical Trial Synthesis

Large-scale patient-trial retrieval requires supervision reflecting *both* clinical relevance and eligibility compatibility. Yet existing datasets are sparse and weakly aligned: each patient is linked to only a small subset of trials, and eligibility criteria entangle primary diagnoses, concomitant diagnoses, and exclusion rules in a single representation. With such limited supervision, retrievers can overfit to superficial note correlations (e.g., shared symptoms or organ systems) instead of the discriminative factors that determine eligibility. This issue is amplified because many trials depend on context beyond diagnosis surface forms, making clinically related conditions highly confusable in text. For instance, two patients may both present with wound infections but differ in etiology (animal bite vs. fungal contamination); a trial targeting animal bite-related infections would include the former but exclude the latter despite high semantic similarity.

These observations motivate a *factorized* synthesis strategy that separates primary and concomitant patient factors and introduces hard contrasts along eligibility-defining attributes. By generating controlled patient-trial pairs that match or violate specific eligibility triggers, the resulting supervision provides discriminative signals that guide retrieval beyond symptom-level matching toward clinically grounded eligibility reasoning.

❶ Patient Clinical Factor Decomposition. Patient notes contain multiple clinical signals, but they do not contribute equally to trial relevance. To mirror clinical decision-making, we decompose a patient into a *primary diagnosis* that dominates treatment planning and a set of *concomitant diagnoses* that may independently affect eligibility. Given a note x_p, we infer the primary diagnosis d_0 using an LLM $\mathcal{M}_{\texttt{pri}}$ with a primary-focused instruction template:

$$d_0 \leftarrow \mathcal{M}_{\texttt{pri}}(x_p). \tag{3}$$

We then extract a diverse concomitant set $\mathcal{D}_p^{\text{con}}$ via a complementary template that encourages coverage while reducing redundancy:

$$\mathcal{D}_p^{\text{con}} = \{d_1, \ldots, d_K\} \leftarrow \mathcal{M}_{\texttt{con}}(x_p), \tag{4}$$

covering comorbidities, chronic conditions, symptoms, treatments, and other eligibility-relevant factors. This factorization reflects multiple enrollment pathways: some trials are driven by d_0, while others hinge on concomitant diagnoses or ongoing treatments. By separating dominant and accompanying factors, we enable retrieval to match diverse

enrollment rationales while preserving which factors should prioritize relevance ranking.

❷ Primary-focused Trial Augmentation (PTA). Among a patient's eligibility-triggering factors, the *primary diagnosis* drives treatment planning, so trials targeting it should be ranked above trials triggered by other factors. To encode this *relevance priority*, we build a *primary view* anchored on the inferred primary diagnosis. We synthesize a primary-target positive trial by prompting $\mathcal{G}$ with the full note x_p and primary diagnosis d_0:

$$t^{+}_{p,\mathsf{PTA}} \leftarrow \mathcal{G}(x_p, d_0). \tag{5}$$

To sharpen primary-diagnosis boundaries, we additionally synthesize *primary-avoiding* negatives $t^{-}_{p,\mathsf{PTA}}$ by prompting $\mathcal{G}$ with (x_p, d_0) while explicitly instructing it to generate patient-consistent trials whose target is *not* the primary diagnosis, yield target the concomitant factors:

$$\mathcal{N}_{\mathsf{PTA}} = \{t^{-}_{p,\mathsf{pri}}\} \leftarrow \{\mathcal{G}(x_p, d_0; \neg d_0)\}, \tag{6}$$

where $\neg d_0$ denotes avoidance of the primary diagnosis. This supervision fixes patient evidence but contrasts target diagnoses, pushing the retriever to allocate similarity to the primary diagnosis rather than concomitant factors in x_p.

❸ Concomitant-varying Trial Augmentation (CTA). Beyond primary-first ranking, high-recall retrieval must also cover trials matched via *concomitant diagnoses* (e.g., comorbidities, chronic diseases, symptoms, ongoing treatments) that independently trigger eligibility. Directly synthesizing trials from the full note x_p often entangles factors and is dominated by the primary diagnosis, yielding weak diagnosis-specific signals. We therefore create concomitant-target positives by prompting the generator with each extracted concomitant diagnosis $d \in \mathcal{D}^{\text{con}}_p$:

$$\{t^{+}_{p,\mathsf{CTA}}\} \leftarrow \{\mathcal{G}(d)\}_{d \in \mathcal{D}^{\text{con}}_p}. \tag{7}$$

Because clinically related diagnoses can be textually confusable, concomitant-target positives alone do not guarantee diagnostic specificity. We thus introduce *near-miss diagnoses*,which are clinically plausible alternatives that share salient evidence with true diagnoses but are diagnostically mismatched, to construct hard negatives. Specifically, we generate a set $\hat{\mathcal{D}}_p = \{\hat{d}_{p,1}, \ldots, \hat{d}_{p,K}\}$ by prompting an LLM with the patient note p and its primary, concomitant diagnoses, instruct the model to satisfy either (1) *System-overlap* (same clinical system, different etiology) or (2) *Symptom-overlap* (shared key symptoms, different organ system). We convert these near-misses into hard negatives by synthesizing trials that explicitly target them:

$$\mathcal{N}_{\mathsf{CTA}} = \{t^{-}_{p,\mathsf{CTA}}\} \leftarrow \{\mathcal{G}(d)\}_{d \in \hat{\mathcal{D}}_p}. \tag{8}$$

During training, each positive $t^{+}_{p,\mathsf{CTA}}$ is contrasted against these negatives, which are semantically close but diagnostically incorrect, encouraging the retriever to distinguish true concomitant factors from confusable alternatives.

❹ Clinical Relevance and Eligibility Verification. LLM-based trial synthesis can be noisy: a generated positive may be textually similar yet target a wrong diagnosis or violate inclusion/exclusion criteria. We therefore verify each synthesized pair (x_p, t) with an LLM $\mathcal{M}_{\mathtt{veri}}$, producing relevance and eligibility judgments:

$$\big(r_{p,t},\, e_{p,t}\big) = \mathcal{M}_{\mathtt{veri}}(x_p, t), \tag{9}$$

where $r_{p,t}$ measures *semantic relevance* (match to a clinical factor supported by x_p) and $e_{p,t}$ measures *clinical eligibility* (satisfies inclusion and avoids exclusion). Both use a 0–3 scale and are grounded in explicit evidence; missing support or conflicts imply ineligibility. We keep t as positive if $r_{p,t} > 0$ and $e_{p,t} > 0$, yielding *9,414* primary-driven and *26,793* concomitant-driven pairs.

3.3 Factorized Clinical Contrastive Training.

The supervision in §3.2 contains two competing signals: (i) a *primary-first ranking* signal that pushes primary-diagnosis trials above alternatives, and (ii) a *coverage* signal that expands recall through concomitant pathways while using near-miss contrasts to enforce diagnostic specificity. A single retriever trained on both signals can be pulled in opposite directions: over-emphasizing primary-first ranking weakens sensitivity to concomitant eligibility triggers, whereas optimizing for broad coverage can dilute the preference for primary-diagnosis trials and increase confusions among clinically related diagnoses. This conflict is especially pronounced in contrastive learning because the definition of "hard negatives" differs across views: primary-view negatives are non-primary targets (to sharpen prioritization), while concomitant-view negatives are near-miss targets (to prevent diagnostic confusion).

We therefore factorize training into two bi-encoder experts (primary-prioritization and concomitant-coverage) and merge them into a single deployable retriever via parametric fusion.

Primary-focused Contrastive Training (θ_{pri}). Although a patient note may support multiple enrollment pathways, retrieval should consistently prioritize trials matched via the primary diagnosis. Generic contrastive training can blur this preference because cues shared across pathways drive gradients toward broad semantic similarity. We therefore train a dedicated expert θ_{pri} on the primary-view supervision (§3.2) to encode primary-diagnosis prioritization. Given a positive trial $t^+_{\texttt{PTA}}$, augmented negative trials $\mathcal{N}_{\texttt{PTA}}$, and in-batch negatives $\mathcal{N}_{\texttt{B}}$, we optimize $\theta_{\texttt{pri}}$ with the InfoNCE loss:

$$\mathcal{L}(p, t^+_{\texttt{PTA}}, \mathcal{N}_{\texttt{PTA}}; \theta_{\texttt{pri}}) = -\log \frac{\exp\left(\mathcal{R}(x_p, t^+_{\texttt{PTA}})/\tau\right)}{\exp\left(\mathcal{R}(x_p, t^+_{\texttt{PTA}})/\tau\right) + \sum_{t \in \mathcal{N}_{\texttt{PTA}} \cup \mathcal{N}_{\texttt{B}}} \exp\left(\mathcal{R}(x_p, t)/\tau\right)}, \tag{10}$$

where τ is a temperature parameter and $\mathcal{R}$ is the retriever scoring function defined in Eq. 2.

Concomitant-focused Contrastive Training (θ_{con}). Primary-first ranking alone is insufficient for high-recall candidate generation because patients may match additional trials via concomitant diagnoses and other eligibility triggers. However, expanding coverage increases exposure to *clinical confounders*: diagnoses outside the patient profile can closely overlap in system or symptom expressions yet differ in eligibility-defining attributes (e.g., etiology or clinical context), causing a similarity-driven retriever to over-retrieve confusable alternatives. To improve concomitant coverage while controlling these confusions, we train a complementary expert θ_{con} on concomitant-view supervision. Following Eq. 10, we optimize θ_{con} with $\mathcal{L}(p, t^+_{\texttt{CTA}}, \mathcal{N}_{\texttt{CTA}}; \theta_{\texttt{con}})$, contrasting a positive $t^+_{\texttt{CTA}}$ synthesized from concomitant diagnoses with negatives $\mathcal{N}_{\texttt{CTA}}$ guided by near-miss diagnoses, encouraging discrimination along clinically defining eligibility attributes.

Parametric Fusion for Unified Retrieval. The two experts capture complementary signals: θ_{pri} enforces primary-first ranking by concentrating similarity on primary-diagnosis evidence, while θ_{con} improves concomitant coverage and suppresses near-miss confusions. Instead of score-level combination (which requires two encoders at inference and is sensitive to calibration), we perform *parametric fusion* to obtain a single retriever:

$$\theta_{\text{fuse}} = \lambda\, \theta_{\text{pri}} + (1 - \lambda)\, \theta_{\text{con}}. \tag{11}$$

We set $\lambda = 0.5$ to balance the objectives, yielding a single deployable bi-encoder with a unified embedding space; empirically, θ_{fuse} outperforms both individual experts and score-level aggregation.

Inference and Deployment. The factorized data construction and diagnosis-level reasoning described above are used solely to synthesize training supervision. At inference time, FACTRIAL operates as a standard dense retriever. Specifically, we use the fused parameters θ_{fuse} to encode all trials offline and build a retrieval index. Given a patient note at query time, the same fused bi-encoder directly encodes the raw note into the shared embedding space, and retrieval is performed via inner-product similarity without any explicit factor extraction or diagnosis inference.

4 Experiment

4.1 Experimental Setup

Evaluation Benchmarks. Following TrialGPT (Jin et al., 2024), we evaluate FACTRIAL on three public benchmarks: SIGIR 2016 (Koopman and Zuccon, 2016; Roberts et al., 2015; Simpson et al., 2014) and TREC CT 2021/2022 (Roberts et al., 2021, 2022b). SIGIR 2016 has 58 synthetic patients and 3,621 trials; TREC CT 2021/2022 have 75, 50 patients and 26,149, 26,581 trials. Each cohort provides expert relevance labels from patient notes and eligibility criteria (excluding geolocation and recruitment status): SIGIR uses (*irrelevant*, *potential*, *eligible*) and TREC CT uses (*irrelevant*, *excluded/ineligible*, *eligible*). To test robustness to many unlabeled trials, we additionally collect $451,538$ public trials from ClinicalTrials.gov to form an expanded corpus and re-evaluate all cohorts with the original annotations. More details in Appendix B.

Training Data. For training, we sample patients from the public PMC-Patients (PMC-v2) corpus (Zhao et al., 2023). We exclude *deceased* patients and those *fully recovered without follow-up symptoms*, who are typically ineligible for or unlikely to enroll in clinical trials. After filtering and the verification in § 3.2, we retain 9,414 patients, all

Method	SIGIR			TREC_2021			TREC_2022			AVG		
	MAP	nDCG@10	R@500	MAP	nDCG@10	R@500	MAP	nDCG@10	R@500	MAP	nDCG@10	R@500
BM25	7.46	7.93	44.79	19.58	32.97	34.94	15.27	26.97	34.31	14.10	22.62	38.01
bge-m3	14.80	12.23	62.96	45.57	46.89	61.02	31.28	38.27	41.18	30.55	32.46	57.05
nomic-embed-text-v1.5	18.28	13.30	69.14	54.87	52.81	67.73	46.20	46.29	59.69	39.78	37.47	69.52
Qwen3-embedding-0.6B	20.73	15.84	72.73	62.75	54.82	71.55	54.96	50.98	66.22	46.15	40.55	70.17
Qwen3-embedding-4B	34.08	25.69	83.86	73.43	60.93	77.29	69.74	57.24	71.66	59.08	47.95	77.60
TrialGPT-retrieval	46.04	31.65	93.57	76.27	55.65	83.71	85.84	61.11	81.41	69.38	49.47	86.23
FACTRIAL-0.6B	35.63	24.55	90.36	88.46	62.89	86.30	79.98	58.15	79.03	68.02	48.53	85.23
FACTRIAL-4B	**55.57**	**36.33**	**97.09**	**97.40**	**66.85**	**88.65**	**95.98**	**68.82**	**83.68**	**82.98**	**57.33**	**89.81**

Table 1: Main retrieval results on three cohorts (%). Best results are in bold.

disjoint from the evaluation cohorts. Each patient is represented by a clinical note, which serves as the query participating the data synthetic process.

Metrics. We treat each patient as a query and rank candidate trials. We report MAP, nDCG@10, and Recall@500 (higher is better), measuring overall ranking, top-10 quality, and relevant-trial coverage in the top 500, respectively. Metrics are computed per patient, averaged within each cohort, and then unweighted-averaged across cohorts.

Baselines. We compare FACTRIAL against representative sparse and dense retrievers and a state-of-the-art LLM-assisted pipeline under a unified setup (patient note as query; trial description as document). BM25 (Robertson et al., 2009) is the sparse lexical baseline. For encoder-only dense retrieval, we evaluate bge-m3 (Chen et al., 2024) and nomic-embed-text-v1.5 (Nussbaum et al., 2024). For decoder-only embedding models, we include Qwen3-embedding-0.6B and Qwen3-embedding-4B (Zhang et al., 2025). We also compare to TrialGPT-Retrieval (Jin et al., 2024), an LLM-assisted pipeline that uses GPT-4 to aid retrieval.

Implementation Details. We use Qwen3-32B (Team, 2025) for data synthesis. For embedding training, FACTRIAL is initialized from Qwen3-Embedding models (Zhang et al., 2025), including both the 0.6B and 4B variants, and fine-tuned on our synthesized data using an InfoNCE-style contrastive objective. We set the temperature to $\tau = 0.1$, enable in-batch negatives. Training is conducted for one epoch with a per-device batch size of 4, gradient accumulation over 2 steps, and a global batch size of 32. We use 2 hard negatives for PTA and 5 hard negatives for CTA. The learning rate is set to 8×10^{-6} with a warmup ratio of 0.1, a cosine learning-rate schedule, weight decay of 0.01, and gradient clipping with a maximum norm of 1.0. For model fusion, we set $\lambda = 0.5$, which already achieves strong and stable performance across benchmarks; therefore, we do not further tune this hyperparameter. All prompt templates are available in Appendix C.

4.2 Main Results

Table 1 reports results on three patient cohorts. FACTRIAL-4B achieves the best performance across all metrics and cohorts, setting a new state of the art. Averaged over cohorts, it improves over the strongest prior pipeline, TrialGPT-retrieval, by **13.6** MAP and **7.9** nDCG@10, while also increasing high-recall coverage by **3.58** Recall@500. The gains hold across benchmarks, including SIGIR (smaller pool, higher lexical variance) and the two TREC CT cohorts (larger pools, more criteria-heavy trials), indicating robust generalization across cohort composition and pool scale.

We observe a clear recall gap between *specialized* systems (TrialGPT-retrieval, FACTRIAL) and general-purpose retrievers. For example, FACTRIAL-4B reaches overall 89.91 R@500, substantially above Qwen3-embedding-4B, 77.60 overall R@500. This reflects *multiple enrollment pathways* in patient notes: relevant trials may be triggered not only by the primary diagnosis but also by concomitant diagnoses and other eligibility factors. Generic retrievers often focus on globally dominant evidence (typically the primary diagnosis), underretrieving secondary-factor matches. TrialGPT-retrieval expands coverage via LLM-based query expansion, whereas FACTRIAL internalizes this behavior through factorized supervision that explicitly trains for concomitant-driven matches, yielding further gains in both R@500 and ranking quality.

FACTRIAL also substantially improves MAP and nDCG@10, suggesting better ordering of clinically plausible candidates rather than sim-

Method	SIGIR			TREC_2021			TREC_2022			AVG		
	nDCG@10	R@500	R@1000	nDCG@10	R@500	R@1000	nDCG@10	R@500	R@1000	nDCG@10	R@500	R@1000
BM25	1.05	6.81	9.19	21.22	16.59	20.40	22.39	17.02	21.04	14.89	13.47	16.88
bge-m3	1.73	13.04	17.75	28.26	30.20	39.33	27.08	25.13	32.67	19.02	22.79	29.92
nomic-embed-text-v1.5	1.41	13.57	19.04	24.91	32.24	42.15	24.40	30.90	39.07	16.91	25.57	33.42
Qwen3-embedding-0.6B	2.17	19.89	25.78	37.20	47.78	57.78	39.61	47.71	57.37	26.33	38.46	46.98
Qwen3-embedding-4B	4.21	31.13	39.23	40.97	51.53	62.54	44.22	53.75	63.26	29.80	45.47	55.01
FACTRIAL-0.6B	1.20	28.21	35.43	39.63	55.61	66.56	41.94	56.43	66.08	27.59	46.75	56.02
FACTRIAL-4B	**4.03**	**40.60**	**50.24**	**42.28**	**58.19**	**69.65**	**50.53**	**62.20**	**71.91**	**32.28**	**53.66**	**63.93**

Table 2: Larger-corpus retrieval results on SIGIR, TREC_2021, and TREC_2022 (%). Best results are in bold.

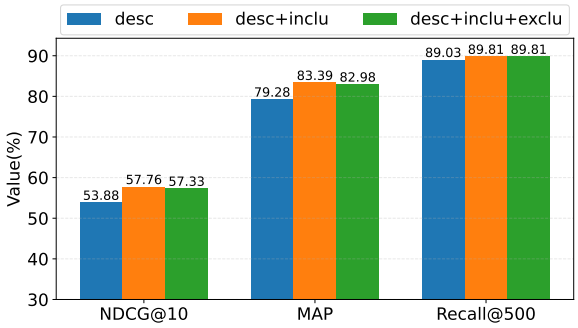

Figure 2: The contribution of different trial fields (e.g., description, inclusion criteria, and exclusion criteria).

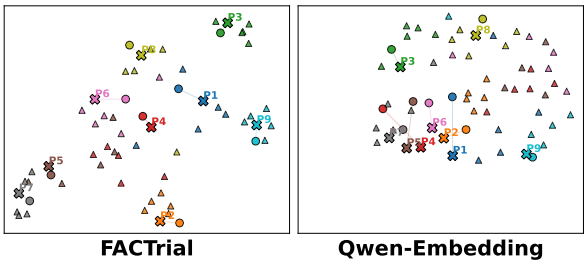

Figure 3: Case study via embedding visualization. Markers indicate types: patient (×), primary-target trial (○), and concomitant-target trial (△). The same color indicating a patient and its matched trials.

ply retrieving more. In contrast, non-specialized dense retrievers (encoder-only and decoder-only) often achieve moderate recall but lag in MAP/nDCG@10, consistent with over-reliance on broad semantic similarity that fails to separate diagnostically confusable alternatives.

Finally, comparing FACTRIAL to its Qwen3-embedding backbones isolates the impact of our training recipe: at both scales, FACTRIAL yields large absolute gains, showing improvements beyond model scale. Notably, FACTRIAL-0.6B is competitive with TrialGPT-retrieval despite being much smaller, suggesting that structured factorized supervision can compensate for parameter count in dense patient–trial retrieval.

4.3 Analysis

Performance Robustness on Large-Scale Trial Collections. Table 2 reports results when each cohort is evaluated against an expanded corpus of 451,538 trials. FACTRIAL-4B remains best overall, achieving the top average nDCG@10 and Recall@500/1000 and substantially improving coverage over strong dense baselines. The gains are most pronounced on recall: FACTRIAL-4B retrieves more judged-relevant trials within the top ranks, indicating robustness when the pool is dominated by unlabeled distractors. Absolute scores (especially on SIGIR) are lower and noisier because large-corpus evaluation has incomplete judgments: newly added trials are unjudged yet treated as non-relevant, which underestimates true performance. Despite this bias, FACTRIAL maintains clear margins over sparse (BM25) and dense (Qwen3-embedding) baselines, supporting that our hierarchical supervision improves scalable candidate generation in large-pool retrieval.

The Impact of Different Trial Fields. Figure 2 analyzes trial-field contributions by progressively adding eligibility text. Augmenting descriptive fields with *inclusion criteria* (`desc` → `desc+inclu`) yields the largest and most consistent gains, suggesting inclusion statements provide strong alignment cues for matching patient notes to trial. In contrast, adding *exclusion criteria* (`desc+inclu` → `desc+inclu+exclu`) offers little benefit and can slightly hurt performance. This reflects a limitation of embedding retrieval: exclusion clauses rely on negation, exceptions, and multi-clause logic that are hard to encode in a single vector and may introduce noisy lexical overlap. Overall, FACTRIAL 's gains appear driven mainly by inclusion-style evidence, motivating reranking or explicit constraint modeling.

Model / Setting	MAP	nDCG@10	R@500
FACTRIAL-4B	**82.98**	**57.33**	**89.81**
4B (Train w.CTA)	75.66	53.61	87.87
4B (Train w.PTA)	80.12	56.24	87.54
4B (Train w.PTA+CTA)	75.17	52.84	87.06
4B (Fusion w.RRF)	81.51	56.83	89.56
4B (Fusion w.Score)	81.33	56.52	88.11

Table 3: Ablation results.

Case Study via Embedding Visualization. We examine how FACTRIAL reshapes the patient–trial embedding space. We randomly sample 10 patients and manually annotate, per patient, one *primary*-target trial and five *concomitant*-triggered trials (60 pairs). We visualize embeddings with UMAP (McInnes et al., 2018), comparing the encoder *before* vs. *after* FACTRIAL training using identical hyperparameters and a fixed seed, and align the coordinates across plots. As shown in Figure 3, before training, concomitant trials are dispersed and sometimes closer to other patients' trials, reflecting weak patient-specific binding. After training, FACTRIAL pulls concomitant trials toward their patient while keeping the primary trial nearest, yielding tighter patient-centric neighborhoods and supporting the patient-trial retrieval goals.

4.4 Ablation Study

Comparison of Different Training Recipes. Table 3 studies training recipes. Single-view training is already effective: `Train w.PTA` improves MAP and nDCG@10 over `Train w.CTA`, consistent with explicitly encoding primary-first prioritization, while `Train w.CTA` provides broader coverage signals. However, jointly training with one encoder (`Train w.PTA+CTA`) is consistently worse than `Train w.PTA` for both sizes and sometimes regresses toward `Train w.CTA`. This indicates supervision conflict: concomitant-triggered trials are positives in CTA but primary-avoiding negatives in PTA, yielding inconsistent gradients under a single objective. In contrast, view-specialized experts plus fusion (FACTRIAL) performs best, showing that decoupling prioritization and coverage is more effective than mixing them.

Parameter-level Fusion vs. Retrieval Fusion. We compare parameter-level fusion to two late-fusion baselines that combine expert outputs at inference: `Fusion w.Score` sums dot-product scores, and `Fusion w.RRF` applies reciprocal rank fusion (RRF) (Cormack et al., 2009). Both improve over individual experts, confirming complementary signals, but FACTRIAL is consistently best on average metrics. This suggests parameter-space fusion better integrates the two competencies into a unified embedding geometry, whereas late fusion only aggregates rankings without a coherent representation. It also reduces inference cost by requiring a single encoder and retrieval run.

5 Related Work

Matching between Patient and Trial. Patient-trial matching is often cast as retrieval between free-text patient records and structured trial descriptions, with two complementary paradigms: *patient-to-trial* (retrieving eligible trials for a patient) (Roberts et al., 2022b) and *trial-to-patient* (retrieving candidate patients for a trial) (Stubbs et al., 2019). Consistent with general retrieval settings (Craswell et al., 2025; Thakur et al., 2021), it is typically a two-stage pipeline: an efficient retriever recalls a high-coverage candidate set from a large trial pool, followed by an expensive reranker for fine-grained eligibility reasoning (e.g., cross-encoders or LLM-assisted judging) (Frihat and Fuhr, 2021; da Costa Pereira, 2022; Kusa et al., 2023; Jin et al., 2021; Nievas et al., 2024). We focus on first-stage retrieval because missed relevant trials cannot be recovered downstream, and retrieval quality largely determines end-to-end performance.

Large-scale Patient–Trial Retrieval. Large-scale patient–trial retrieval largely follows two paradigms. Early work uses sparse lexical methods (TF–IDF, BM25), sometimes with heuristic field weighting or rule-based eligibility matching (Frihat and Fuhr, 2021; Caucheteur et al., 2021; Ji et al., 2021). While efficient, sparse retrieval is sensitive to lexical mismatch between narrative notes and structured trial text, hurting recall. Later work adopts off-the-shelf encoders, including biomedical models (BioBERT, PubMedBERT) (Lee et al., 2020; Gu et al., 2021) and general embedding models (Chen et al., 2024). Despite semantic generalization, dense retrieval struggles with long trial documents (especially eligibility criteria), limited context, and scarce supervision, making clinically confusable diagnoses and context-dependent triggers hard to separate (da Costa Pereira, 2022; Kusa et al., 2023; Jin et al., 2021).

A second line uses task-specific query rewriting/evidence extraction, often with LLMs, to pro-

duce structured features or patient-specific expansions. TrialGPT-retrieval (Jin et al., 2024) uses GPT-4 to generate and rank up to 32 keywords; related work similarly relies on LLM-generated keywords or features (Datta et al., 2025). These approaches are effective but costly and may depend on closed models. In contrast, we build diagnosis-structured synthetic supervision that models multiple eligibility pathways and hard confounders, enabling a single dense retriever to improve coverage and ranking without LLM inference.

6 Conclusion

We address a practical gap in patient-trial matching: affordable, open retrievers for large-scale candidate generation. We propose FACTRIAL, which factorizes patient factors for supervision to prioritize primary-diagnosis matches, expand concomitant-pathway coverage, and suppress near-miss confounders, then unifies the two experts via parametric fusion. Across three public cohorts, FACTRIAL consistently improves ranking quality and high-recall coverage, providing a privacy-friendly retriever for large-scale patient-trial retrieval.

7 Limitations

While FACTRIAL provides an effective task-specific retriever for large-scale patient–trial candidate generation, it remains an embedding-based dense retriever and thus is primarily driven by semantic similarity. As a result, it is not well-suited for fine-grained logical eligibility reasoning, especially constraints involving negation, exceptions, or multi-clause logic commonly expressed in exclusion criteria. In practice, trials that should be filtered out by exclusion rules may still be retrieved and require a downstream reranking or eligibility-checking stage to enforce logical constraints.

In addition, compared to lightweight lexical or general-purpose embedding retrievers (e.g., BM25 or bge-style models), the 4B-scale encoder used in FACTRIAL incurs higher computational cost at inference time. However, this overhead remains within the budget of first-stage retrieval and is justified by the substantial gains in ranking quality and high-recall coverage observed across benchmarks. An important direction for future work is to design retrieval pipelines that better integrate explicit logical constraint modeling, either by incorporating structured reasoning modules during retrieval or by tighter coupling between retrieval and rule-based or LLM-based eligibility verification.

Although FACTRIAL does not rely on LLMs at inference time, its training supervision is derived from LLM-generated clinical reasoning, which may introduce systematic biases or stylistic artifacts. To mitigate this risk, we apply explicit relevance and eligibility verification and evaluate the resulting retriever on independently annotated public benchmarks, where FACTRIAL consistently achieves strong performance. Nevertheless, it remains an open research question how biases originating from LLM-generated supervision propagate into learned embedding spaces, how they interact with contrastive objectives, and to what extent they affect generalization beyond the evaluated cohorts.

8 Ethical Considerations

This work studies patient-trial retrieval using publicly available benchmark datasets and openly released clinical trial documents. All experiments are conducted on de-identified data provided by the dataset creators, and we do not collect, access, or process any private patient records beyond what is already included in these public resources. Our models are trained and evaluated solely within this setting, and no additional personally identifiable information is introduced.

Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 62572034, U22B2021, and U24B20143), State Key Laboratory of Complex & Critical Software Environment (SKLCCSE), and CIE–Tianyi Cloud Research Program.

References

Marliese Alexander, Benjamin Solomon, David L Ball, Mimi Sheerin, Irene Dankwa-Mullan, Anita M Preininger, Gretchen Purcell Jackson, and Dishan M Herath. 2020. Evaluation of an artificial intelligence clinical trial matching system in australian lung cancer patients. *JAMIA Open*, 3(2):209–215.

Mette Brøgger-Mikkelsen, Zarqa Ali, John R Zibert, Anders Daniel Andersen, and Simon Francis Thomsen. 2020. Online patient recruitment in clinical trials: systematic review and meta-analysis. *Journal of medical Internet research*, 22(11):e22179.

Déborah Caucheteur, Emilie Pasche, Luc Mottin, Anaïs Mottaz, Julien Gobeill, and Patrick Ruch. 2021. SIB text mining at TREC clinical trials 2021. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M Voorhees, and Ian Soboroff. 2025. Overview of the trec 2022 deep learning track. *arXiv preprint arXiv:2507.10865*.

Joao da Costa Pereira. 2022. Neural retrieval models for matching patients to clinical trials.

Surabhi Datta, Kyeryoung Lee, Liang-Chin Huang, Hunki Paek, Roger Gildersleeve, Jonathan Gold, Deepak Pillai, Jingqi Wang, Mitchell K Higashi, Lizheng Shi, and 1 others. 2025. Patient2trial: From patient to participant in clinical trials using large language models. *Informatics in Medicine Unlocked*, 53:101615.

Sameh Frihat and Norbert Fuhr. 2021. Trec 2021 clinical trials retrieval, duisburg-essen university submission. In *TREC*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Shashi Gupta, Aditya Basu, Mauro Nievas, Jerrin Thomas, Nathan Wolfrath, Adhitya Ramamurthi, Bradley Taylor, Anai N Kothari, Regina Schwind, Therica M Miller, and 1 others. 2024. Prism: Patient records interpretation for semantic clinical trial matching system using large language models. *NPJ digital medicine*, 7(1):305.

Deloitte Insights. 2020. Intelligent clinical trials transforming through ai-enabled engagement. *Retrieved November*, 16:2020.

Yanqing Ji, Yun Tian, Hao Ying, and John Tran. 2021. Clinical trial search using lucene and UMLS. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Zheng Yuan, and Songfang Huang. 2021. Alibaba damo academy at trec clinical trials 2021: Exploringembedding-based first-stage retrieval with trialmatcher. In *TREC*.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. 2024. Matching patients to clinical trials with large language models. *Nature communications*, 15(1):9074.

Rashmi Ashish Kadam, Sanghratna Umakant Borde, Sapna Amol Madas, Sundeep Santosh Salvi, and Sneha Saurabh Limaye. 2016. Challenges in recruitment and retention of clinical trial subjects. *Perspectives in clinical research*, 7(3):137–143.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Bevan Koopman and Guido Zuccon. 2016. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 669–672.

Wojciech Kusa, Óscar E Mendoza, Petr Knoth, Gabriella Pasi, and Allan Hanbury. 2023. Effective matching of patients to clinical trials using entity extraction and neural re-ranking. *Journal of biomedical informatics*, 144:104444.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

National Institutes of Health. 2024. Nih's definition of a clinical trial. https://grants.nih.gov/policy-and-compliance/policy-topics/clinical-trials/definition. Last updated: Sep 18, 2024. Accessed: 2026-01-05.

National Library of Medicine. 2025. Clinicaltrials.gov: A 25-year journey to a half-million registered studies. https://nlmdirector.nlm.nih.gov/2025/04/02/clinicaltrials-gov-a-25-year-journey-to-a-half-million-registered-studies/. Accessed: 2026-01-05.

Mauro Nievas, Aditya Basu, Yanshan Wang, and Hrituraj Singh. 2024. Distilling large language models for matching patients to clinical trials. *Journal of the American Medical Informatics Association*, 31(9):1953–1963.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Aidan Nuttall. 2012. Considerations for improving patient recruitment into clinical trials. *Clinical Leader Newsletter*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick, and William R. Hersh. 2022a. Overview of the trec 2022 clinical trials track. In *Text Retrieval Conference*.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and Willian R Hersh. 2021. Overview of the trec 2021 clinical trials track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and Willian R Hersh. 2022b. Overview of the trec 2022 clinical trials track. In *Proceedings of the Thirty-first Text REtrieval Conference (TREC 2022)*.

Kirk Roberts, Matthew S Simpson, Ellen M Voorhees, and William R Hersh. 2015. Overview of the trec 2015 clinical decision support track. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Matthew S Simpson, Ellen M Voorhees, and William R Hersh. 2014. Overview of the trec 2014 clinical decision support track. In *Proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014)*.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. 2019. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W Mahaffey, and Nigam H Shah. 2025. Zero-shot clinical trial patient matching with llms. *NEJM AI*, 2(1):AIcs2400360.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(1):909.

A Examples of Patient Note and Clinical Trial

Patient Note (Example)

1. A 67-year-old woman comes to the clinic due to recent episode of choking, dysphagia, and cough.
2. Her other medical problems include hypertension, dyslipidemia, and osteoarthritis.
3. She does not smoke or use alcohol.
4. She lives with her husband and she is able to do her own daily activities.
5. She used to teach elementary school. Blood pressure is 135/80 mm Hg.
6. The patient's breath smells bad. Other physical examinations are normal.
7. A barium swallow study reveals an abnormality in the upper esophagus with an outpouching at the junction of the lower part of the throat and the upper portion of the esophagus.

Trial (Example): Esophageal High Resolution Manometry and Dysphagia

Title	Esophageal High Resolution Manometry and Dysphagia
Diseases	Dysphagia
Drugs/Interventions	Conventional manometry; High resolution manometry
Inclusion	• Male or female older than 18 years • Patient with unexplained dysphagia • Patient without cause of dysphagia on eso-gastro-duodenal endoscopy • Patient referred for esophageal manometry • Patient with health insurance • Informed consent signed
Exclusion	• Patient younger than 18 years • Allergy to one component of manometry catheter • Drug intake which can modify esophageal motricity within 12 hours preceding manometry • Patient unable to give consent or legally incompetent • Patient non qualified according to the investigator • Patient refusal or absence of informed consent signed • Concomitant participation to another study
Brief Summary	Two to 15% of subjects present dysphagia. In case of normal eso-gastro-duodenal endoscopy, patients with dysphagia are referred for esophageal motility testing. Esophageal manometry is the gold standard to evaluate esophageal motility in absence of esophageal obstruction. Two different techniques are available: conventional manometry and high resolution manometry. The second one may improve the diagnostic yield and tolerance of examination in patients with dysphagia.

B Details of Three Patient Cohorts

Cohort	SIGIR	TREC 2021 CT	TREC 2022 CT
Number of patients	58	75	50
Age (year)	38.5 ± 23.7	41.6 ± 19.4	35.3 ± 20.2
Sex (male:female)	29:29	38:37	28:22
Note length (words)	88.7 ± 36.8	156.2 ± 45.4	109.9 ± 21.6
Eligible trials/patient	7.3 ± 6.7	74.3 ± 49.0	78.8 ± 67.3
Potential trials/patient	11.7 ± 10.2	None	None
Excluded trials/patient	None	80.3 ± 60.3	60.7 ± 65.5
Irrelevant trials/patient	47.1 ± 19.5	323.2 ± 93.2	568.4 ± 164.1
Considered initial trials	3621	26149	26581

Table 4: Cohort statistics for the three evaluation datasets.

C Prompt for Clinical Trial Synthesis

C.1 Prompt for Primary Diagnosis Decomposition ($d_0 \leftarrow \mathcal{M}_{\texttt{pri}}(x_p)$)

Task Overview
You are a professional doctor with a strong background in pharmacology and rich experience in clinical diagnosis.Given a patient note, you should follow the formulated workflow and accomplish the required tasks.

Patient

```
{Patient}
```

Instructions
Your workflow has 2 steps:

- Step1:Identify if the note contains an explicit current diagnosis (not past medical or surgical history) for the patient's active condition. If present, extract the diagnosis verbatim or summarize it concisely. If absent, infer the most likely current diagnosis based on the documented symptoms, clinical findings, and course of events.
- Step2:Analyze the note and classify the patient's status: 1.If the note explicitly states or provides clear evidence that the patient has died. 2.If the note clearly indicates that the primary condition being treated has been resolved.(If the patient is cured.)

Output Format (strict JSON)

```
{
   "if_death": str(yes|no),
   "if_cure": str(yes|no),
   "diagnosis":str,
   "rationale":str
}
```

"if_death" is yes means the patient has dead, otherwise no.
"if_cure" is yes means the patient has cured, otherwise no.
"diagnosis" is the current diagnosis extracted from the patient note or the inferred current diagnosis.

Figure 4: The Prompt for Primary Diagnosis Decomposition.

C.2 Prompt for Concomitant-Diagnosis Decomposition ($\mathcal{D}_p^{\text{con}} = \{d_1, \ldots, d_K\} \leftarrow \mathcal{M}_{\textsf{con}}(x_p)$)

Task Overview
You are a clinical abstraction assistant. Given a patient note, your task is to extract multiple clinically meaningful positive factors that jointly characterize the patient's medical profile.

Patient

```
{Patient}
```

Instructions
You should generate exactly **five concomitant factors** based on the patient note, following these principles:

- Each factor should represent **one major clinical or population-relevant dimension** derived from the patient text.
- The five factors must cover **different aspects** of the patient's medical profile (e.g., presenting symptom, lifestyle factor, structural abnormality, demographic characteristic, clinical behavior, or chronic comorbidity).
- Avoid redundancy: do not restate the same disease or perspective in multiplefactors.
- Maintain both **clinical accuracy** and **diversity** across the generated factors.

Output Format (strict JSON)

```
{
   "positive_factors": list[str],
   "rationale": str
}
```

Output Requirements

- `positive_factors` should be a list of five concise but informative clinical statements.
- `rationale` should briefly explain the reasoning behind the selected positive factors and how diversity across clinical dimensions was ensured.

Figure 5: The Prompt for Concomitant Diagnosis Decomposition

C.3 Prompt for Near-miss Diagnoses Construction ($\hat{\mathcal{D}}_p = \{\hat{d}_{p,1}, \ldots, \hat{d}_{p,K}\}$)

Task Overview
You are a clinical abstraction assistant. Given a patient note, the primary diagnosis, and a set of positive factors, your task is to construct clinically realistic **near-miss (negative) diagnoses** that may resemble the patient's presentation but do not apply to the patient.

Patient

```
{Patient}
```

Primary Diagnosis

```
{Primary Diagnosis}
```

Positive factors

```
{Positive_Factors}
```

Instructions
You should generate exactly **five negative factors**, following these constraints:

- Each negative factor must be a **stand-alone clinical diagnosis**, written as a complete sentence describing a coherent disease entity with its typical presentation.
- Each factor should describe a diagnosis whose symptoms or manifestations may **resemble aspects of the patient's presentation**, but the diagnosis itself does **not apply** to the patient.
- Do **not** discuss or imply any differences between the negative factors and the patient's true condition.
- Negative factors may originate from the **same organ system with different etiologies** (e.g., neoplastic, infectious, inflammatory, structural, functional) or from **different systems** that can produce overlapping symptoms.
- Do **not** include negations, contrasts, exclusions, or meta-level statements such as "without…", "no evidence of…", "but lacking…", or references to the positive factors.
- Ensure **etiologic diversity** while maintaining **clinical realism**.

Output Format (strict JSON)

```
{
   "negative_factors": list[str],
   "rationale": str
}
```

Output Requirements

- `negative_factors` must contain exactly five independent clinical diagnoses.

Figure 6: The Prompt for Near-miss Diagnosis Construction

C.4 Prompt for Primary-focused Eligible Trial Generation ($t^{+}_{p,\textsf{PTA}} \leftarrow \mathcal{G}(x_p, d_0)$)

Task Overview
You are given a patient's clinical note and their current diagnosis. Your task is to generate one medically plausible and realistically eligible clinical trial that explicitly targets the patient's current diagnosis.

Patient

```
{Patient}
```

Current Diagnosis

```
{Primary Diagnosis}
```

Instructions
You should generate exactly **one eligible clinical trial** according to the following requirements:

- The trial must be **realistic**, **medically plausible**, and explicitly designed to treat or study the given current diagnosis.
- The patient should clearly satisfy **most inclusion criteria** and violate **none of the exclusion criteria**.
- The trial should include a concise and authentic title (e.g., "A Phase II Study of . . . "), a brief summary (2–3 sentences), and realistic drugs, diseases, inclusion criteria, and exclusion criteria.
- Use **natural clinical language** consistent with real clinical trial registry entries (e.g., ClinicalTrials.gov).
- Do **not** include unrelated diseases, past conditions, or vague or speculative details.

Output Format (strict JSON)

```
{
   "title": str,
   "brief_summary": str,
   "drugs": list[str]  or  [ ],
   "diseases": list[str],
   "inclusion_criterion": list[str],
   "exclusion_criterion": list[str]  or  [ ],
   "rationale": str
}
```

Output Requirements

- `title` should reflect a realistic clinical study style (e.g., phase, intervention, disease focus).
- `brief_summary` should concisely describe the study objective and intervention.
- `drugs` and `diseases` must be clinically appropriate for the given diagnosis.
- `rationale` should clearly explain why the patient is eligible for the trial and how the trial directly targets the patient's current diagnosis.

Figure 7: The Prompt for Primary-focused Eligible Trial Generation

C.5 Prompt for Primary-focused Ineligible Trial Generation ($\mathcal{N}_{\textbf{PTA}} = \{t^{-}_{p,\textbf{pri}}\} \leftarrow \{\mathcal{G}(x_p, d_0; \neg d_0)\}$)

Task Overview
You are given a patient's clinical note and their current diagnosis. Your task is to generate **two clinically realistic but ineligible clinical trials** that may appear related to the patient's record, yet the patient is explicitly disqualified due to disease scope, timing, or clinical context.

Patient

```
{Patient}
```

Current Diagnosis

```
{Primary Diagnosis}
```

Instructions
You should generate exactly **two ineligible clinical trials** that satisfy the following constraints:

- The trials must **not directly target** the patient's current diagnosis.
- Each trial should focus on a **related disease, procedure, or physiological finding** mentioned in the patient note (e.g., past medical history, prior surgery, laboratory abnormality, or a clinically adjacent condition).
- Each trial should appear plausibly connected to the patient's record, such as referencing:
  - a treatment the patient has already received,
  - a similar but distinct disease subtype,
  - or a diagnostic or procedural element described in the note.
- The patient must be **explicitly ineligible**, for example:
  - the trial targets preoperative patients while the patient is postoperative;
  - the trial studies a related but distinct disease stage or subtype;
  - the trial excludes patients with prior surgery, chemotherapy, or other interventions already received by the patient.
- Maintain **clinical realism** and use terminology consistent with real-world clinical trial descriptions.

Output Format (strict JSON)

```
[
  {
    "title": str,
    "brief_summary": str,
    "drugs": list[str]  or  [ ],
    "diseases": list[str],
    "inclusion_criterion": list[str],
    "exclusion_criterion": list[str]  or  [ ],
    "rationale": str
  },. . .
]
```

Figure 8: The Prompt for Primary-focused Ineligible Trial Generation

C.6 Prompt for Concomitant-varying Trial Generation (Both Concomitant ($\{t^{+}_{p,\mathsf{CTA}}\} \leftarrow \{\mathcal{G}(d)\}_{d\in\mathcal{D}_p^{\text{con}}}$) and Near-miss Diagnoses ($\mathcal{N}_{\mathsf{CTA}} = \{t^{-}_{p,\mathsf{CTA}}\} \leftarrow \{\mathcal{G}(d)\}_{d\in\hat{\mathcal{D}}_p}$)

Task Overview

You are an experienced clinical trial protocol writer with expertise in regulatory-compliant study design, medical accuracy, and realistic trial structuring consistent with entries found on ClinicalTrials.gov. Your task is to generate one medically plausible and realistic clinical trial based on a given patient factor.

Factor

```
{Concomitant diagnosis / near-miss diagnosis}
```

Instructions

You should generate exactly **one clinical trial** that satisfies the following requirements:

- The trial must be **medically plausible** and may vary in study type (e.g., interventional, observational, or diagnostic).
- Use **realistic interventions**, avoiding fictional drugs or implausible study designs.
- The trial structure and language should resemble authentic entries in clinical trial registries such as ClinicalTrials.gov.
- The `rationale` must clearly explain how the given factor determines patient eligibility and how the trial's objective aligns with that factor.
- Maintain a **clean and consistent format** suitable for downstream processing.

Output Format (strict JSON)

```
{
   "title": str,
   "brief_summary": str,
   "drugs": list[str]  or  [ ],
   "diseases": list[str],
   "inclusion_criterion": list[str],
   "exclusion_criterion": list[str]  or  [ ],
   "rationale": str
}
```

Figure 9: The Prompt for Concomitant-varying Trial Generation (Both Concomitant and Near-miss Diagnoses)

C.7 Prompt for Clinical Relevance and Eligibility Verification ($(r_{p,t},\ e_{p,t}) = \mathcal{M}_{\texttt{veri}}(x_p, t)$)

Task Overview

You are a senior clinical trial evaluation specialist with over 15 years of professional experience. Given a patient note and a clinical trial description, your task is to evaluate the relationship between the patient and the trial along two independent dimensions: **relevance** and **eligibility**.

Patient

```
{Patient}
```

Clinical Trial

```
{Trial}
```

Evaluation Dimensions

1. Relevance (Semantic / Clinical Alignment) Assess whether the trial's target condition(s) align with the patient's actual medical problems. Choose **exactly one** relevance level:

- **Strong match (3)**: The trial's primary target disease directly matches the patient's confirmed factor(s), or targets a clinically significant comorbidity or complication explicitly present.
- **Moderate match (2)**: The trial targets a medically related condition within the same disease family or organ system, and the patient represents a plausible and meaningful subpopulation.
- **Weak match (1)**: Overlap exists only at the level of general symptoms, risk factors, or broad categories, and the patient is not a clearly intended target population.
- **No match (0)**: The trial's target disease area is clinically irrelevant to the patient.

2. Eligibility (Rule-Based Fit)
Determine whether the patient would realistically qualify for the trial based on explicit inclusion and exclusion criteria. Choose **exactly one** eligibility level:

- **Eligible (3)**: The patient satisfies all major inclusion criteria and triggers no major exclusion criteria.
- **Likely eligible (2)**: The patient satisfies most inclusion criteria, with only minor or limited conflicts.
- **Possibly eligible (1)**: Some inclusion criteria cannot be verified due to missing information, or minor conflicts exist, but no definitive exclusion is triggered.
- **Not eligible (0)**: The patient clearly fails essential inclusion criteria, meets a major exclusion criterion, or has a diagnosis incompatible with the trial's target condition.

Output Format (strict JSON)

```
{
   "relevance_score": str(0|1|2|3),
   "relevance_reason": str,
   "eligibility_score": str(0|1|2|3),
   "eligibility_reason": str
}
```

Figure 10: The Prompt for Clinical Relevance and Eligibility Verification