EC-RAFT: Automated Generation of Clinical Trial Eligibility Criteria through Retrieval-Augmented Fine-Tuning

Nopporn Lekuthai[1,2] Nattawit Pewngam[3] Supitcha Sokrai[3] Titipat Achakulvisut[1]

[1] Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Bangkok, Thailand
[2] Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand
[3] Ravis Technology, Bangkok, Thailand
nopporn.lek@student.mahidol.edu, titipat.ach@mahidol.edu

Abstract

Eligibility criteria (EC) are critical components of clinical trial design, defining the parameters for participant inclusion and exclusion. However, designing EC remains a complex, expertise-intensive process. Traditional approaches to EC generation may fail to produce comprehensive, contextually appropriate criteria. To address these challenges, we introduce EC-RAFT, a method that utilizes Retrieval-Augmented Fine-Tuning (RAFT) to generate structured and cohesive EC directly from clinical trial titles and descriptions. EC-RAFT integrates contextual retrieval, synthesized intermediate reasoning, and fine-tuned language models to produce comprehensive EC sets. To enhance clinical alignment evaluation with referenced criteria, we also propose an LLM-guided evaluation pipeline. Our results demonstrate that our solution, which uses Llama-3.1-8B-Instruct as a base model, achieves a BERTScore of 86.23 and an EC-matched LLM-as-a-Judge score of 1.66 out of 3, outperforming zero-shot Llama-3.1 and Gemini-1.5 by 0.41 and 0.11 points, respectively. On top of that, EC-RAFT also outperforms other fine-tuned versions of Llama-3.1. EC-RAFT was trained in a low-cost setup and, therefore, can be used as a practical solution for EC generation while ensuring quality and relevance in clinical trial design. We release our code on GitHub at https://github.com/biodatlab/ec-raft/.

1 Introduction

Eligibility Criteria (EC) are essential components of clinical trial design, specifying the parameters for participant inclusion and exclusion (Su et al., 2023). These criteria ensure trials are scientifically valid, ethically sound, and capable of meeting their objectives. However, designing EC remains a labor-intensive and expertise-driven process (Su et al., 2023). Tools that can suggest or generate relevant EC have the potential to significantly facilitate researchers' work in trial design (Kim et al., 2024). Generating these criteria is inherently complex because consistency and clinical validity are needed throughout the criteria set. Despite advances in using large language models (LLMs) for summarization or specialized tasks in the biomedical domain, several barriers remain to creating a fully automated, contextually accurate system that can generate comprehensive sets of EC directly from trial descriptions. Recent developments in instruction fine-tuning for LLMs have shown promise in generating logical reasoning outputs through techniques like chain-of-thought prompting and rationale generation (Wei et al., 2022). Retrieval-augmented generation (RAG) (Lewis et al., 2021) has also emerged as an effective mechanism for grounding model outputs with external domain knowledge, thereby improving factual correctness (Ram et al., 2023). Retrieval-augmented fine-tuning (RAFT) (Zhang et al., 2024) extends RAG by incorporating instruction fine-tuning to improve both domain adaptation and retrieval robustness. These developments allow the development of an end-to-end system to generate a complete set of EC while preserving essential clinical context and domain relevance. To address these gaps, we propose EC-RAFT, a novel approach that leverages Retrieval-Augmented Fine-Tuning (RAFT) (Zhang et al., 2024) for automated EC generation. EC-RAFT aims to produce complete EC sets directly from trial titles and descriptions without requiring user-prompted EC categories or a recommendation system. Our key features include:

1. RAFT (Zhang et al., 2024) incorporates relevant external clinical trial information (existing trial details and eligibility criteria) and generates intermediate reasoning steps to fine-tune LLM.

2. Generating a complete set of eligibility criteria results in a fully structured set of inclusion and exclusion criteria. We demonstrate that

Findings of the Association for Computational Linguistics: ACL 2025, pages 9432–9444
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

synthesized intermediate reasoning steps produced by LLM, enhance the performance of the base models during fine-tuning for EC generation. Our results show that EC-RAFT exceeds zero-shot baseline approaches across multiple evaluation metrics, including semantic similarity and LLM-as-a-judge scoring.

Our training setup was also optimized for cost efficiency using the Parameter-Efficient Fine-Tuning technique (PEFT) (Xu et al., 2023; Hu et al., 2021). Specifically, training our best model required 380 GPU hours on NVIDIA A100 costing approximately 452.20 USD while achieving superior performance compared to the baseline.

2 Related Work

2.1 Eligibility Criteria Generation and Recommendation.

Over the past decade, various methods have been proposed to facilitate EC design. Trial2Vec (Wang and Sun, 2022) introduced trial-level representation using contrastive learning to recommend relevant clinical trials to researchers, providing a foundation for trial similarity assessment. Based on trial representation approaches, CReSE (Kim et al., 2024) applied contrastive learning and rephrasing strategies to recommend relevant EC for a given trial context, focusing on high semantic similarity. However, CReSE's recommendation-centric approach lacks mechanisms for generating complete EC sets. AutoTrial (Wang et al., 2023) generates EC at the trial level using large language models (LLMs), offering interpretability through structured outputs presented as chains of criteria. While described as "reasoning chains", these reflect sequences of related criteria rather than explicit, step-by-step rationales. The method incorporates instruction-based prompting and retrieval-augmented generation, enabling controllable and context-aware outputs. However, the reliance on structured instruction categories and list of ECs as their reasoning chain may limit flexibility in addressing nuanced or unconventional trial. Autocriteria (Datta et al., 2024) employs prompting on GPT-4 to extract fine-grained EC from clinical trial documents. To address remaining challenges, EC-RAFT integrates retrieval-augmented fine-tuning with synthesized chain-of-thought reasoning, offering a more adaptive and comprehensive solution for generating structured EC.

2.2 LoRA and Supervised Fine-Tuning (SFT).

Adapting LLMs to specialized tasks, such as EC generation, often requires fine-tuning on domain-specific datasets. Low-rank adaptation (LoRA) (XTuner Contributors, 2023; Hu et al., 2021) efficiently integrates domain knowledge into pre-trained models and has been successfully applied to biomedical tasks (Liao et al., 2024). Similarly, supervised fine-tuning (SFT) has demonstrated efficacy in automated medical report generation (Guo et al., 2024). However, while these methods show significant performance, they lack integrated retrieval mechanisms essential for producing comprehensive, domain-specific outputs such as complete EC sets. EC-RAFT overcomes this gap by incorporating retrieval-augmented fine-tuning, enabling it to leverage external knowledge effectively to generate detailed and structured eligibility criteria.

2.3 Retrieval-Augmented Fine-Tuning (RAFT).

RAFT (Zhang et al., 2024) has shown promise across various domains, including biomedical tasks, by simulating an "open-book" scenario in which models consult external documents during training and inference. Traditionally, RAFT methods involve providing the model with "golden" and "distractor" texts, helping it learn when to utilize external information effectively. However, standard RAFT approaches predominantly focus on short-form QA tasks and rarely extend to generating extensive outputs like structured EC sets. EC-RAFT expands RAFT's applicability by integrating domain-specific retrieval with synthesized chain-of-thought reasoning, thereby providing a novel method capable of generating comprehensive, trial-level eligibility criteria.

3 Methods

In this section, we introduce our approach, which leverages clinical trial data from ClinicalTrials.gov and integrates state-of-the-art techniques in embedding, retrieval, and fine-tuning to automate the generation of EC (Figure 1). We then describe the experiments designed to evaluate our system.

3.1 ClinicalTrials.gov Dataset

We collected 267,347 clinical trials from ClinicalTrials.gov, covering 2000 to 2024. To facilitate analysis, we split these trials into three datasets: 213,877 trials for training, 26,735 trials for valida-

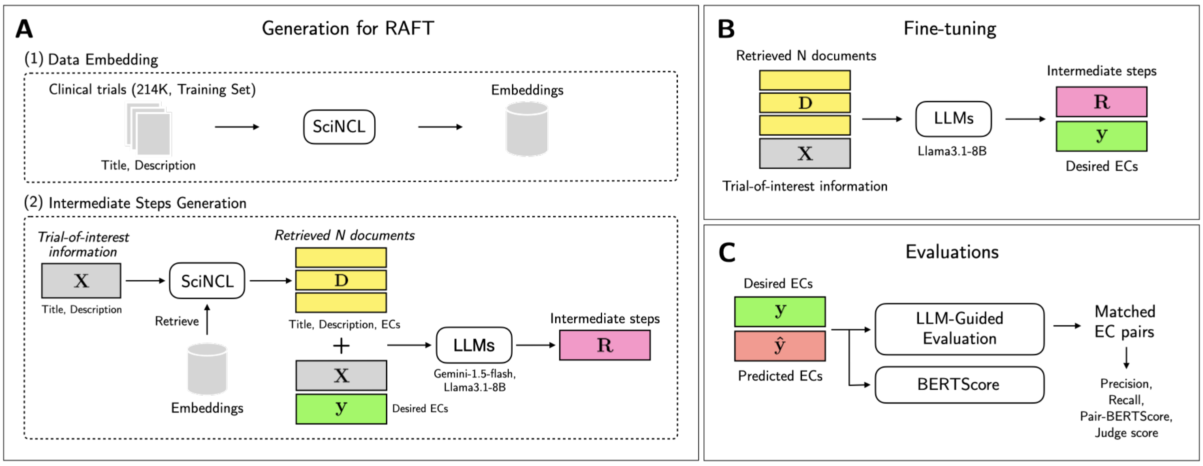

Figure 1: **Overview of the EC-RAFT pipeline. A.** (1) Data embedding using SciNCL (2) Retrieve relevant trials and their ECs ($\mathbf{D}$) for the trial of interest ($\mathbf{X}$), then combine with desired ECs ($\mathbf{y}$) to generate intermediate steps ($\mathbf{R}$). **B.** Fine-tune the model to generate a single response that includes both reasoning and final eligibility. **C.** Evaluate using two approaches: (1) **BERTScore** (Zhang et al., 2020) for semantic similarity, and (2) **LLM-Guided Evaluation** for clinical relevance of matched EC pairs.

tion, and 26,735 trials for testing (Table 1). The training, validation, and test set contains around 168.4k, 20.9k, 21.1k interventional and 45.4k, 5.8k, 5.6k observational trials respectively. The training data contain 1.25M interventional criteria with an average of 4.98 ± 5.11 inclusions and 7.46 ± 7.05 exclusions per trial and 137k observational criteria with an average of 3.02 ± 2.85 inclusions and 3.44 ± 3.68 exclusions per trial.

Our dataset consists of three primary sections: title, description, and ECs. The description section includes a brief summary, a detailed description, and intervention details, including the type, name, description, and alternative names. The EC section, which extracts from `eligibilityModule` within the `protocolSection`, contains key participant criteria, including both structured fields and free-text criteria. `eligibilityCriteria` within `eligibilityModule` section provides key eligibility details, including inclusion and exclusion criteria. While most trials specify age and gender requirements within the `eligibilityCriteria` section, some studies omit explicit references to these factors. Instead, these details are provided in dedicated fields within the same module: `sex` for gender information, `minimumAge` and `maximumAge` for age ranges, and `healthyVolunteers` for whether healthy volunteers are accepted. We extracted and processed these fields from both structured metadata and free-text EC to ensure that all EC are included.

3.2 Data Embedding and Retrieval

The first step involves obtaining comprehensive clinical trial data, including titles, descriptions, and eligibility criteria, from ClinicalTrials.gov (Figure 1A). We employ the SciNCL embedding model (Ostendorff et al., 2022) to embed clinical trials, which are subsequently retrieved to generate intermediate steps ($\mathbf{R}$). The rationale for selecting SciNCL is its ability to embed semantics in domain-specific text. After embedding, we retrieve relevant trials—specifically their titles, descriptions, and ECs ($\mathbf{D}$) using Euclidean distance. Importantly, only the training split was embedded. During testing and evaluation, we retrieved trials exclusively from the embedded training split. Our experiments vary the relevant trials (top-N) from N = 1 to 5 for generating the intermediate step ($\mathbf{R}$) (Section 5.4).

3.3 Intermediate Steps Generation

In EC-RAFT, the generation of intermediate reasoning steps ($\mathbf{R}$) plays an important role in creating a structured pathway for training models. This process begins by integrating the retrieved trial information ($\mathbf{D}$) which includes the title, description and ECs, the trial-of-interest information ($\mathbf{X}$), consisting of its title and description, and the desired eligibility criteria ($\mathbf{y}$) for the target study (Figure 1A). The $\mathbf{D}$ is retrieved from the vector database using $\mathbf{X}$'s title and description while filtering out $\mathbf{X}$ out of retrieved documents, with different top-N values applied based on the experimental configuration. The desired eligibility criteria ($\mathbf{y}$) serve

Statistic	Train (N = 213,877)		Validation (N = 26,735)		Test (N = 26,735)	
	Interventional	Observational	Interventional	Observational	Interventional	Observational
Number of Clinical Trials	168,429	45,448	20,928	5,807	21,129	5,606
Total Inclusion Criteria	838,948	137,234	103,910	17,531	103,982	16,990
Total Exclusion Criteria	1,256,242	156,298	154,896	20,470	156,212	19,000
Mean Inclusion Criteria per Trial (± SD)	4.98 ± 5.11	3.02 ± 2.85	4.97 ± 5.04	3.02 ± 2.77	4.92 ± 5.01	3.03 ± 2.99
Mean Exclusion Criteria per Trial (± SD)	7.46 ± 7.05	3.44 ± 3.68	7.40 ± 7.00	3.53 ± 3.64	7.39 ± 7.05	3.39 ± 3.61

Table 1: **Statistics of clinical trials and EC**. We calculate an average and a standard deviation of the number of EC of interventional and observational trials as these study types differ in their structure, particularly in the number of exclusion criteria.

as a *hint* that guides the LLM in breaking down each criterion, connecting them to evidence derived from retrieved studies ($\mathbf{D}$) and the study information ($\mathbf{X}$).

The primary objective is to generate intermediate reasoning steps ($\mathbf{R}$) that justify how each eligibility criterion ($\mathbf{y}$) is logically constructed and justified based on the retrieved trials ($\mathbf{D}$) and the target trial information ($\mathbf{X}$). The process can be written as:

$$\mathbf{D} + \mathbf{X} + [\text{Hint} : \mathbf{y}] \rightarrow \mathbf{R} \tag{1}$$

These intermediate steps will later be used in the fine-tuning steps formulated in (2). (see 3.4 for more details). These intermediate steps allow the model to learn how to derive eligibility criteria ($\mathbf{y}$) from trial information ($\mathbf{X}$) and retrieved studies information ($\mathbf{D}$).

Including retrieved trials as part of the input provides the LLM with domain-specific examples, offering insights into established clinical practices. These examples enable the model to identify patterns and infer appropriate criteria for the target study. However, discrepancies may arise when the desired EC conflict with information from the retrieved trials. For instance, a retrieved trial might exclude patients with mild hypertension, whereas the target study explicitly includes them. In such cases, the LLM is tasked with identifying and articulating these conflicts, justifying deviations from established norms.

This conflict-resolution mechanism aims to ensure that the generated eligibility criteria ($\hat{\mathbf{y}}$) are likely to be both contextually relevant and aligned with the specific goals of the target study, even when they may diverge from traditional practices. Our experiments explore the use of models including `Gemini-1.5-flash-002` (Gemini Team, 2024) and `Llama-3.1-8b-instruct` (Grattafiori et al., 2024) to synthesize intermediate steps ($\mathbf{R}$).

3.4 RAFT for Generating EC

RAFT in EC-RAFT enhances the model's ability to generate eligibility criteria ($\mathbf{y}$) by leveraging relevant context retrieved from clinical trial data ($\mathbf{D}$). Unlike traditional RAFT methods (Zhang et al., 2024) that classify documents as golden or distractors, EC-RAFT utilizes all retrieved trials holistically to account for varying levels of relevance. This ensures that the model is informed by diverse clinical contexts during fine-tuning. In this step, we utilized `Llama-3.1-8b-instruct` as a base model for supervised fine-tuning. We utilize Low-Rank Adaptation (LoRA) training techniques for cost efficiency. This fine-tuning process is structured as follows:

$$\mathbf{D} + \mathbf{X} \rightarrow \mathbf{R} + \mathbf{y} \tag{2}$$

This approach aligns the model's training process with real-world scenarios, allowing it to learn directly from domain-specific documents in an open-book setting (Zhang et al., 2024). By integrating reasoning steps ($\mathbf{R}$), the model is encouraged to generate both eligibility criteria ($\mathbf{y}$) and output logical intermediate steps generated in the section above.

3.5 Generation of Eligibility Criteria

During inference, the fine-tuned model inputs the target trial's title and description (Figure 1B). It retrieves relevant trials from the vector database and uses the combined information to generate a complete set of EC. The output includes how eligibility criteria are derived ($\hat{\mathbf{R}}$) and the whole set of predicted eligibility criteria ($\hat{\mathbf{y}}$). Similar to the fine-tuning process, we can write this as:

$$\mathbf{D} + \mathbf{X} \rightarrow \hat{\mathbf{R}} + \hat{\mathbf{y}} \tag{3}$$

We generate both the reasoning path and the predicted criteria. This allows the model to produce a reasoning process before predicting EC, which may improve results compared to direct inference (Wu et al., 2024).

To evaluate the effectiveness of our approach, we compare the performance of EC-RAFT with zero-shot inference from `Llama-3.1-8b-instruct` and `Gemini-1.5-flash`. We also vary the number of top-N during the generation of $\hat{\mathbf{R}}$ to evaluate its performance across different numbers of retrieved documents ($\mathbf{D}$).

4 Evaluation

Due to the challenging nature of semi-structured Eligibility Criteria, we employ three metrics to compare our predicted output ($\hat{y}$) with the ground truth (y) to measure: 1) **BERTScore** for overall semantic similarity, and 2) **LLM-Guided evaluation** which only evaluate the matched pair, **Pair-BERTScore**, identified by LLMs and utilize **LLM-as-a-Judge** to judge capability to assess clinical relevance for each matched pair.

4.1 BERTScore

We utilize BERTScore (Zhang et al., 2020) with the DistilBERT (uncased) (Sanh et al., 2020) model to assess the semantic similarity between the desired and predicted EC. BERTScore evaluates alignment based on token-level matches between the reference and predicted criteria, weighting these matches by their contextual embeddings to produce a similarity score. However, BERTScore may overestimate similarity due to the semi-structured nature of EC and may fail to distinguish logical inversions between inclusion and exclusion criteria.

4.2 LLM-Guided Evaluation

We propose an LLM-guided evaluation pipeline to assess how well-generated EC aligns with their corresponding reference criteria. This pipeline combines (1) **Pairing-and-scoring step** matching EC and calculating Pair-BERTScore (Section 4.2.1) and (2) **An additional match score** using an LLM-as-a-Judge (Section 4.2.2). Below, we provide a general overview of the pipeline, followed by the unique details of each metric.

1. **Initial Evaluation** We use `Gemini-1.5-flash-002` to identify the most semantically and clinically relevant predicted criterion for each reference criterion. The model matches each reference criterion with the most pertinent predicted criterion, regardless of order, ensuring that all potential matches are considered. This process captures nuanced relationships between reference and predicted EC by explicitly accounting for inclusion-exclusion inversions, clinical parameters, and eligibility thresholds. The evaluation is generated in free-text format, prioritizing matching accuracy and judgment without enforcing a structured response, which could hinder accuracy (Tam et al., 2024). The evaluation prompt is provided in Figure A.

2. **Structured Output** We use `watt-tool-8B`'s (watt-ai, 2023) structured response functionality to convert free-text evaluations into a JSON schema, ensuring consistency for accuracy calculations (Figure B). We utilized `watt-tool-8B` due to its state-of-the-art performance in tool-calling despite its size (Yan et al., 2024).

4.2.1 Pair-BERTScore

After getting the structured pairs of inclusion and exclusion, we calculate semantic similarity using BERTScore (Fig 2). This process enhances evaluation accuracy by removing any inflated scores that may arise from structural similarities. Note that Pair-BERTScore only accounts for the paired EC but not the excess generation of predicted criteria.

4.2.2 LLM-as-a-Judge

While Pair-BERTScore measures semantic similarity, it may not capture clinically significant distinctions between desired and predicted EC $(\mathbf{y}, \hat{\mathbf{y}})$. To address this, we introduce LLM-as-a-Judge, which evaluates the clinical and logical alignment between matched EC pairs. For each matched pair, `Gemini-1.5-flash` assigns a clinical relevance score, referred to as `match_score`, on a scale of 0-3, based on the degree of alignment (Figure A).

Adapted from the Evaluation Guideline for Assessing Clinical Relevance between an EC Pair (Su et al., 2023), the scoring methodology categorizes EC similarity as follows:

- Clinical relevance 3 → Clinically identical EC.
- Clinical relevance 2 → Strongly relevant due to factors like disease progression or epidemiology.
- Clinical relevance 1 → Loosely relevant due

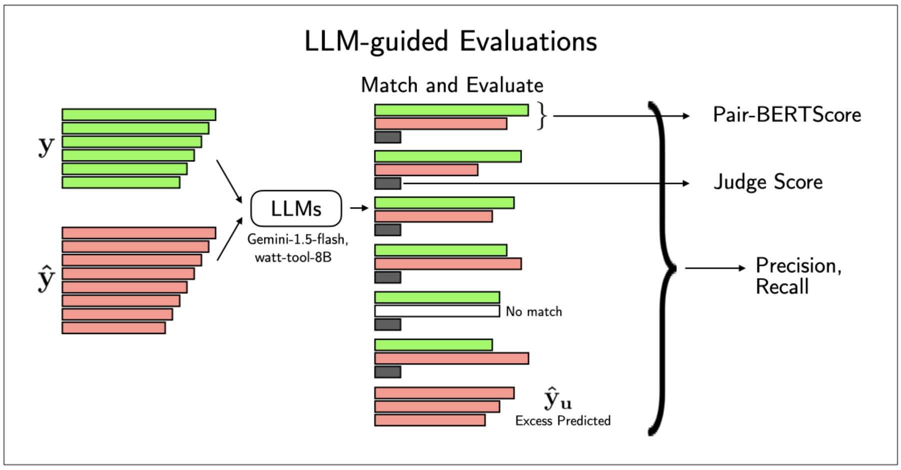

Figure 2: **LLM-guided Evaluation Metrics.** We align only generated EC, with corresponding reference EC and compute precision, recall, pair-BERTScore, and judge score. **Note** that in the actual pipeline, we also instruct the model to reason before evaluating each judge score (Fig. A, B)

to general treatment plans, disease progression, or epidemiological factors.

- Clinical relevance 0 → Irrelevant from a clinical perspective.

We calculate the mean `match_score` to measure how well the generated EC ($\hat{\mathbf{y}}$) align with the desired EC ($\mathbf{y}$).

4.2.3 Precision-Recall

Similar to Pair-BERTScore, the judge's score does not account for the excess EC generated. Thus, we also computed precision and recall to quantitatively measure the agreement between predicted and reference EC as follows

$$Precision = \frac{N_M}{N_P}, Recall = \frac{N_M}{N_R}, \quad (4)$$

where N_M represents the number of matched reference criteria with a positive match score (`match_score` > 0), N_R denotes the total number of reference criteria, and N_P is the total number of predicted criteria after de-duplication and filtering. De-duplication removes the exact predicted EC or a part of the same EC.

5 Results

5.1 Comparison with zero-shot baselines

Using the ClinicalTrials.gov test split, we compare EC-RAFT performance against two zero-shot baselines: `Llama-3.1-8B-Instruct` and `Gemini-1.5-flash`. As shown in Table 2, EC-RAFT achieves a BERTScore of 86.35, 4.93 higher than base model `Llama-3.1-8B-Instruct` and 4.17 higher than `Gemini-1.5-flash` which is a larger model (Table 2). This indicates improved overall semantic similarity between the generated and reference eligibility criteria. Regarding clinical relevance, EC-RAFT with Gemini's $\mathbf{R}$ obtains the highest precision and mean judge score, along with a superior mean Pair-BERTScore. This means that EC-RAFT can generate precise EC to the referenced EC. Although `Gemini-1.5-flash` registers a slightly higher recall, this advantage comes at the expense of precision—likely due to its tendency to generate excess criteria. On top of that, our model was self-improved by using base model `Llama-3.1-8B-Instruct` to generate $\mathbf{R}$ that could match the performance of `Gemini-1.5-flash` in some areas. Our results underscore the effectiveness of incorporating retrieval-augmented fine-tuning with intermediate reasoning steps, as it enables the model to generate eligibility criteria that are both semantically and clinically relevant.

5.2 Comparison with finetune baselines

To further benchmark EC-RAFT, we fine-tuned two strong LLMs—`Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) and `Meditron-7B` (Chen et al., 2023), a domain-specific variant fine-tuned on biomedical corpora (DSF), serves as a particularly strong baseline in medical NLP tasks—using a conventional approach where where $\mathbf{X}$ is the concatenation of trial title and description and $\mathbf{y}$ is

Model	BERTScore ↑	LLM-guided Evaluations			
		Precision ↑	Recall ↑	Mean Judge Score ↑	Mean Pair-BERTScore ↑
LLaMA 3.1-8B-Instruct (Zero-shot)	81.42	77.16	67.63	1.3097	51.95
LLaMA 3.1-8B-Instruct (Fine-tuned)	83.01	61.42	70.16	1.5114	61.49
Meditron-7b (Biomedical DSF + Fine-tuned)	82.96	60.72	71.23	1.5748	62.59
Gemini-1.5-flash-002 (Zero-shot)	82.18	72.47	**78.34**	1.6004	63.66
EC-RAFT ($\mathbf{R}$ from LLaMA 3.1-8B)	**86.35**	72.55	61.20	1.5932	66.92
EC-RAFT ($\mathbf{R}$ from Gemini-1.5-flash)	86.23	**78.84**	75.89	**1.7150**	**67.76**

Table 2: Comparison of zero-shot and fine-tuned baselines alongside EC-RAFT.

the eligibility criteria.

$$\mathbf{X} \rightarrow \mathbf{y} \tag{5}$$

Despite these models being explicitly fine-tuned for the task, EC-RAFT consistently outperforms them across key evaluation metrics, including BERTScore and LLM-guided clinical relevance assessments (Table 2). These results affirm the effectiveness of retrieval-augmented fine-tuning not only over zero-shot baselines but also over task-specific supervised learning, further validating EC-RAFT's performance.

While we acknowledge AutoTrial as a strong baseline for this task, we encountered challenges in replicating their work, as the paper emphasis on generation at the criteria level.

5.3 Effect of Larger model Intermediate steps

Here, we want to see if reasoning steps can affect the fine-tuned performance of EC-RAFT. We compare two variations differing in the model used to generate intermediate reasoning steps ($\mathbf{R}$): `Llama-3.1-8B-Instruct` and `Gemini-1.5-flash`. As shown in Table 2, both approaches significantly improve BERTScore over the baselines, with EC-RAFT using `Llama-3.1-8B-Instruct` achieving a slightly higher BERTScore than the Gemini-based variant. However, EC-RAFT with `Gemini-1.5-flash` exhibits superior overall performance across LLM-guided evaluations, achieving the highest precision, recall, mean Pair-BERTScore, and mean judge score, suggesting that its generated criteria are more clinically aligned. These results highlight the impact of selecting a strong LLM for generating intermediate reasoning steps, reinforcing that larger models like `Gemini-1.5-flash` can improve the accuracy and clinical relevance of EC generation.

5.4 Effect of LoRA hyper-parameters

LoRA (Low-Rank Adaptation) enables efficient fine-tuning by introducing trainable low-rank updates. We evaluate the impact of Rank (r) and Scaling Factor (α) on Eligibility Criteria generation using BERTScore and LLM-guided evaluations. Results in Table 3 show slightly better in BERTScore, precision, and judge score when increasing r from 64 to 128 and α from 16 to 64, while recall remains stable, indicating that increasing LoRA's rank does not significantly enhance EC generation $\hat{\mathbf{y}}$.

5.5 Effect of top-N retrieval

We evaluate EC-RAFT with different top-N settings to examine the impact of retrieved documents on eligibility criteria generation. We generate $\mathbf{R}$ using `Llama-3.1-8B-Instruct` by varying N retrieved documents. Table 4 shows that increasing N initially improves performance. BERTScore peaks at top-N of 4 before stabilizing, and precision follows a similar trend, suggesting excess documents may introduce noise. Recall remains stable with minor fluctuations, while Mean Pair-BERTScore and Mean Judge Score show slight variations. Overall, retrieving around four relevant documents provides modest benefits, but the overall impact remains limited.

5.6 LLM-as-a-Judge agreement with human

In healthcare, the application of artificial intelligence systems necessitates careful assessment of their quality and applicability before they are used in practice (de Hond et al., 2022). To validate the reliability of our LLM-guided evaluation, we conducted a human evaluation with a licensed physician. We sampled 20 trials from the test set, comprising 302 matched EC pairs. The physician reviewed each pair, rematched when necessary, and independently assigned `match_score` using the same 0–3 rubric applied by the LLM-as-a-Judge. Notably, no EC pairs required rematching during manual review.

Table 5 demonstrate strong alignment between the LLM-as-a-Judge and expert human judgment,

Model	BERTScore ↑	LLM-guided Evaluations			
		Precision ↑	Recall ↑	Mean Judge Score ↑	Mean Pair-BERTScore ↑
EC-RAFT ($r = 64$, $\alpha = 16$)	86.17	70.69	67.76	1.6039	61.73
EC-RAFT ($r = 128$, $\alpha = 64$)	86.24	71.08	67.70	1.6046	61.76

Table 3: Comparison between different LoRA configuration (Rank r and Alpha α) with top-N = 2

Top-N	BERTScore ↑	LLM-guided Evaluations			
		Precision ↑	Recall ↑	Mean Judge Score ↑	Mean Pair-BERTScore ↑
1	86.17	70.73	67.70	1.6003	61.69
2	86.17	70.67	67.76	1.6039	61.76
3	86.31	72.05	66.82	1.5897	61.10
4	86.35	72.55	66.92	1.5932	61.21
5	86.35	72.47	67.12	1.5981	61.37

Table 4: Comparison of different top-N configurations for EC-RAFT

Metric	Value
Krippendorff's alpha	89.26
Spearman's rank correlation	90.26% (p = 0.00)

Table 5: Agreement between physician and LLM-as-a-Judge on clinical relevance scoring.

with the Krippendorff's alpha of 89.26% affirming the clinical reliability of our automatic evaluation pipeline.

5.7 Qualitative and Error Analysis

We sample a clinical trial on stroke and generate EC using EC-RAFT and Gemini-1.5-flash (Table 6). We found that EC from EC-RAFT are closely matches the reference in age and thrombectomy eligibility but omits intracranial vertebral artery involvement. Meanwhile, Gemini-1.5-flash are more restrictive, requiring prior endovascular therapy and a strict 90-day follow-up. It also excludes patients with a history of stroke/TIA and severe co-morbidities, further reducing eligibility.

Overall, EC-RAFT tracks the reference more closely, while Gemini-1.5-Flash generates a more lengthy EC, having higher recall but lower precision. This highlights the trade-off between precision and recall in automated EC generation.

6 Conclusion

In this work, we introduced EC-RAFT, a framework that leverages retrieval-augmented fine-tuning and synthesized intermediate reasoning to automate the generation of clinical trial eligibility criteria. EC-RAFT generates structured, robust, and clinically relevant eligibility criteria directly from trial descriptions. Our experiments on a large-scale ClinicalTrials.gov dataset demonstrate that EC-RAFT outperforms zero-shot baselines without retrieval, despite being much smaller in model size, and outperforms fine-tuned versions of LLama, achieving higher BERTScores and clinical alignment, as evidenced by BERTScore and LLM-guided evaluations. Notably, the LLM-guided evaluations show high agreement with a licensed physician, further validating the clinical relevance of the generated criteria. While challenges remain, EC-RAFT represents a significant step towards automating the complex process of clinical trial design. Future work will refine the intermediate steps generation process and scale up model size and compute for better performance.

Limitations

While EC-RAFT demonstrates promising results in automated EC generation, several limitations should be acknowledged. First, our approach relies on LLMs, which can produce plausible but inaccurate or inconsistent criteria that require human expert validation. Second, training data comes primarily from public clinical trial registries, which may not fully represent the diversity of trial designs or specialized medical domains. Third, our evaluation metrics (BERTScore and LLM-as-a-Judge) provide approximations of generated EC but may not fully capture clinical relevance or practical applicability. Specifically, the latter may suffer from inconsistent

Multicentre ObseRvational Study in Posterior mecHanical thrombEctomy: conscioUs Sedation vs General aneSthesia: MORPHEUS STROKE REGISTRY (NCT05082896)

Reference Criteria	EC-RAFT	Gemini-1.5-flash
Inclusion Criteria: - Subject aged $\geq$ 18 years - Evidence of posterior arterial occlusion - Patients eligible for mechanical thrombectomy **Exclusion Criteria:** - Patient's refusal to participate - Arterial recanalization prior to endovascular thrombectomy **Sex:** ALL **Ages:** - Minimum Age: 18 Years - Age Group (Child: birth-17, Adult: 18-64, Older Adult: 65+): OLDER_ADULT, ADULT **Accepts Healthy Volunteers:** No	**Inclusion Criteria:** - Age $\geq$ 18 years - Acute ischemic stroke with occlusion of the basilar artery or P1 segment of the posterior cerebral artery - Treatment with endovascular therapy - Treatment initiation within 24 hours after stroke onset - Informed consent from patient or legal representative **Exclusion Criteria:** - Pre-existing conditions with significant impact on functional outcome (pre-stroke mRS > 1) - Intracranial hemorrhage - Pregnancy - Participation in other clinical trials **Sex:** ALL **Ages:** - Minimum Age: 18 Years - Age Group: OLDER_ADULT, ADULT **Accepts Healthy Volunteers:** No	**Inclusion Criteria:** - Confirmed diagnosis of ischemic stroke due to posterior circulation large vessel occlusion (PC LVO) demonstrated by imaging (CTA or MRA). - Underwent endovascular therapy (EVT) for the PC LVO. - Received either conscious sedation OR general anesthesia during EVT. - Able to complete at least 90 days of follow-up. - Provided informed consent to participate in the study. **Exclusion Criteria:** - Contraindications to endovascular therapy (EVT). - Ischemic stroke not due to PC LVO (e.g., anterior circulation stroke). - Hemorrhagic stroke. - Incomplete follow-up data at 90 days post-EVT. - History of previous stroke or Transient Ischemic Attack (TIA). - Severe co-morbidities that may confound outcome assessment. **Sex:** ALL **Ages:** - Minimum Age: 18 Years - Maximum Age: 90 Years - Age Group: ADULT, OLDER_ADULT **Accepts Healthy Volunteers:** No

Table 6: Comparison of NCT05082896's Eligibility Criteria: Reference, EC-RAFT, and Gemini-1.5-flash

responses and bias. The performance may vary between different medical specialties and trial types, particularly for rare diseases or novel therapeutic approaches where training data are limited. Future work should address these limitations through expanded training datasets, domain-specific fine-tuning, and more rigorous clinical validation processes.

Acknowledgments

The authors acknowledge the Faculty of Medicine Ramathibodi Hospital and the NSTDA Supercomputer Center (ThaiSC) project #pv814001 for providing funding and computing resources for this work. The authors also thank the team at RAVIS Technology for providing insightful feedback during our development.

References

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Surabhi Datta, Kyeryoung Lee, Hunki Paek, Frank J Manion, Nneka Ofoegbu, Jingcheng Du, Ying Li, Liang-Chin Huang, Jingqi Wang, Bin Lin, et al. 2024. Autocriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *Journal of the American Medical Informatics Association*, 31(2):375–385.

Anouk A.H. de Hond, Amber M. Leeuwenberg, Lotty Hooft, et al. 2022. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Medicine*, 5(1):2.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Li Guo, Anas M. Tahir, Dong Zhang, Z. J. Wang, and Rabab Ward. 2024. Automatic medical report generation: Methods and applications. *ArXiv*, abs/2408.13988.

J. E. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Siun Kim, Jung-Hyun Won, David Lee, Renqian Luo, Lijun Wu, Tao Qin, and Howard Lee. 2024. CReSE: Benchmark data and automatic evaluation framework for recommending eligibility criteria from clinical trial information. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2243–2273, St. Julian's, Malta. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Yusheng Liao, Shuyang Jiang, Yu Wang, and Yanfeng Wang. 2024. Ming-moe: Enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts. *ArXiv*, abs/2404.09027.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. *Preprint*, arXiv:2202.06671.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, A. Shashua, Kevin Leyton-Brown, and Y. Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Qianmin Su, Gaoyi Cheng, and Jihan Huang. 2023. A review of research on eligibility criteria for clinical trials. *Clinical and experimental medicine*, 23(6):1867–1879.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *Preprint*, arXiv:2408.02442.

Zifeng Wang and Jimeng Sun. 2022. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6377–6390.

Zifeng Wang, Cao Xiao, and Jimeng Sun. 2023. Autotrial: Prompting language models for clinical trial design. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12461–12472.

watt-ai. 2023. watt-tool-8b. `https://huggingface.co/watt-ai/watt-tool-8B`. Accessed: 14 February 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, E. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Tianhao Wu, Janice Lan, Weizhe Yuan, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Thinking llms: General instruction following with thought generation. *Preprint*, arXiv:2410.10630.

XTuner Contributors. 2023. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`.

Lingling Xu, Haoran Xie, S. J. Qin, Xiaohui Tao, and F. Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *ArXiv*, abs/2312.12148.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Berkeley function calling leaderboard.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *Preprint*, arXiv:2403.10131.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Appendix

This appendix provides documentation of the prompts used in EC-RAFT for EC generation. The following sections detail the exact prompts, implementation notes, and practices developed during our research.

A.1 LLM-guided Evaluation Prompt

To ensure a standardized and clinically grounded evaluation, we adapt the scoring methodology from the Evaluation Guideline for Assessing Clinical Relevance between an EC Pair (Su et al., 2023).

The adapted framework categorizes EC similarity into four levels:

- Clinical relevance 3 → Clinically identical ECs.
- Clinical relevance 2 → Strongly relevant due to factors like disease progression or epidemiology.
- Clinical relevance 1 → Loosely relevant due to general treatment plans, disease progression, or epidemiological factors.
- Clinical relevance 0 → Irrelevant from a clinical perspective.

The actual prompt can be found in figure A. The matched EC pairs and their scores can be found in figure B.

A.2 LLM-guided Evaluation JSON Schema

After the initial evaluation, we utilize `watt-tool-8B` to convert the free-text evaluation into a structured JSON format for quantitative analysis in Section 4.2. Since each reference criterion can match multiple predicted criteria, the predicted values are stored as a list of strings to accommodate the one-to-many relationship.

A.3 Implementation Details & Computational Cost

Our default LoRA configuration includes a Rank of 64, α of 16, and dropout of 0.1, except in section 5.4. We train on four NVIDIA A100 GPUs, requiring 192 to 470 GPU-hours per model, depending on the top-N value, totaling around 2,200 GPU-hours across this paper. Our best-performing model is trained in 380 hours, costing approximately 452.20 USD at a market rate of 1.19 USD per GPU-hour.

A.4 Therapeutic Area Breakdown

We provide additional results broken down by therapeutic areas below. EC-RAFT consistently outperforms base LLaMA and Gemini baselines, confirming robust and generalizable performance improvements across diverse domains (Table 7). Such generalizability is critical in clinical AI applications, where performance must remain reliable across settings and populations (de Hond et al., 2022). The therapeutic area classification is based on the CReSE paper (Kim et al., 2024). However, since CReSE did not provide the information on how they classify the therapeutic area (and neither did

LLM-guided Evaluation Prompt

Please evaluate the clinical relevance of the following two eligibility criteria on a 4-point scale (0–3). Below is an example of a clinical situation by `match_score` and the corresponding EC pair.
Clinical relevance 3: The two eligibility criteria are essentially identical clinically.
Examples:

- EC1: "[exclusion] serum albumin is 2.4 g/dL or less"
  EC2: "[inclusion] serum albumin is 2.4 g/dL or more"
- EC1: "Minimum Age : 18 Years"
  EC2: "Minimum Age : 18 Years"

Clinical relevance 2: The two eligibility criteria have strong relevance due to factors such as disease progression or epidemiology.
Example: ***...omitted for brevity...***
...omitted for brevity...
Evaluation Process
For each reference criterion, compare it to the relevant predicted criteria. If no relevant predicted criterion exists, state this explicitly. The evaluation process is as follows:

1. Recite the reference exact criterion and state explicitly if it is from [inclusion] or [exclusion].
2. Search the predicted criteria list to identify the relevant matches, regardless of order (comma-separated), and explicitly state which part of the predicted criteria each match comes from ([inclusion], [exclusion], [age], [sex], [accepts healthy volunteers]).
3. Recite the reference **Sex**, **Ages**, and **Accepts Healthy Volunteers** one at a time and compare them with the relevant predicted values.
4. Provide a reason explaining how the criteria match or differ.
5. Assign a match score (0–3) based on the clinical relevance of the predicted criterion to the reference criterion.
6. If no predicted criterion matches the reference, state that explicitly and assign a score of 0.

At the end of the evaluation, please provide:

- **Unmatched Predicted Criteria:**
  - **Unmatched Predicted Inclusion Criteria:** List all predicted inclusion criteria that were not matched to any reference criteria (relevance score = 0). No explanation is needed—just list them (comma-separated).
  - **Unmatched Predicted Exclusion Criteria:** List all predicted exclusion criteria ***...Same as before, omitted for brevity...***

Figure A: **LLM-guided Evaluation Prompt**

Therapeutic Area	**Mean Judge Score**			**Pair-BERTScore**		
	LLaMA-8b-Instruct	**Gemini-1.5-Flash**	**EC-RAFT**	**LLaMA-8b-Instruct**	**Gemini-1.5-Flash**	**EC-RAFT**
Oncology	1.1792	1.5286	**1.6267**	46.60	60.10	**62.36**
Neurology	1.3446	1.6231	**1.6894**	53.53	64.23	**65.11**
Metabolic	1.3021	1.5573	**1.7124**	51.87	62.77	**66.44**
Cardiology	1.3539	1.6250	**1.6698**	52.81	**64.29**	63.97
Rheumatology	1.3045	1.5850	**1.6681**	52.90	63.90	**64.55**
Infectious	1.3025	1.5980	**1.7936**	50.68	62.87	**67.54**
Hematology	1.2949	1.5800	**1.6922**	51.28	63.34	**64.40**
Immunology	1.2214	1.5593	**1.8144**	49.62	62.97	**69.70**
Dermatology	1.2217	1.5450	**1.8008**	49.59	62.83	**63.13**
Nephrology	1.3471	1.6496	**1.6851**	52.24	63.98	**68.17**
Pulmonology	1.3340	1.5728	**1.7203**	52.26	62.40	**66.10**
Gastroenterology	1.3602	1.6635	**1.7159**	53.69	**65.69**	65.45
Others	1.3694	1.6462	**1.7600**	54.69	65.91	**67.94**

Table 7: **Performance by therapeutic area.** EC-RAFT consistently outperforms both baselines across diverse domains.

the ClinicalTrials.gov include this information), we used `Gemini-1.5-flash` to classify trial information into the appropriate category.

LLM-guided Evaluation JSON Schema

```
{
  ”inclusion_criteria”: [
    {
      ”reference”: ”criteria”,
      ”predicted”: [”match”],
      ”reason”: ”explanation”,
      ”match_score”: 3
    }
  ],
  ”exclusion_criteria”: [
    {
      ”reference”: ”criteria”,
      ”predicted”: [”match”],
      ”reason”: ”explanation”,
      ”match_score”: 2
    }
  ],
  ”sex”: {
    ”reference”: ”value”,
    ”predicted”: [””],
    ”reason”: ”explanation”,
    ”match_score”: 0
  },
  ”age”: {
    ”reference”: ”value”,
    ”predicted”: [”match”],
    ”reason”: ”explanation”,
    ”match_score”: 2
  },
  ”accept_healthy_volunteer”: {
    ”reference”: ”value”,
    ”predicted”: [”match”],
    ”reason”: ”explanation”,
    ”match_score”: 1
  },
  ”unmatched_predicted_criteria”: {
    ”unmatched_predicted_inclusion
    _criteria”: [”unmatched”],
    ”unmatched_predicted_exclusion
    _criteria”: [”unmatched”]
  }
}
```

Figure B: **JSON Schema parsed from free-text judge response:**