MIMIC-CTRL-2026: A Realistic Clinical Trial Dataset to Benchmark LLM performance

Noshin Nawal[1] **Robert Margaryan**[2] **Moustafa Abdalla**[3,4]

Justin A. Ezekowitz[2] **Mohamed Abdalla**[2]

[1]Department of Computing Science, University of Alberta
[2]Department of Medicine, University of Alberta
[3]Department of Surgery, Massachusetts General Hospital
[4]Department of Global Health and Social Medicine, Harvard Medical School
mabdall2@ualberta.ca

Abstract

Patient recruitment remains a major barrier in clinical trials, as identifying eligible participants often requires manual screening of complex health records, contributing significantly to trial delays and increased costs. While Large Language Models (LLMs) offer potential to automate this process, existing public datasets are overly simplistic in their patient representation: they pair full criterion lists with single-paragraph synthetic summaries, failing to capture the longitudinal complexity of real clinical records. We introduce MIMIC-CTRL-2026, a dataset of 800 manually labeled patient-criterion pairs grounded in real-world longitudinal EHR data from MIMIC-IV, annotated by clinically trained experts. Each instance pairs one eligibility criterion with a full patient record spanning multiple discharge notes, radiology reports, lab values, and medication histories. While simple prompting achieves high performance ($\kappa = 0.84$) on existing benchmarks, it collapses ($\kappa = 0.06$) on our dataset, revealing a substantial benchmark-difficulty gap between simplified and real-world clinical evaluation. Lastly, we present a framework combining structured pre-processing with a clinician-validated expanded prompt that substantially outperforms naive baselines without any model fine-tuning, establishing a realistic benchmark for automated clinical trial screening on real EHR data.

Our dataset and code are available at: *https://doi.org/10.5281/zenodo.22103011*

> *A 51-year-old woman is seen in clinic for advice on osteoporosis. She has past medical history of significant hypertension and diet-controlled diabetes mellitus. She currently smokes 1 pack of cigarettes per day. She was documented by previous LH and FSH levels to be in menopause within the last year. She is concerned about breaking her hip as she gets older and is seeking advice on osteoporosis prevention.*

Figure 1: An example patient case (from the SIGIR dataset) which is used to match or exclude patients based on trials' inclusion and exclusion criteria.

1 Introduction

Clinical trials are essential to ensuring safe and effective healthcare innovations (Friedman et al., 2015). Patient recruitment remains one of the most resource-intensive and time-consuming steps in clinical trials (Coorevits et al., 2013; Cowie et al., 2017). Identifying eligible participants often requires manual screening of complex health records, a challenge that contributes significantly to trial delays and costs (Huang et al., 2018). This cost is not equitably distributed: data extraction is more complex and costly when screening older patients (Sedrak et al., 2021), those with intellectual and developmental disabilities (higher co-morbidity burden) (McDonald et al., 2022) and children (higher rare disease burden) (US Food and Drug Administration, 2019).

Recent advances in large language models (LLMs) offer transformative potential to streamline clinical trial workflows, including automated interpretation of eligibility criteria, natural language search across heterogeneous data sources, and adaptive support tools for trial teams. Previous studies have demonstrated that AI-assisted patient matching can improve the process of enrollment (Unlu et al., 2025).

Unfortunately for researchers, the sensitive nature of the data required to develop and evaluate these methods creates hurdles for reproducibility. Much of the past research is conducted on private data which makes direct comparisons practically impossible. To address this, the research community has generated some public datasets for patient trial matching. Unfortunately, the currently publicly available datasets are overly simplistic in

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 27428–27452
October 24-29, 2026 ©2026 Association for Computational Linguistics

their representation of patients. For example, a shared task from SIGIR-2016 (Koopman and Zuccon, 2016) represents patients as a single paragraph as illustrated in Figure 1. This overly simplifies the task of patient matching as real-world textual medical records for individual patients span multiple documents and formats (e.g., admissions notes, care notes, discharge notes, lab values, etc.).

The simplicity of these datasets results in an over-estimation in the performance of LLM-based methods for criterion-level eligibility assessment. To address these limitations, the contributions of our work are as follows:

- We introduce **MIMIC-CTRL-2026**, a manually annotated dataset of 800 patient–criterion pairs for criterion-level eligibility assessment using real-world longitudinal EHR data. Unlike prior datasets where patients are represented with synthetic summaries, for our dataset, each instance pairs *one* eligibility criterion with a full patient record spanning multiple admission or discharge notes, radiology reports, lab values, and medication histories.

- We benchmark models on both existing simplified (single-paragraph) and realistic longitudinal patient data, quantifying the performance gap and demonstrating that simplified datasets yield inflated results even under naive prompting strategies.

- We perform qualitative evaluation of model errors, exposing systematic failure modes on realistic clinical text and highlighting core limitations of existing approaches.

- We present an end-to-end framework for automated eligibility matching. The pipeline that combines structured pre-processing, targeted section retrieval via structural boundary extraction, and LLM-based criterion evaluation.

2 Related Work

Below we discuss: a) existing public patient-clinical trial matching datasets, and b) existing work using LLMs for patient-trial matching.

2.1 Public Patient-Clinical Trial Matching Datasets

NLP research for patient-trial matching uses a variety of public and private datasets. While private datasets offer high ecological validity, they hinder reproducibility because they contain personally identifiable information (PII) which prevent them from being shared with the research community. Thus many studies leverage public datasets which we discuss below:

2.1.1 '*JV Health Search*' (SIGIR)

'*JV Health Search*' (Koopman and Zuccon, 2016) (henceforth: SIGIR) is a dataset developed for the express purpose of evaluating information retrieval models' performance at the task of patient matching. In this set-up patients serve as the query to be matched with relevant trials. Patients (n=60) are represented using a description (average of 78 words) and a short summary (an average of 22 words) adopted from previous work (Simpson et al., 2014). Trials are represented using the inclusion and exclusion criteria (conditions that determine whether a person is eligible to participate) from real-world clinical trials listed at *ClinicalTrials.gov*. A total of approximately 4000 manually annotated patient-trial pairs are available from this dataset.

2.1.2 Text REtrieval Conference (TREC) (2021–2023)

These three datasets were published as part of the Clinical Trials Track of the Text REtrieval Conference (TREC) in 2021, 2022, and 2023.

TREC-2021 and TREC-2022 use the same underlying corpus (Roberts et al., 2022). With the task framed again as information retrieval, patients are topics and are represented using a short paragraph similar to what would be observed in an admission note. A few patient profiles have some selected information formatted as a list (often of lab values) appended to the prose. The 2021 iteration of the task had 75 patient profiles which was reduced to 50 in the 2022 iteration. Like the SIGIR dataset, inclusion and exclusion criteria from *ClinicalTrials.gov* were used. Annotators ranked patient-trial pairings (n=35,832) using one of three labels: Not Relevant, Excluded, or Eligible.

TREC-2023, while very similar to the previous two iterations, instead of patients being represented using synthetic patient cases written in full prose, patients are instead represented using a simulated questionnaire that would match what the physician of the patient would fill out to identify eligible trials (TREC Biomedical Tracks Organizers, 2023).

2.1.3 *n2c2*

Unlike the previously described datasets, the n2c2 2018 shared task: *Cohort selection for clinical trials* (Stubbs et al., 2019) uses real-patient data. Each patient (n=288) is represented by 2-5 notes (with an average of 2711 tokens/patient). Each patient was manually annoted to see if they met a pre-selected set of 13 criteria extracted from real-studies with minor modifications to maximize the utility of the patient notes.

2.1.4 Limitations of Public Datasets

While these works have made substantial contributions to the field, they share several limitations that restrict their applicability to real-world clinical trial–patient matching. First, the SIGIR, TREC 2021, and 2022 datasets rely on synthetic paragraphs, substantially oversimplify the task and fail to reflect the complexity of real clinical narratives. Although the TREC 2023 track represents a realistic patient checklist, requiring the filling of such a checklist does not address the main bottleneck faced by clinicians which needs us to deal with the natural, longitudinal clinical text. Second, the n2c2 dataset, while more realistic in that it is derived from non-synthetic clinical data and remains limited in scope, as it: a) includes only a small number of discharge summaries per patient (typically two to five) , and b) excludes other clinically relevant modalities, such as imaging reports, additional note types, lab tables, and medications. More importantly, the n2c2 dataset is no longer publicly available, further limiting its utility for reproducible and extensible research.

2.2 Automated Approaches to Trial Matching

Researchers from the information retrieval subfield frame the task as an information retrieval task where the patient is a query and is meant to be matched with all eligible trials (Roberts et al., 2022; Simpson et al., 2014; TREC Biomedical Tracks Organizers, 2023; Rybinski et al., 2024). In such framing, models are evaluated using metrics such as average precision@10, recall@10 and other retrieval metrics (Truong et al., 2022; Lahiri et al., 2024).

Researchers in NLP often frame the task as a classification task where, given a patient's profile and the inclusion and exclusion criteria for a specific trial, the model would classify whether the patient is eligible, ineligible, or not relevant to the particular trial. Such approaches tend to report traditional classification metrics (Tai and Tannier, 2025).

Some researchers combine both retrieval and classification in a multi-stage application (where retrieval is conducted to gather relevant trials followed by matching and sometimes a ranking as well) (Nievas et al., 2024; Jin et al., 2024).

For our work, we will focus on studies which have a classification component and highlight some particular relevant research. Jin et al. (2024), propose *TrialGPT*, a three-stage zero-shot framework built on OpenAI's GPT-4 and GPT-3.5 models to match patients to clinical trials. The system first generates trial-relevant keywords to enable large-scale retrieval, efficiently filtering tens of thousands of active trials (e.g., over 23,000 U.S. trials). In the second stage, the model evaluates patient eligibility at the individual inclusion and exclusion criterion level, producing natural-language explanations for its decisions. Finally, these criterion-level predictions are aggregated into trial-level eligibility scores. TrialGPT achieves near-expert performance, reporting an accuracy of 87.3% while reducing screening time by 42.6%. Despite these promising results, TrialGPT relies on synthetic patient summaries that do not reflect the complexity of real-world clinical records.

Wornow et al. (2024) present a zero-shot LLM-based clinical trial screening system that achieves state-of-the-art (SOTA) performance. Their approach combines strategic prompting with a two-stage retrieval pipeline that first filters relevant clinical notes before applying eligibility assessment. The system, evaluated using the n2c2 dataset, reports an 81% Macro-F1 score and reduces screening costs by approximately 95%, addressing key practical challenges related to data scarcity, scalability, and deployment feasibility.

Chowdhury et al. (2024) proposed Siamese-PTM, a Siamese Neural Network using Llama 2 embeddings to model the semantic correlation between patient EHR and trial criteria. This architecture achieved a significant 40% performance gain over a rule-based baseline. Guluzade et al. (2025) (ELMTEX) implemented fine-tuning on smaller open-source LLMs to optimize the extraction of structured clinical entities (diagnoses, treatments, etc.) from reports. Their primary limitation is that their system is designed for general clinical entity extraction using a purpose-built dataset, lacking validation on an established clinical benchmark and not addressing the holistic reasoning required

Inclusion Criteria: *Established chronic heart failure ($\geq$ 6 months)*
Exclusion Criteria: *Unable to undergo CMR imaging or brain MRI*

Figure 2: Sample inclusion and exclusion criteria.

for clinical trial eligibility assessment.

3 Creating MIMIC-CTRL

In contrast to current publicly available datasets, real-world clinical data is inherently messy, noisy, and voluminous, with a single patient profile comprising extensive narrative notes, structured tables, and medication histories. To address this gap, we introduce MIMIC-CTRL-2026, a dataset featuring real-world patient records for criterion-level eligibility assessment against three heart-failure trials. The dataset comprises 800 expert-annotated patient-criterion pairs from 50 unique patients across three trials: 240 from Trial 1 (CONVERGE-HF) 30×8, 260 from Trial 2 (COMET) 20×13, and 300 from Trial 3 (FINALITY-HF) 20×15. The 20 patients annotated for COMET are a subset of the 30 sampled for CONVERGE-HF, while the 20 patients annotated for FINALITY-HF are disjoint from both. Each data instance pairs *one* eligibility criterion with a single patient record. Trials are described in section 3.1.

The patient records are from the MIMIC-IV dataset version 3.1 and the MIMIC-IV-Note module version 2.2[1][2]. A cohort of 5,525 unique heart failure patients was identified by filtering for relevant International Classification of Diseases (ICD) codes, specifically those starting with I50 (ICD-10) and 428 (ICD-9). For these patients, free-text discharge notes and radiology reports were extracted alongside structured longitudinal data, including prescriptions and laboratory results for BNP and NT-proBNP. Because structured echocardiography data were unavailable during the extraction phase, these measurements were sourced directly from the clinical notes. Additionally, medication data were refined to focus specifically on heart failure management, capturing relevant prescriptions for 5,270 patients within the cohort.

3.1 Clinical Trials

Sample inclusion and exclusion criteria are presented in Figure 2. Given the medical expertise of the annotators who partook in this task, we limited our analysis to three cardiology trials: CONVERGE-HF[3], which investigates vericiguat in heart-failure patients with concomitant mild-to-moderate cognitive impairment; FINALITY-HF[4], which investigates the efficacy and safety of finerenone in select patients with heart failure intolerant to or ineligible for steroidal mineralocorticoid receptor antagonist therapy; and COMET (Poole-Wilson et al., 2002), which investigates whether beta-blocking agents with differing pharmacological profiles exert different effects on morbidity and mortality in patients with chronic heart failure. The inclusion and exclusion criteria for all three trials can be found in Appendix Tables 12, 13, and 14.

3.2 Annotation Procedure

While we initially targeted 30 patients for all three trials, given the time constraints of our clinical annotators, we were forced to limit two trials to 20 patients each. Manual annotation was performed by a resident physician and a medical student under the supervision of a clinician-trialist, who reviewed each patient-criterion pair independently before reaching consensus. Out of the total of 800 patient-criterion combinations, 149 cases (18.6%) had their classification altered during adjudication, with the biggest differences occurring in changes between the met/not-met classifications (55 cases) and between insufficient information/met classifications (46 cases).

For the annotation schema, we used 4 labels: 0 (the patient fails to meet the criterion or has a disqualifying condition), 1 (the patient meets the criterion or does not have the forbidden condition), 2 (the medical records do not contain enough data to determine eligibility), and 3 (the criterion is irrelevant to this patient's clinical context, e.g., pregnancy criteria for a documented male patient). Eligibility decisions were made directly at the patient–criterion level using the full patient record.

To enable manual annotation discharge notes were combined into a single document in *reverse chronological order* since eligibility assessments often prioritize a patient's current state rather than historical episodes. Radiology reports were simi-

[1] https://physionet.org/content/mimiciv/3.1/
[2] https://physionet.org/content/mimic-iv-note/2.2/
[3] www.clinicaltrials.gov/study/NCT06601465
[4] www.clinicaltrials.gov/study/NCT06033950

larly processed. Prescription records were filtered to include only heart failure-relevant drugs as suggested by the clinical experts. Vital signs were summarized per visit and specific lab values were extracted as requested by annotators (e.g., ejection fraction).

Initial pairwise agreement between the two annotators was 78% (Cohen's $\kappa = 0.57$), indicating moderate-to-substantial agreement and reflecting the genuine difficulty of the task for trained clinical specialists.

3.3 Data Pre-processing for LLM Ingestion

Patient-level data was extracted and matched with the relevant trial data. The raw MIMIC-IV patient records are extremely token-dense, often exceeding 20,000 tokens per patient due to multiple admissions. Appendix A presents the full pre-processing details.

4 Patient Matching Experiments

The majority of the experiments were executed using Llama 3.3 70B loaded with 4-bit NF4 quantization (Dettmers et al., 2023).

First, we evaluated four prompting configurations of increasing sophistication: Naive Prompt (NP), Naive Prompt with Processed Data (NP-PD), Expanded Prompt (EP), and Expanded Prompt with Processed Data (EP-PD).

Second, we evaluate if performance of the best-performing configuration (EP-PD) through the addition of a retrieval-augmented generation (RAG) component (Lewis et al., 2020) (EP-PD-RAG).

Third, we explore if clinical domain pre-training or explicit chain-of-thought reasoning provide additional gains by evaluating the best configuration (EP-PD) using MedGemma-27B (Sellergren et al., 2025), a model fine-tuned on biomedical text, and DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI, 2025), a reasoning-specialised model. All models were deployed locally to comply with the MIMIC responsible use agreement.

Finally, beyond quantitative evaluation, we perform qualitative analysis of model errors and generated reasoning to expose systematic failure modes on realistic clinical text. As the primary contribution of this paper is the new dataset, full technical and prompt details are provided in the Appendix.

4.1 Naive Prompt (NP)

Naive prompting uses the trial criteria with minimal modification as the prompt to an LLM with the patient information as context. This evaluates model robustness without trial-specific prompt customization.

For SIGIR_2016, to match previous works, we construct a generic prompt that accommodates a system instruction, criteria (which differ across patients), an eligibility score rubric, and an output format. On a per-trial basis, the model outputs JSON with a strict four-point eligibility score: Inclusion" (meets inclusion or does not meet exclusion), Not Included" (does not meet inclusion or meets exclusion), Not Applicable" (e.g., pregnancy tests for males), or Not Enough Information" (insufficient data). Further technical details can be found in the Appendix.

For TREC_2021, we applied the same set-up, aggregating all inclusion and exclusion criteria for a specific patient–trial pair into a single comprehensive prompt to facilitate comparison with past work. The model assigns one of three labels: non-relevant (0), excluded (1), or eligible (2).

For MIMIC-CTRL, eligibility is categorized using numerical scale: **1** denotes inclusion or non-exclusion, **0** denotes non-inclusion or exclusion, **2** indicates insufficient information, and **3** signifies that the criterion is not applicable. The naive prompt for MIMIC-CTRL closely followed the structure used in SIGIR and TREC experiments, serving as a faithful local re-implementation of the TrialGPT paradigm using an open-weight model.

4.2 Naive Prompt with Processed Data (NP-PD)

For the MIMIC-CTRL dataset, we address the high token density of patient records by implementing a hybrid, section-aware preprocessing pipeline. Rather than providing the model with raw, unstructured narratives, our system performs clinical segmentation to decompose discharge notes into standardised categories such as 'History of Present Illness' and 'Discharge Medications'.

Deterministic regular expressions isolate individual sections and extract key clinical indicators, such as ejection fraction and laboratory results, from headline-content pairs. For complex or unstructured entities, we employ the LLM as fallback to ensure comprehensive data extraction. These components are concatenated into a refined representation, significantly reducing noise compared to the raw narratives.

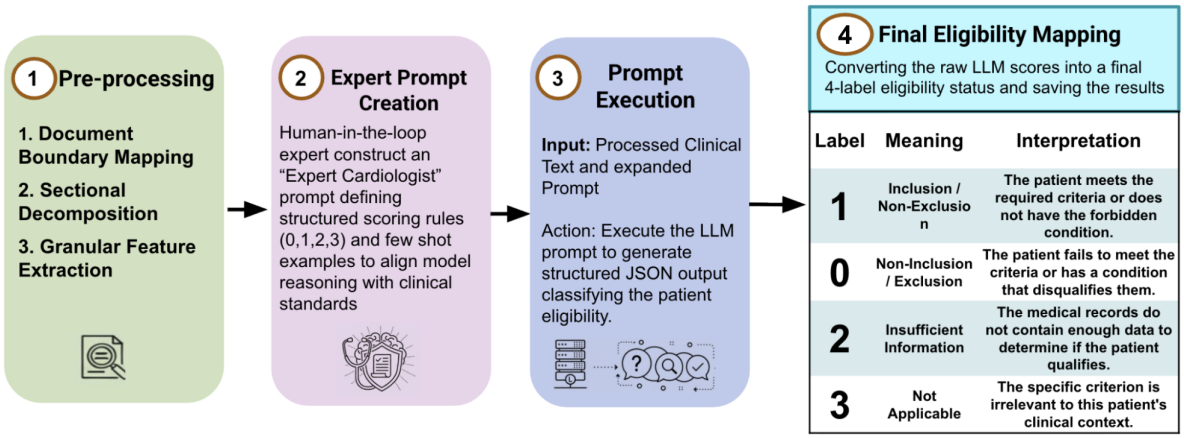


Figure 3: Illustration of Expanded Prompt-Processed Data methodology as applied to MIMIC-CTRL.

4.3 Expanded Prompt (EP)

Our clinical annotators developed enriched prompts referred to as the *Expanded Prompt*. These expanded prompts leveraged our clinical experts' knowledge to improve the original prompt using natural language. Prompts were expanded to identify where to look (i.e., specifying which section of the record to search), what to consider when assessing criterion (e.g., 'able to undergo MRI' meaning no implanted electronic devices, foreign metallic bodies, pregnancy, or claustrophobia), or identifying relevant medications. Prompts were validated and iterated upon using two manually labelled patients not included in the final dataset.

4.4 Expanded Prompt with Processed Data (EP-PD)

Here, we integrated the enriched prompt with the structured preprocessing pipeline (Figure 3), simultaneously maximising both input quality and reasoning instruction quality. The full inference-time control flow is shown in Figure 5 in Appendix A.

4.5 EP-PD with RAG

We incorporated a retrieval-augmented generation (RAG) component into the EP-PD pipeline, adding sentence-level semantic retrieval within the sections identified by the existing section-level mapping. Each sentence within the mapped sections is encoded using BioLORD-2023-C (Remy et al., 2023) to identify the most relevant sentences, and the top-ranked sentences are then passed to the model. Only the mapped sections are indexed, so retrieval operates over candidate sections rather than the full record, with ChromaDB (Chroma Core, 2022) used as the vector store for the sentence embeddings. The aim was to test whether passing only the relevant sentences, rather than the full relevant sections, would preserve sufficient evidence while reducing the number of tokens the model must process. Results are discussed in Section 5.3.

4.6 Extended Model Evaluation

To assess whether clinical domain pre-training or explicit chain-of-thought reasoning provide gains beyond prompt engineering, we re-ran the EP-PD configuration with two additional models.

MedGemma-27B (Sellergren et al., 2025) is fine-tuned on biomedical and clinical text. It was deployed in native `bfloat16` precision.

DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI, 2025) generates an explicit chain-of-thought trace before outputting structured output. `max_new_tokens` was raised to 2,048 to accommodate the reasoning trace, and a five-stage parser strips the internal monologue before JSON parsing.

5 Quantitative Results

Dataset	ACC	M-Prec	M-Rec	M-F1	Kappa
SIGIR	0.90	0.87	0.85	0.85	0.84
TREC	0.74	0.75	0.64	0.65	0.54
MIMIC*	0.28	0.33	0.34	0.25	0.06

Table 1: Performance variation of the same naive prompt structure across datasets. LLaMA 3.3-70B is evaluated on SIGIR, TREC, and MIMIC-CTRL* — the mean across three label-set trials (CONVERGE-HF, COMET, FINALITY). *Acc.: Accuracy, M-Prec.: Macro Precision, M-Rec.: Macro Recall, M-F1: Macro F1-Score, κ: Cohen's Kappa.*

5.1 Cross-Dataset Performance Comparison:

Table 1 highlights substantial performance degradation as the evaluation moves from simplistic to real-world clinical data: while the model achieves good performance on SIGIR_2016 with a Cohen's Kappa of 0.84, this drops to 0.54 for TREC_2021 and collapses to 0.06 for MIMIC_CTRL[5]. We present this as a benchmark-difficulty gap, since the most prominent difference between the datasets is realism. Technically the datasets also differ in label structure (SIGIR/TREC use three labels vs. our four but this is a minor part of the overall dataset, Table 7 in Appendix). Differences between the tasks/datasets are discussed in more detail in Appendix E. These findings suggest naive prompting is effective for structured, non-complex patient-trial data but fails to provide clinically reliable eligibility decisions on real-world patient records.

5.2 Performance Across Prompting Strategies

Table 3 reports the mean performance of the Llama-3.3-70B-Instruct model for different prompting strategies across all three trials. We can see that processing the dataset led to consistent improvements – an increase in κ of 0.14 for the Naive Prompt and an increase of 0.28 for the Expanded Prompt. When comparing NP vs EP or NP_PD vs EP_PD, we also saw a consistent improvement. This demonstrates that both prompting strategies resulted in performance gains with EP_PD performing the best.

5.3 Cross-Model Performance Evaluation:

Table 2 presents the experimental results of MIMIC_CTRL using the best performing prompting strategy (EP_PD) for an expanded selection of models.

We did not find the addition of the RAG component to have a consistent effect, with κ dropping in one trial, increasing in another, and remaining unchanged in a third. A case-level analysis of RAG reveals that this inconsistency is criterion-dependent rather than reflecting a uniform improvement or degradation (see Appendix D for details). We did not see any performance improvements when comparing Qwen to Llama. The larger Qwen model (72B) performed better than the smaller (3B) model on two of the trials.

Using an LLM pre-trained in the clinical domain did not result in any visible improvements. We observe lowest performance on FINALITY, whose 15 criteria include several requiring multi-step temporal and pharmacological reasoning (e.g., CRT implantation within 30 days, CYP3A4 co-administration), where clinical pre-training does not compensate for insufficient reasoning capacity at 27 billion parameters. Similarly, DeepSeek-R1-Distill-Llama-70B, despite its explicit chain-of-thought reasoning trace, did not result in any visible improvements in performance.

5.4 Confidence Intervals

95% confidence intervals were calculated for each measure using bootstrap resampling of the data at the patient level with 1,000 resamples, Table 3. Resampling was done at the level of the individual patients and not at the level of individual patient-criterion combinations, since multiple criteria obtained from one and the same patient are not statistically independent samples. In each resample, patients were selected by random sampling with replacement and their corresponding criterion-level predictions used to calculate measures.

Table 4 shows these confidence intervals along with the corresponding point estimates for each combination of the pipelines. Importantly, the confidence interval for NP's Cohen's Kappa contains zero (-0.00, 0.07), which means that naive prompting is not significantly better than random agreement. This result strengthens the main claim of this research on the overestimation of naive LLM performance on simpler benchmarks. Furthermore, the confidence intervals of EP_PD are generally not overlapping with those of EP and NP_PD for most metrics, which means that the performance difference could not have been obtained by mere sampling fluctuations.

6 Qualitative Results

Disagreements between annotators and the model were mainly due to: a) differences in judgment, b) differences in knowledge, and c) differences in the ability to identify key information in complex free-text. In the process of screening patients the models struggled at times to identify implanted cardiac devices (e.g., pacemakers, ICDs, CRT), conditions or medications. For example when evaluating a patient's discharge summary the model identified sublingual nitroglycerin as a long acting nitrate, but this particular formulation is short acting and would not fulfill the criteria for exclusion in this

[5]We prioritise Cohen's Kappa as it accounts for chance agreement and class imbalance in our multi-label setting.

Model	Strategy	CONVERGE-HF			COMET			FINALITY		
		Acc.	M-F1	κ	Acc.	M-F1	κ	Acc.	M-F1	κ
LLaMA 3.3-70B	EP_PD	**0.83**	**0.56**	**0.63**	**0.70**	**0.53**	**0.47**	0.71	0.57	0.38
LLaMA 3.3-70B	EP_PD + RAG	0.76	0.50	0.46	0.68	0.52	0.46	**0.72**	**0.65**	**0.46**
MedGemma 27B	EP_PD	0.79	0.52	0.56	0.51	0.23	0.02	0.25	0.37	−0.03
DeepSeek	EP_PD	0.48	0.23	0.39	0.39	0.22	−0.07	0.40	0.21	−0.04
LLaMA 3.1	EP_PD	0.35	0.28	−0.02	0.44	0.43	0.16	0.57	0.31	0.21
Qwen 72B	EP_PD	0.55	0.46	0.28	0.42	0.43	0.14	0.33	0.43	0.08
Qwen 2.5 3B	EP_PD	0.26	0.27	0.10	0.48	0.46	0.20	0.25	0.13	0.03

Table 2: Performance across all label sets on MIMIC-CTRL. Best accuracy per label set in **bold**. Full results including M-Prec. and M-Rec. are in the Appendix (supplementary material Table: 5). *Acc.: Accuracy, M-F1: Macro F1-Score, κ: Cohen's Kappa.*

Strategy	Acc.	M-P	M-R	M-F1	κ
NP	0.28	0.33	0.34	0.25	0.06
NP_PD	0.43	0.39	0.45	0.34	0.20
EP	0.51	0.55	0.42	0.40	0.22
EP_PD	0.75	0.63	0.57	0.55	0.50

Table 3: Mean LLaMA 3.3-70B-Instruct performance averaged across all three trials. *Acc.: Accuracy, M-P: Macro Precision, M-R: Macro Recall, M-F1: Macro F1, κ: Cohen's Kappa.*

patient's example. In other instances the model correctly identified relevant factors considered by the annotator, but came to the incorrect conclusion. For example, it correctly identified that a patient had multiple cardiovascular comorbidities such as coronary artery disease and hypercholesterolemia, but incorrectly assumed that this suggested on its own that the patient may have fulfilled the criteria for having chronic heart failure which is a related, but separate condition. Instances where the model was determined to be correct as opposed to the gold label resulted in an updating of the gold label.

The following observations were documented through structured post-annotation discussions with the resident physician and medical student annotators. They mentioned that there were numerous instances where the model effectively identified information which was missed by annotators in their review that ultimately led to the more correct determination by the model. There are also some instances of hallucinations by the model regarding the information contained in the patient documents, leading it to make incorrect conclusions.

The model's explanations were usually correct when it had correctly classified the pairing. Explanations for incorrect predictions were somewhat incoherent and failed to address the core qualities of the criteria that determined if it would be fulfilled or not – the model's explanations cannot be used to reliably infer likelihood of correct answer.

7 Discussion

Screening patients for participation in clinical trials is a challenging, time-intensive task. There is a clear role for model assisted screening in reducing the burden of trial screening particularly with complex well demarcated criteria that require little clinical interpretation. Interpretation of patient information and trial criteria is challenging even for clinicians; even with a consensus screening design determinations are marred with ambiguity. More specifically, during human annotations, disagreements in the screening process arose both from the identification of disparate information and conflicting interpretations of identical data, underscored by the inherent complexity of clinical documentation. These variations are exacerbated by the ambiguity of eligibility criteria, which often lack standardized thresholds, such as specific hemoglobin levels for "severe anemia," thereby allowing subjective clinical experience to dictate divergent interpretations.

The amount of information available to inform clinicians in trial screening decision-making is inversely proportional to the amount of ambiguity within these decisions. However, this volume of data presents specific challenges for automated systems; our results indicate that performance degrades as trial complexity increases. Trials with a greater number of criteria, such as FINALITY-HF with 15 criteria, consistently yield lower Kappa scores than simpler trials such as CONVERGE-HF with 8 criteria. This suggests sensitivity to overall prompt complexity rather than any per-instance processing of multiple criteria.

Even in unified electronic medical record systems, the completeness of documentation is variable and clinicians may be forced to follow up with patients and other care providers. For this reason, we disagree with Stubbs et al. (2019), which concluded that all instances of "possibly met" could

Strategy	Acc.	M-P	M-R	M-F1	κ
NP	0.28 [0.24, 0.30]	0.33 [0.28, 0.35]	0.34 [0.30, 0.41]	0.25 [0.23, 0.31]	0.06 [−0.00, 0.07]
NP_PD	0.43 [0.39, 0.46]	0.39 [0.35, 0.40]	0.45 [0.37, 0.54]	0.34 [0.31, 0.37]	0.20 [0.15, 0.23]
EP	0.51 [0.47, 0.54]	0.55 [0.53, 0.58]	0.42 [0.36, 0.47]	0.40 [0.38, 0.50]	0.22 [0.12, 0.23]
EP_PD	**0.75 [0.71, 0.77]**	**0.63 [0.59, 0.74]**	**0.57 [0.52, 0.63]**	**0.55 [0.53, 0.63]**	**0.50 [0.44, 0.55]**

Table 4: Mean LLaMA 3.3-70B-Instruct performance averaged across all three trials, with 95% bootstrap confidence intervals (patient-level resampling, $N = 1000$ iterations). Best result per column in **bold**. *Acc.: Accuracy, M-P: Macro Precision, M-R: Macro Recall, M-F1: Macro F1, κ: Cohen's Kappa.*

be replaced with a binary decision. We believe that an intermediate determination such as "possibly" is the result of a cautious approach to uncertainty which maintains the integrity of clinical trials and reduces the risk of subjecting patients to preventable harms. In a non-clinical research setting where information availability has substantial limitations, such an intermediate determination is even more critical.

While the current work focuses on criterion-level eligibility evaluation, the final downstream usage is at the level of trial aggregation. Having the criteria-level labels allows researchers to easily aggregate their predictions based on their particular approaches. More details are provided in Appendix A.2.

8 Conclusion

In this work, we introduced **MIMIC-CTRL-2026**, to our knowledge the first publicly available dataset for criterion-level eligibility assessment that pairs individual eligibility criteria with full longitudinal EHR records spanning multiple admissions. Our experiments revealed that LLM performance on synthetic data is over-estimated and doesn't generalize to the complexity of real-world clinical records. Our EP-PD model setup highlights the gains from both prompt engineering as well as structured preprocessing, paving a new path for research on automated clinical trial matching.

9 Limitations

Because of the sensitive nature of the data, there are not many publicly available datasets which could be annotated to create our dataset. This has various impacts on our work. First, our specific findings of model performance are less robust to performance on other datasets – though we think the substantial difference in performance indicates that trends will hold but the degree may differ.

Second, the sensitive nature of this dataset prevented us from directly comparing with existing baselines (e.g., TrialGPT) which route patient data through external servers, violating the responsible use agreement governing MIMIC access. While we believe our naive prompting baseline, a local re-implementation of the TrialGPT paradigm using an open-weight model, is a methodologically sound comparator, it is not equivalent in performance.

Our data source has an impact on disease burden (i.e., prevalence) and thus expected performance baselines: MIMIC is a dataset of patients admitted to the ICU which are more severe patients unlikely to be sought for the average cardiology trial (when compared to traditional cardiology outpatients). MIMIC also redacted some social information to minimize the risk of re-identification – information which could change a patients eligibility (and our model's performance). The patterns observed in our work (e.g., which model family performs better) is thus not necessarily going to hold on other datasets.

Given the clinical expertise of our collaborators and the great cost of manual annotations, we had to limit ourselves to only three cardiology trials and observed results may not hold to other medical specialties (e.g., trials from other specialties may have harder or easier criteria for current LLMs). To accommodate a broader range of use, future work should focus on expanding the number of trials of different specialties in the dataset. Furthermore, the expanded prompt requires domain expertise to develop and must be adapted for each new trial's criteria, which limits its scalability without clinical collaboration. This is an important consideration for future work aiming to generalise the approach to new trials and specialties.

Our work is restricted to the English language (due to data availability), and the findings may not generalize well to other languages or multilingual clinical contexts. Further, our annotation of the dataset was manual, requiring expert physician annotators to go over entire patient files (which could span years) which limited the number of patient-trial pairs and the overall number of trials that could

be annotated.

The application of LLMs to enable trial matching is generally low-risk. However, the technology to find people based on description's from large, complex, and unstructured texts could be used in surveillance and other negative scenarios. While we don't believe that our work directly contributes to this (i.e., the data and code won't be useful for surveillance purposes), learning gained (e.g., the framework developed, and general findings) could be used for nefarious actions. Unfortunately, we don't think there is a way of preventing this while enabling our work. There are also other externalities to our work. For example, while less severe, this work required substantial time from physicians which theoretically could be spent providing care. Likewise, processing large amounts of data using LLMs may have negative effects on the environment.

Acknowledgments

Mohamed Abdalla is supported by a CIFAR AI chair and an NSERC Discovery Grant. Research reported in this work was partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-2024C3-42525). The views in this work are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

References

Shaika Chowdhury, Sivaraman Rajaganapathy, Yue Yu, Cui Tao, Maria Vassilaki, and Nansu Zong. 2024. Matching patients to clinical trials using Llama 2 embeddings and Siamese neural network. *medRxiv*.

Chroma Core. 2022. Chroma: The open-source embedding database. https://github.com/chroma-core/chroma. Accessed: 2026-05-22.

Pascal Coorevits, Mats Sundgren, Gunnar O Klein, Anne Bahr, Brecht Claerhout, Christel Daniel, Martin Dugas, Danielle Dupont, Andreas Schmidt, Peter Singleton, and 1 others. 2013. Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6):547–560.

Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, and 1 others. 2017. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. 2015. *Fundamentals of clinical trials*. Springer.

Aynur Guluzade, Naguib Heiba, Zeyd Boukhers, Florim Hamiti, Jahid Hasan Polash, Yehya Mohamad, and Carlos A Velasco. 2025. ELMTEX: Fine-tuning large language models for structured clinical information extraction. a case study on clinical reports. *arXiv preprint arXiv:2502.05638*.

Grant D. Huang, Jonca Bull, Kelly Johnston McKee, Elizabeth Mahon, Beth Harper, and Jamie N. Roberts. 2018. Clinical trials recruitment planning: A proposed framework from the Clinical Trials Transformation Initiative. *Contemporary Clinical Trials*, 66:74–79.

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.

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.

Aritra Kumar Lahiri, Emrul Hasan, Qinmin Vivian Hu, and Cherie Ding. 2024. Tmu at trec clinical trials track 2023. *arXiv preprint arXiv:2403.12088*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Katherine E McDonald, Ariel E Schwartz, and Maya Sabatello. 2022. Eligibility criteria in nih-funded clinical trials: Can adults with intellectual disability get in? *Disability and Health Journal*, 15(4):101368.

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.

Philip A Poole-Wilson, John GF Cleland, Andrea Di Lenarda, Peter Hanrath, Michel Komajda, Marco Metra, Willem J Remme, Karl Swedberg, and Christian Torp-Pedersen. 2002. Rationale and design of the carvedilol or metoprolol european trial in patients with chronic heart failure: Comet. *European Journal of Heart Failure*, 4(3):321–329.

François Remy, Kris Demuynck, and Thomas Demeester. 2023. Biolord-2023: Semantic textual representations fusing llm and clinical knowledge graph insights. *arXiv preprint arXiv:2311.16075*.

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

Maciej Rybinski, Wojciech Kusa, Sarvnaz Karimi, and Allan Hanbury. 2024. Learning to match patients to clinical trials using large language models. *Journal of Biomedical Informatics*, 159:104734.

Mina S Sedrak, Rachel A Freedman, Harvey J Cohen, Hyman B Muss, Aminah Jatoi, Heidi D Klepin, Tanya M Wildes, Jennifer G Le-Rademacher, Gretchen G Kimmick, William P Tew, and 1 others. 2021. Older adult participation in cancer clinical trials: a systematic review of barriers and interventions. *CA: a cancer journal for clinicians*, 71(1):78–92.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. MedGemma technical report. *arXiv preprint arXiv:2507.05201*.

Matthew S Simpson, Ellen M Voorhees, and William Hersh. 2014. Overview of the trec 2014 clinical decision support track. In *23rd 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.

Chi-en Amy Tai and Xavier Tannier. 2025. Clinical trial cohort selection using large language models on n2c2 challenges. *arXiv preprint arXiv:2501.11114*.

TREC Biomedical Tracks Organizers. 2023. 2023 TREC clinical trials track. https://www.trec-cds.org/2023.html. Accessed: 2025-12-18.

Thinh Hung Truong, Yulia Otmakhova, Rahmad Mahendra, Timothy Baldwin, Jey Han Lau, Trevor Cohn, Lawrence Cavedon, Damiano Spina, and Karin Verspoor. 2022. Ittc@ trec 2021 clinical trials track. *arXiv preprint arXiv:2202.07858*.

Ozan Unlu, Matthew Varugheese, Jiyeon Shin, Samantha M Subramaniam, David Walter Jacques Stein, John J St Laurent, Charlotte J Mailly, Marian J McPartlin, Fei Wang, Michael F Oates, and 1 others. 2025. Manual vs ai-assisted prescreening for trial eligibility using large language models—a randomized clinical trial. *JAMA*, 333(12):1084–1087.

US Food and Drug Administration. 2019. Rare pediatric disease priority review vouchers: Draft guidance for industry. Draft guidance.

Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W. Mahaffey, and Nigam H. Shah. 2024. Zero-shot clinical trial patient matching with LLMs. *arXiv preprint arXiv:2402.05125*.

A Pre-processing Details

SIGIR_2016: From the publicly available SIGIR dataset[6], we extracted patient summary text from <TOP> and <NUM> tags, while trial metadata and eligibility criteria were parsed from the clinical trial records. No heavy preprocessing was required, as each patient note consists of approximately 70-80 words; we simply numbered each sentence. These sources were combined into a table of patient–trial pairs. Ultimately, 1,016 unique patient–trial combinations were selected for experimental evaluation, using annotations from Jin et al. (2024).

TREC_2021: TREC provides 75 patient descriptions (referred to as "topics"[7]) and approximately 31,000 expert-annotated patient–trial labels. From these, we selected a subset of 2,074 unique combinations due to computational limitations. Patient notes and trial information were combined into a single table following the same format as SIGIR.

MIMIC: The MIMIC-IV dataset contains detailed clinical narratives including history of present illness, physical examinations, discharge medications, and family history. Since patients often have multiple admissions, a single patient record can exceed 20,000–25,000 tokens. Our initial processing consisted of passing these raw notes directly into the model, which consistently produced contexts too heavy for effective processing, consistent with findings on long-context limitations reported by Liu et al. (2024).

A.1 Data Preprocessing and Patient Selection

To address the token density problem, we developed a three-stage sequential pipeline (Figure 4) that transforms raw discharge notes into structured, criterion-relevant representations.

[6] https://data.csiro.au/collection/csiro:17152
[7] https://www.trec-cds.org/2021.html

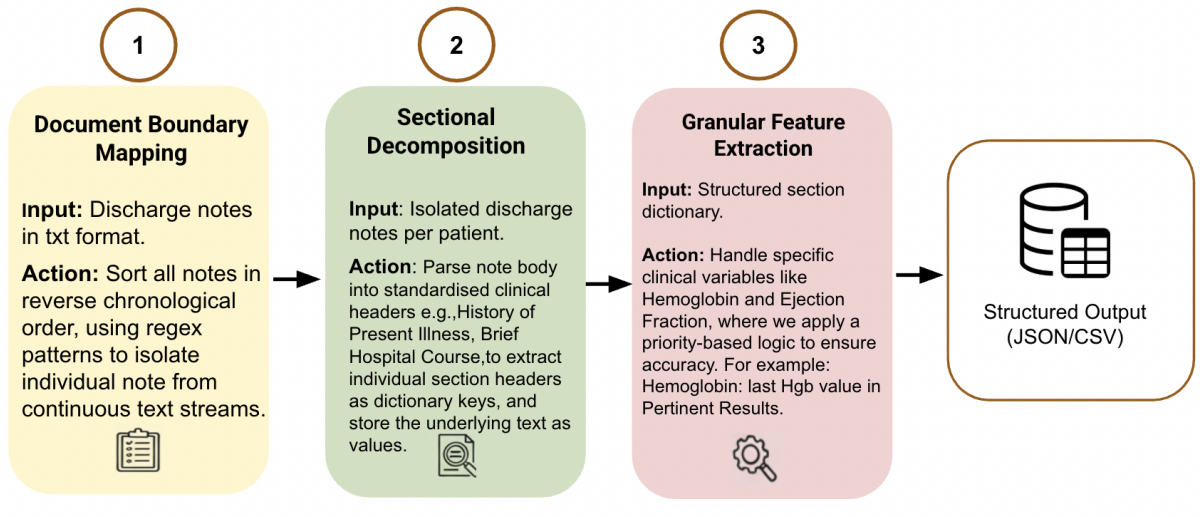


Figure 4: Illustration of Pre-Processing methodology for creating the MIMIC-CTRL dataset.

Stage 1 (Document Boundary Mapping): The pipeline identifies the boundaries of individual discharge notes within each input file, since multiple admissions are concatenated into a single document. Notes are then sorted in *reverse chronological order*, placing the most recent admission first, as eligibility assessments should reflect a patient's current clinical state rather than historical episodes.
Stage 2 (Sectional Decomposition): Each note is split into labelled clinical sections by reading through the text line by line. When a pre-defined header is encountered (e.g., *History of Present Illness*, *Past Medical History*, *Discharge Medications*, *Vital Signs*, *Brief Hospital Course*, *Pertinent Results*), the pipeline begins collecting the following lines as the content for that section until the next known header appears. The result is a nested dictionary of the form `{patient_id: {ds_number: {section: content}}}`, enabling constant-time retrieval of any section for any patient during inference. A criterion-to-section lookup table maps each eligibility criterion to the sections most likely to contain relevant evidence (e.g., the sGC contraindication criterion maps to Discharge Medications and Pertinent Results), preventing the full record from being passed to the model for every criterion.
Stage 3 (Granular Feature Extraction): Key clinical values are extracted from the structured sections using the following strategies:

- **Ejection Fraction (EF):** A four-level priority hierarchy is applied. Priority 1 searches the TTE/echo block within Pertinent Results, as this represents the measured echocardiogram from the current admission and is the most authoritative source. Priority 2 searches for any `LVEF=` pattern within Pertinent Results, taking the last match when multiple echo studies are present. Priority 3 checks the Brief Hospital Course. Priority 4 searches the History of Present Illness as a last resort, as this section often contains historical values from previous admissions.
- **Haemoglobin (Hgb):** The last recorded value in Pertinent Results is extracted rather than the first, ensuring the most recent lab draw is captured.
- **HFrEF Classification:** Rather than searching for the term `HFrEF` directly (which risks false positives), the label is derived numerically from the extracted EF value: EF $\leq 40\%$ = HFrEF, 41–50% = HFmrEF, $> 50\%$ = preserved ejection fraction. Cases where no EF value could be extracted fall back to keyword search and are flagged for manual verification.
- **Biological Sex:** Inferred from pronoun and noun patterns (e.g., male", gentleman", "female") within the History of Present Illness section and passed to the model as part of the patient-level metadata in the prompt, alongside patient ID, discharge note number, and matched section headers, to support sex-specific scoring (e.g., automatic Score 3 for pregnancy-related criteria in documented male patients).

For criteria involving explicit temporal windows (e.g., events within two months or thirty days), only

the most recent discharge note was considered, as earlier admissions cannot reliably satisfy recency constraints.

A.2 Criterion-Level Task Formulation

A fundamental design decision in this work is the formulation of eligibility assessment at the level of individual criteria rather than at the level of the full trial. We made this decision because, by providing labels at the criteria level, it becomes straightforward for researchers to create patient-trial labels. Moreover, it enables us to more effectively study current model limitations (e.g., identifying exactly where they underperform). Understanding this particular issue may have a significant impact on how criterion-level eligibility assessments are conducted and how LLMs are incorporated into this process.

Each data instance consists of a single eligibility criterion paired with a patient's clinical record; the system then produces a score using the four-label scheme defined already in Section 3.2. This granularity enables fine-grained error analysis and allows downstream users to aggregate criterion-level predictions into trial-level eligibility decisions according to their own strategies.

This formulation places complexity on the patient side rather than the criterion side. Each criterion is a single, specific clinical condition, while the patient record may span dozens of sections across multiple hospitalisations. This contrasts with prior datasets (SIGIR, TREC) that presented full criterion lists paired with simplified patient summaries. In MIMIC-CTRL-2026, the criterion representation is deliberately simple and singular while the patient representation is deliberately complex and realistic.

The asymmetry in patient counts across trials reflects the time constraints of the clinical annotators: although we started with the goal of annotating 30 patients for all three trials, after completing the first trial (CONVERGE-HF) and seeing how long it took for both annotators to complete the task and resolve disagreements, we could only annotate 20 additional patients for FINALITY-HF given our remaining budget, and reused a 20-patient subset of the CONVERGE-HF cohort for COMET rather than sampling entirely new patients.

All datasets, code, and models used in this study were collected and used in accordance with their respective licences and access requirements.

B Technical Details

B.1 SIGIR_2016

Inference uses temperature, T = 0, a 16,384-token context window, and a 512-token limit for new generations.

B.2 TREC_2021

We set the temperature 0 within a 16,384-token context window to ensure deterministic reasoning and classification results.

B.3 MIMIC

The prompt was limited to a maximum of 20,000 tokens (with irrelevant content trimmed if necessary), and inference was performed with a temperature of 0 and `max_new_tokens` set to 512. This configuration assessed the model's robustness when evaluated against all three trials' criteria using full clinical narratives, without trial-specific prompt customization or retrieval. Overall, this naive prompting strategy yielded poor performance, achieving low classification accuracies for both inclusion and exclusion criteria.

C Llama Model Details (General & Project-Specific)

The `Llama-3.3-70B-Instruct` model, a 70-billion-parameter LLM developed by Meta, is typically deployed using `bfloat16` or `float16` precision with higher-temperature sampling to support creative text generation. In this project, however, the model was configured for efficient, resource-constrained multi-GPU inference by applying 4-bit NF4 quantization via the `BitsAndBytes` library in conjunction with `float16` precision. This quantization strategy reduced the model's VRAM requirements from approximately 140 GB to roughly 35–40 GB, enabling deployment on limited hardware resources. Furthermore, text generation was performed using a low temperature ($T = 0$) to ensure consistent and deterministic outputs, which are critical for maintaining high fidelity in clinical reasoning tasks involving long and complex patient records.

D Case-Level RAG Analysis

A case-level analysis across all three trials reveals that RAG's effect is highly trial-dependent, explaining the inconsistent aggregate κ scores observed in Table 2. Table 6 summarises the number of cases where each approach is uniquely correct.

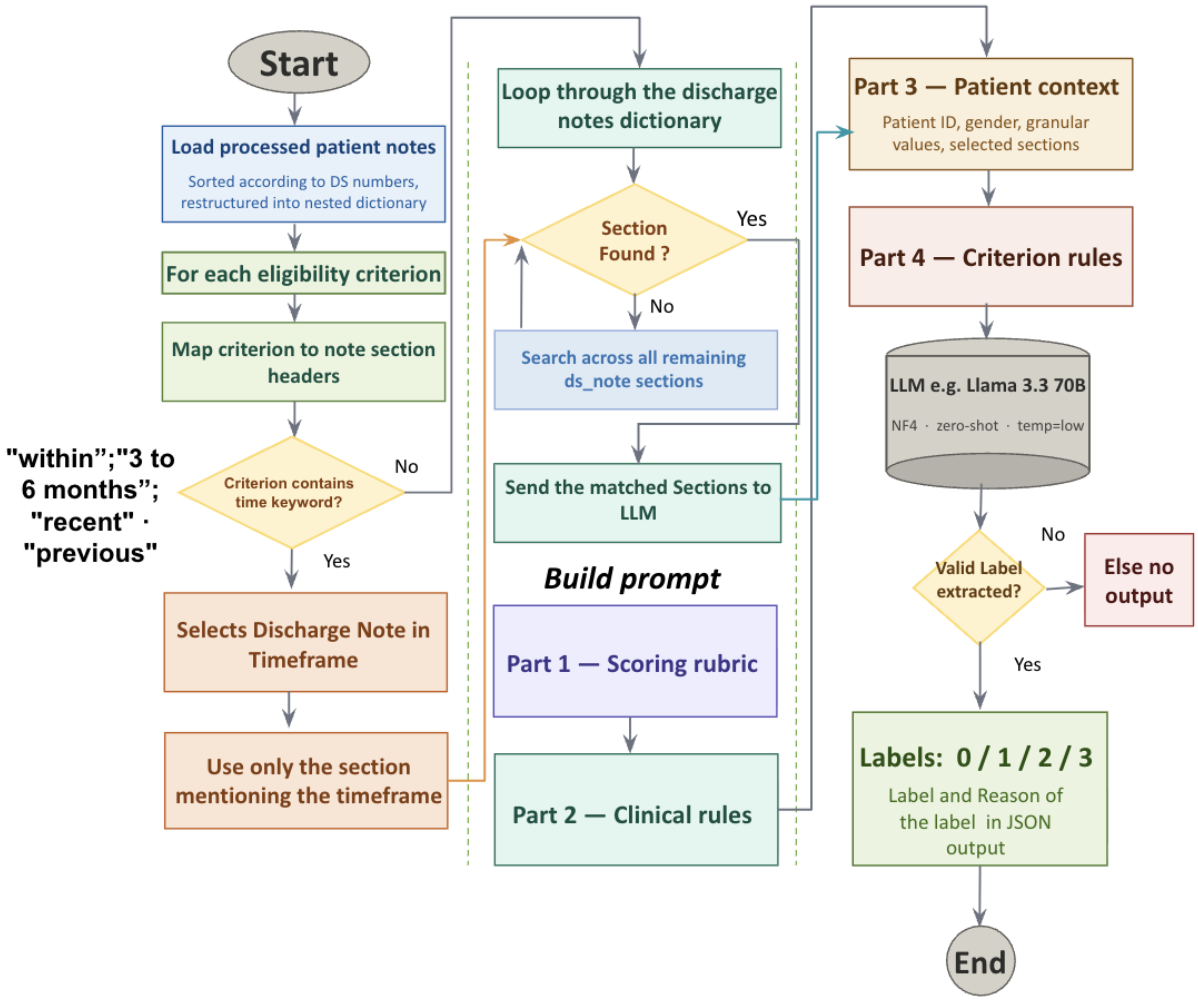


Figure 5: System workflow for expanded prompting with processed discharge note data. For each criterion, the pipeline first loads the pre-processed notes, organised as a structured nested dictionary, and maps the criterion to its relevant sections, selecting the note from the appropriate time frame when the criterion is time-sensitive. It then retrieves the matching sections, searching the remaining sections if none are found, and assembles a prompt comprising the scoring rubric, the clinical rules, the patient context, and the criterion-specific rules. The prompt is passed to the LLM, which returns the final label and its reasoning as structured JSON.

This pattern is explained by the nature of the criteria in each trial. For CONVERGE-HF, EP-PD's advantage stems from criteria that require holistic reasoning across multiple sentences. For example, assessing mild-to-moderate cognitive impairment requires integrating mentions of MCI (Mild Cognitive Impairment), dementia severity, or MoCA (Montreal Cognitive Assessment) scores across the Brief Hospital Course and Past Medical History sections of clinical notes. Section-level mapping, by passing the entire relevant section to the model, preserves this cross-sentence context more effectively than sentence-level retrieval. Conversely, for COMET, RAG demonstrates particular strength on criteria involving precise numerical thresholds and specific medication checks, even though EP-PD remains uniquely correct on a somewhat larger share of cases overall. For example, determining whether furosemide dose meets the $\geq$40mg threshold, whether amiodarone exceeds 200mg/day, or whether LVEF $\leq$35% is documented all depend on a single precise sentence in the clinical note, where sentence-level retrieval locates the exact evidential sentence more reliably than passing entire sections to the model. For FINALITY, the mix of numerically precise criteria (eGFR $<$25, potassium $>$5.0) and holistic reasoning criteria (probable alternative cause of HF, any other condition making the patient unsuitable) results in RAG and EP-PD correcting a comparable number of each other's errors, producing no consistent net gain. These findings suggest that RAG works well for precise numerical criteria,

Model	Strategy	CONVERGE-HF					COMET					FINALITY				
		Acc.	M-P	M-R	M-F1	κ	Acc.	M-P	M-R	M-F1	κ	Acc.	M-P	M-R	M-F1	κ
LLaMA 3.3-70B	EP_PD	**0.83**	**0.56**	**0.56**	**0.56**	**0.63**	**0.70**	**0.67**	**0.51**	**0.53**	**0.47**	0.71	0.67	0.63	0.57	0.38
LLaMA 3.3-70B	EP_PD + RAG	0.76	0.49	0.51	0.50	0.46	0.68	0.66	0.51	0.52	0.46	**0.72**	**0.71**	**0.69**	**0.65**	**0.46**
LLaMA 3.3-70B	EP	0.61	0.47	0.41	0.41	0.25	0.60	0.60	0.53	0.56	0.35	0.33	0.29	0.31	0.24	0.05
LLaMA 3.3-70B	NP_PD	0.47	0.42	0.33	0.33	0.20	0.47	0.40	0.50	0.39	0.26	0.35	0.34	0.52	0.31	0.13
LLaMA 3.3-70B	NP	0.30	0.36	0.35	0.26	0.03	0.38	0.43	0.50	0.36	0.19	0.15	0.21	0.18	0.12	−0.05
MedGemma 27B	EP_PD	0.79	0.52	0.53	0.52	0.56	0.51	0.22	0.26	0.23	0.02	0.25	0.52	0.41	0.37	−0.03
DeepSeek	EP_PD	0.48	0.23	0.23	0.23	0.39	0.39	0.26	0.23	0.22	−0.07	0.40	0.25	0.26	0.21	−0.04
LLaMA 3.1	EP_PD	0.35	0.32	0.32	0.28	−0.02	0.44	0.54	0.42	0.43	0.16	0.57	0.31	0.32	0.31	0.21
Qwen 72B	EP_PD	0.55	0.53	0.53	0.46	0.28	0.42	0.54	0.40	0.43	0.14	0.33	0.44	0.52	0.43	0.08
Qwen 2.5 3B	EP_PD	0.26	0.50	0.40	0.27	0.10	0.48	0.56	0.44	0.46	0.20	0.25	0.25	0.27	0.13	0.03

Table 5: Full performance across all label sets on MIMIC-CTRL, including Macro Precision and Macro Recall. Best accuracy per label set in **bold**. *Acc.: Accuracy, M-P: Macro Precision, M-R: Macro Recall, M-F1: Macro F1-Score, κ: Cohen's Kappa.*

Trial	Total	Both	RAG only	EP-PD only
CONVERGE-HF	240	167	13	33
COMET	260	155	21	28
FINALITY	300	185	31	27

Table 6: Case-level comparison of RAG and EP-PD across trials. *Both*: cases where both approaches are correct. *RAG only*: cases where RAG is correct and EP-PD is not. *EP-PD only*: cases where EP-PD is correct and RAG is not.

while EP-PD works better when broader clinical reasoning is needed across multiple sentences.

E Cross-Dataset Confound Details

This section expands on the differences between SIGIR_2016, TREC_2021, and MIMIC-CTRL that contribute to the benchmark-difficulty gap discussed in Section 5.1.

Label structure: SIGIR and TREC use a 3-label annotation schema (met, not met, insufficient information), while MIMIC-CTRL uses a 4-label schema that adds a "Not Applicable" category for criteria that are biologically or clinically irrelevant to a given patient (e.g., pregnancy-related criteria for a documented male patient). Table 7 reports the gold-label distribution across our 800 patient-criterion pairs. The "Not Applicable" label comprises only 3.2% of gold instances, suggesting its direct contribution to the observed cross-dataset performance drop is limited relative to the other confounds discussed below, though we do not attempt to fully decompose this confound from the comparison alone.

Task granularity: SIGIR and TREC present the model with the full list of trial criteria alongside a single, simplified patient summary, requiring the model to reason over all criteria jointly. MIMIC-CTRL instead formulates the task at the level of a single criterion per inference call (Appendix A), paired with the patient's full longitudinal record. This finer granularity changes what the model must do at each step and is not, by itself, a measure of record realism.

Criterion complexity: The eligibility criteria used in MIMIC-CTRL's three trials (CONVERGE-HF, COMET, FINALITY-HF) include multi-part, temporally-qualified, and numerically-thresholded conditions (e.g., criteria requiring both a specific lab value within a time window and the absence of a named medication class). SIGIR and TREC's criteria are comparatively less structurally complex, independent of the patient representation used to evaluate them.

Patient representation: SIGIR and TREC pair each criterion list with a short, single-paragraph synthetic patient summary. MIMIC-CTRL pairs each criterion with a full, real, longitudinal patient record spanning multiple discharge notes, radiology reports, laboratory results, and medication histories.

Gold Label	Meaning	Count	%
0	Not Met / Excluded	183	22.9%
1	Met / Not Excluded	498	62.2%
2	Insufficient Information	93	11.6%
3	Not Applicable	26	3.2%

Table 7: Gold-label distribution across all 800 patient-criterion pairs in MIMIC-CTRL-2026.

F Licenses and Terms of Use

MIMIC-IV (v3.1) and MIMIC-IV-Note (v2.2) are released under the PhysioNet Credentialed Health Data License 1.5.0; all experiments were conducted locally to comply with these terms and no patient data was transmitted to external servers. The SIGIR-2016 and TREC-2021 datasets were used in accordance with their respective data portal terms. Pre-trained models (LLaMA 3.3-70B,

MedGemma-27B, DeepSeek-R1-Distill-LLaMA-70B, Qwen 2.5) were used in accordance with their respective community licenses. As MIMIC-CTRL-2026 is derived from processing MIMIC-IV data, it will be published under the PhysioNet Credentialed Health Data License 1.5.0. All datasets and models were used solely for research purposes, consistent with their intended use as specified in their respective licenses and access agreements. MIMIC-IV data is de-identified by PhysioNet prior to release; all patient names, dates, and other direct identifiers are removed or replaced in accordance with HIPAA Safe Harbor standards. Our annotation process introduced no new personally identifying information, as annotators worked only with de-identified records and assigned numerical eligibility labels.

G Compute

The experiments described in the paper were performed on NVIDIA A100-SXM4-40GB GPUs. All fine-tuning and inference experiments were conducted on 2-3 GPUs. Running all models took approximately 15-20 hours in total.

H Use of AI Assistants

An AI assistant was used only for spelling, grammar, and phrasing.

I All Prompts and Criteria of the Three Trials

Component	Content
System Instruction	You are an expert physician matching patients to clinical trials. Evaluate the patient note against the following inclusion or exclusion criterion.
Instructions	1. Determine whether the criterion is applicable to the patient. 2. Identify whether there is direct evidence supporting or contradicting the criterion. 3. If information is unclear or missing, label as *not enough information*.
Output Format	Return **only** a JSON object with the keys: `Label` and `reasoning`. The value of `Label` must be one of `0`, `1`, `2`, or `3`.
Patient Note	*[patient's all clinical information]*
Criterion	*[Clinical trial inclusion or exclusion criterion]*
Eligibility Labeling Rubric	**Label** — **Interpretation** **1** — Meets inclusion or does not meet exclusion **0** — Does not meet inclusion or meets exclusion **2** — Insufficient information available **3** — Criterion not applicable

Table 8: `Naive_Prompt_for_MIMIC_IV`. Note: each inference call passes a single criterion at a time; one JSON output is produced per criterion–patient pair.

Component	Content
System Instruction	You are a medical expert evaluating patient eligibility for the *[trial_name]* clinical trial. Based on the provided discharge note, determine if the patient meets the following *[criteria_type]* criterion.
Scoring System	**Score** — **Interpretation** **1** — Inclusion met OR exclusion not triggered **0** — Inclusion not met OR exclusion triggered **2** — Genuinely insufficient or ambiguous information — use **only** when evidence is present but unclear **3** — Biologically not applicable — for example use **only** for sex-specific criteria for a documented male patient
Critical Rules	• **Never** use score 3 for "condition not found"; use 0 or 1 as appropriate. • **Never** use score 2 as a default when information is absent. Absent information means the condition is not present. Exception: use score 2 only when there is some mention but it is ambiguous. • Negation: "no history of X", "denies X", "no X" → treat X as absent. • Temporal: for time-windowed criteria, if an event appears in Past Medical History without a required time duration, assume it is outside the window.
Patient Metadata	Patient ID: *[patient_id]* Discharge Note: *[patient_id]*-DS-*[ds_number]* Gender: *[gender]* Matched Sections: *[matched_headlines]*
Patient Note	*[Relevant processed note content]*
Criterion (*[criteria_type]*)	*[Single inclusion or exclusion criterion — one per inference call]*
Criterion-Specific Decision Rules	*The full set of decision rules for this trial is included as context to guide scoring. Only the rule matching the active criterion above is applied per call.* • **Adult patients**: Always score 1. • **Established chronic heart failure ($\geq$ 6 months)**: Check HPI, PMH, Discharge Diagnosis for `CHF`, `HFrEF`, `heart failure`. Score 1 if found with chronicity implied; 0 if not found; 2 if onset explicitly <6 months. • **Mild-to-moderate cognitive impairment**: Check BHC, PMH for `MCI`, `cognitive impairment`, `Alzheimer`, `dementia` (mild/moderate only). Score 1 if found; 0 if absent. • **Contraindications for sGC stimulator / vericiguat**: Check Discharge Medications for `riociguat`, `sildenafil`, `tadalafil`, `Hgb < 8`, `isosorbide dinitrate` (long-acting only). Score 0 if found; 1 if absent. • **Unable to undergo CMR / Brain MRI**: Check for `pacemaker`, `ICD`, `CRT`, `PPM`. Score 0 if confirmed implanted; 1 if absent. • **CMR exclusions**: Check for `GFR < 30`, `ESRD`, `ICD`, `CRT`, uncontrolled `AF`, recurrent `VT/VF`. Score 0 if found; 1 if absent. • **General medical conditions**: Check for cancer (except basal cell), hepatic failure, uncontrolled thyroid, CABG/PCI within 3 months, moderate–severe dementia. Score 0 if found; 1 if absent. • **Allergies to study products**: Check for allergy to vericiguat specifically. Score 0 if found; 1 if absent or no allergies mentioned.
Output Format	Return **only** a valid JSON object. No text outside the JSON.
	Example: `{"score": "1", "reasoning": "..."}`

Table 9: Expanded Prompt for CONVERGE-HF Trial. The `Criterion` field receives a **single** inclusion or exclusion criterion per inference call; one JSON object is returned per criterion–patient pair. Decision rules for all criteria are included as prompt context but only the rule matching the active criterion is applied.

Component	Content
System Instruction	You are a medical expert evaluating patient eligibility for the COMET clinical trial. Based on the provided discharge note, determine if the patient meets the following *[criteria_type]* criterion.
Scoring System	**Score** **Interpretation** **1** Inclusion met OR exclusion not triggered **0** Inclusion not met OR exclusion triggered **2** Genuinely insufficient or ambiguous information — use **only** when evidence is present but unclear **3** Biologically not applicable — for example use **only** for sex-specific criteria for a documented male patient
Critical Rules	• **Never** use score 3 for "condition not found"; use 0 or 1 as appropriate. • **Never** use score 2 as a default when information is absent. Absent information means the condition is not present. Exception: use score 2 only when there is some mention but it is ambiguous. • Negation: "no history of X", "denies X", "no X" → treat X as absent. • Temporal: for time-windowed criteria, if an event appears in Past Medical History without a required time duration, assume it is outside the window.
Patient Metadata	Patient ID: *[patient_id]* Discharge Note: *[patient_id]*-DS-*[ds_number]* Gender: *[gender]* Matched Sections: *[matched_headlines]*
Patient Note	*[Relevant processed note content]*
Criterion (*[criteria_type]*)	*[Single inclusion or exclusion criterion — one per inference call]*
Criterion-Specific Decision Rules (Part 1)	*The full set of decision rules for this trial is included as context to guide scoring. Only the rule matching the active criterion above is applied per call.* • **NYHA Class II–IV (inclusion)**: Search HPI, Discharge Diagnosis, BHC for NYHA class or symptoms (dyspnea, orthopnea, volume overload). Score 1: NYHA II–IV or consistent symptoms. Score 0: class I or no HF symptoms. Score 2: HF mentioned but functional class entirely unclear. • **$\geq$1 CV Admission Past 2 Years (inclusion)**: Search PMH, HPI for HF, NSTEMI, STEMI, CAD, CABG, stroke. Current admission counts if cardiovascular. PMH CV event without date → assume within 2 years → Score 1. Score 0: no CV admissions found. • **Stable HF Treatment (inclusion)**: Search Discharge Medications for ACE inhibitor (lisinopril, ramipril, enalapril) AND diuretic $\geq$40 mg furosemide equivalent. Medication without start date → assume ongoing. Score 1: both met or appropriately substituted. Score 0: one or both absent. Score 2: medication list absent. • **Cardiac Function (inclusion)**: Search Pertinent Results, BHC, PMH for EF value. EF $\leq$35% → Score 1. EF >35% or preserved/normal → Score 0. Score 2: no cardiac function data. • **Recent Change in HF Treatment (exclusion)**: Search Discharge Medications, BHC for diltiazem, verapamil, amiodarone (>200 mg/day), flecainide, IV inotropes, investigational drug. Score 0: any found. Score 1: none found. Score 2: medication list entirely absent. • **Unstable Angina/MI/CABG/Stroke $\leq$2 Months (exclusion)**: Search HPI, PMH, Discharge Diagnosis for unstable angina, NSTEMI, STEMI, MI, CABG, stroke. Current admission → Score 0. PMH without date → Score 1. Score 2: ambiguous timing. • **Uncontrolled Hypertension SBP >170 or DBP >105 (exclusion)**: Search Pertinent Results for BP values. Score 0: consistently elevated. Score 1: within limits or none documented. Score 2: no BP documentation.

Table 10: Expanded Prompt for COMET Trial (Part 1 of 2). The `Criterion` field receives a **single** inclusion or exclusion criterion per inference call; one JSON object is returned per criterion–patient pair.

Component	Content
Criterion-Specific Decision Rules (Part 2)	*Continued from Part 1. Only the rule matching the active criterion is applied per call.* • **Haemodynamically Significant Valvular Disease (exclusion)**: Search Discharge Diagnosis, BHC, PMH. Only "severe" or "hemodynamically significant" qualifies. Mild/moderate does NOT exclude. Score 0: severe documented. Score 1: none or mild/moderate only. • **Sustained VT/VF <2 Months Untreated (exclusion)**: Search Discharge Diagnosis, BHC, PMH. Current admission → Score 0. ICD or antiarrhythmic confirmed → Score 1. PMH without date → Score 1. • **Pregnancy or Inadequate Contraception, Females Only (exclusion)**: Score 3: documented male. Score 1: female with documented contraception, post-menopausal, or surgically sterile. Score 0: pregnant or inadequate contraception. Score 2: female, contraception status unknown. • **Drug or Alcohol Misuse (exclusion)**: Search Social History, PMH, HPI for alcohol abuse, substance abuse, AUD, SUD. Score 0: active misuse. Score 1: none documented. Score 2: Social History entirely absent. • **Poor Compliance or Serious Systemic Disease (exclusion)**: Search BHC, PMH, Discharge Diagnosis for non-compliant, metastatic cancer, terminal illness, hospice, end-stage, severe liver disease, renal failure on dialysis. Score 0: any found. Score 1: none found. • **Contraindication to β-Blockers (exclusion)**: Search PMH, Discharge Diagnosis, Pertinent Results for HR <60, sick sinus syndrome, 2nd/3rd degree AV block, SBP <85, asthma, COPD, PAD, unstable insulin-dependent diabetes. Pacemaker confirmed → AV block override → Score 1. Score 0: any found without override. Score 1: none found.
Output Format	Return **only** a valid JSON object. No text outside the JSON.
	Example: `{"score": "1", "reasoning": "..."}`

Table 10: Expanded Prompt for COMET Trial (Part 2 of 2). Decision rules for all criteria are included as prompt context but only the rule matching the active criterion is applied.

Component	Content
System Instruction	You are an expert cardiologist evaluating patient eligibility for the FINALITY clinical trial. Based on the provided discharge note, determine if the patient meets the following *[criteria_type]* criterion.
Scoring System	**Score** **Interpretation** **1** Inclusion met OR exclusion not triggered **0** Inclusion not met OR exclusion triggered **2** Genuinely insufficient or ambiguous information — use **only** when evidence is present but unclear **3** Biologically not applicable — for example use **only** for sex-specific criteria for a documented male patient
Critical Rules	• **Never** use score 3 for "condition not found"; use 0 or 1 as appropriate. • **Never** use score 2 as a default when information is absent. Absent information means the condition is not present. Exception: use score 2 only when there is some mention but it is ambiguous. • Negation: "no history of X", "denies X", "no X" → treat X as absent. • Temporal: for time-windowed criteria, if an event appears in Past Medical History without a required time duration, assume it is outside the window. • Temporal: current admission event → falls within any time window.
Patient Metadata	Patient ID: *[patient_id]* Discharge Note: *[patient_id]*-DS-*[ds_number]* Gender: *[gender]* Matched Sections: *[matched_headlines]*
Patient Note	*[Relevant processed note content]*
Criterion (*[criteria_type]*)	*[Single inclusion or exclusion criterion — one per inference call]*
Criterion-Specific Decision Rules (Part 1)	*The full set of decision rules for this trial is included as context to guide scoring. Only the rule matching the active criterion above is applied per call.* • **Age ≥18 (inclusion)**: Always Score 1. All MIMIC-IV patients are adults. • **Symptomatic HFrEF (inclusion)**: Search Discharge Diagnosis, BHC, Pertinent Results for `HFrEF`, reduced EF, EF ≤40, plus symptoms (dyspnea, NYHA II–IV, orthopnea). Score 1: HFrEF and symptoms both confirmed. Score 0: HFpEF stated or no HF documented. Score 2: HFrEF present but symptomatic status genuinely unclear. • **Not on sMRA due to intolerance/contraindication/ineligibility (inclusion)**: Search Discharge Medications, PMH for `spironolactone`, `eplerenone`. Score 1: sMRA absent AND note states intolerance/contraindication/ineligibility. Score 0: sMRA currently prescribed. Score 2: sMRA absent but no reason documented. • **Negative pregnancy test and contraception, females only (inclusion)**: Score 3: documented male. Score 1: female with documented contraception, post-menopausal, or surgically sterile. Score 0: pregnant or no contraception documented. Score 2: female, contraception status unknown. • **Treatment with nsMRA finerenone/esaxerenone (exclusion)**: Search Discharge Medications, Medications on Admission. Score 1: none found. Score 0: any found. Score 2: medication list absent. • **eGFR <25 or potassium >5.0 (exclusion)**: Search Pertinent Results for most recent eGFR and potassium values. CKD Stage III → eGFR >30 → Score 1. CKD IV or ESRD → Score 0. Score 1: eGFR ≥25 AND K^+ ≤5.0. Score 0: either threshold breached. Score 2: labs absent. • **Type 1 AMI/revascularization/valve surgery/CRT within 30 days or planned (exclusion)**: Search Discharge Diagnosis, BHC, PMH for STEMI, NSTEMI, CABG, PCI, valve replacement, CRT. Current admission → Score 0. PMH without date → Score 1. "Planned" or "scheduled" → Score 0. Score 2: procedure present but timing entirely unclear.

Table 11: Expanded Prompt for FINALITY Trial (Part 1 of 2). The `Criterion` field receives a **single** inclusion or exclusion criterion per inference call; one JSON object is returned per criterion–patient pair.

Component	Content
Criterion-Specific Decision Rules (Part 2)	*Continued from Part 1. Only the rule matching the active criterion is applied per call.* • **Prior or planned heart transplant (exclusion)**: Search PMH, Discharge Diagnosis, BHC for `heart transplant`, `cardiac transplant`, `OHTx`. Score 1: none found. Score 0: documented. Score 2: no documentation at all. • **Severe uncorrected primary valvular disease as primary HF cause (exclusion)**: Search Discharge Diagnosis, BHC, Pertinent Results. Requires both: (a) severe or hemodynamically significant AND (b) explicitly primary HF cause. Prior surgical correction → Score 1. Score 1: no severe valvular disease as primary cause. Score 0: both requirements met. Score 2: severity or causation unclear. • **Symptomatic bradycardia or 2nd/3rd degree heart block without pacemaker (exclusion)**: Search Discharge Diagnosis, BHC, PMH. If none found → Score 1. If found, check for confirmed pacemaker/PPM/CRT → Score 1 (override). Block without pacemaker → Score 0. Block with ambiguous pacemaker status → Score 2. • **Excluded cardiomyopathies (exclusion)**: Search Discharge Diagnosis, PMH, BHC for myocarditis, amyloid, sarcoidosis, HOCM, congenital heart disease, constrictive pericarditis, Fabry, hemochromatosis, muscular dystrophy. Score 1: none found. Score 0: any documented. • **Probable alternative cause of HF (exclusion)**: Score 0 only if note explicitly attributes HF to alternative cause (e.g., Takotsubo, peripartum, chemotherapy-induced). Score 1: primary HFrEF without explicit alternative. Score 2: etiology entirely undocumented. • **Concomitant CYP3A4 inhibitors or inducers (exclusion)**: Search Discharge Medications only. Inhibitors: `ketoconazole`, `itraconazole`, `clarithromycin`, `ritonavir`, `posaconazole`, `voriconazole`. Inducers: `rifampin`, `phenytoin`, `carbamazepine`, `phenobarbital`. Discontinued or allergy → do not count. Score 1: none found. Score 0: any found. Score 2: medication list absent. • **Any other condition making patient unsuitable (exclusion)**: Search Discharge Diagnosis, BHC, PMH for metastatic cancer, hospice, palliative care, terminal illness, severe hepatic failure, severe pulmonary hypertension. Apply conservatively. Score 1: none found. Score 0: explicit disqualifying condition. Score 2: uncertain. • **Concurrent participation in another interventional study (exclusion)**: Search BHC, PMH, HPI for `clinical trial`, `investigational drug`, `study drug`. Score 1: none found. Score 0: active participation documented. Score 2: no information.
Output Format	Return **only** a valid JSON object. No text outside the JSON.
	Example: `{"score": "1", "reasoning": "..."}`

Table 11: Expanded Prompt for FINALITY Trial (Part 2 of 2). Decision rules for all criteria are included as prompt context but only the rule matching the active criterion is applied.

Table 12: Clinical Trial Criteria: CONVERGE-HF

Component	Content
Inclusion Criteria	
	• Adult patients
	• Established chronic heart failure ($\geq$ 6 months)
	• Mild-to-moderate cognitive impairment (as per the diagnosis of cognitive impairment or a Montreal Cognitive Assessment (MoCA) score 10–25)
Exclusion Criteria	
	• Patients who have contraindications for sGC stimulator and vericiguat therapy (i.e. use of long-acting nitrates, other soluble guanylate cyclase stimulators (e.g., riociguat), or phosphodiesterase type 5 (PDE-5), severe anemia, pregnancy or breast-feeding)
	• Unable to undergo CMR imaging or brain MRI
	• CMR exclusions: renal failure (a glomerular filtration rate $<$30 mL/min), implantable cardiac device (ICD or CRT), uncontrolled atrial fibrillation or recurrent ventricular arrhythmias
	• General medical conditions: uncontrolled thyroid disorders, hepatic failure, or myocardial revascularization procedures (coronary angioplasty and/or surgical revascularization in the previous 3 months), cancer/malignancy, or moderate-severe dementia
	• Patients with allergies to the study products

Table 13: Clinical Trial Criteria: COMET

Component	Content
Inclusion Criteria	
	• New York Heart Association [NYHA] class II–IV
	• At least one cardiovascular admission during the previous 2 years
	• On stable heart failure treatment with ACE inhibitors for at least 4 weeks unless contraindicated, and on treatment with diuretics ($\geq$ 40 mg of frusemide or equivalent) for at least 2 weeks.
	• Left-ventricular ejection fraction had to be 0·35 or lower measured within the previous 3 months by echocardiography or radionuclide ventriculography. If the ejection fraction was not determined, a left-ventricular end diastolic diameter of greater than 6–0 cm and a fractional shortening of less than 20% as measured by echocardiography could qualify patients..
Exclusion Criteria	
	• Major exclusion criteria were a recent change of treatment—defined as the introduction of a new class of drug for heart failure or treatment with oral β-adrenergic or α-adrenergic receptor blockers within the 2 weeks before randomisation—requirement for intravenous inotropic therapy, current treatment with calcium channel blockers (of the diltiazem or verapamil class), amiodarone ($>$ 200 mg per day) or class-I antiarrhythmic drugs, or administration of any investigational drug (specifically metoprolol or carvedilol) within the preceding 30 days.
	• Excluded patients with unstable angina, myocardial infarction (STEMI, NSTEMI, MI), or coronary revascularisation (CABG) or stroke within the previous 2 months.
	• Other exclusion criteria were uncontrolled hypertension (blood pressure systolic >170 mm Hg or diastolic >105 mm Hg).
	• Haemodynamically significant valvular disease
	• Symptomatic and sustained ventricular arrhythmias within the past 2 months not adequately treated with antiarrhythmic drugs or implantation of an automatic defibrillator.
	• Pregnancy, women with childbearing potential on inadequate contraception.
	• Known drug or alcohol misuse.
	• Poor compliance with treatment or any other serious systemic disease that might complicate management and reduce life expectancy.
	• Excluded patients if there was a contraindication to use of β-blockers, such as resting heart rate $<$ 60 beats per min, sick sinus syndrome, bifascicular block, second or third degree atrioventricular block unless treated with a pacemaker, sitting systolic blood pressure $<$ 85 mm Hg, history of asthma or chronic obstructive pulmonary disease (COPD), peripheral arterial disease (PAD) with symptoms at rest, or unstable insulin-dependent diabetes mellitus.

Table 14: Clinical Trial Criteria: FINALITY

Component	Content
Inclusion Criteria	
	• Age $\geq$ 18 years or legal age of majority
	• Symptomatic heart failure with reduced ejection fraction (HFrEF) per protocol-defined criteria
	• Not receiving steroidal mineralocorticoid receptor antagonists (sMRA) due to prior intolerance, contraindication, or ineligibility for treatment
	• Negative pregnancy test and agreement to use adequate contraception during the trial (female participants only)
Exclusion Criteria	
	• Treatment with non-steroidal mineralocorticoid receptor antagonists (nsMRA)
	• Estimated glomerular filtration rate (eGFR) $<$ 25 mL/min/1.73 m^2 and/or serum or plasma potassium $>$ 5.0 mmol/L
	• Type 1 acute myocardial infarction, coronary revascularization, valve replacement or repair, or implantation of a cardiac resynchronization therapy device within 30 days prior to enrollment or planned during the study
	• Prior or planned heart transplantation
	• Hemodynamically significant (severe) uncorrected primary cardiac valvular disease considered by the investigator to be the primary cause of heart failure
	• Symptomatic bradycardia or second- or third-degree heart block in the absence of a pacemaker
	• Cardiomyopathy due to known acute inflammatory heart disease, infiltrative diseases, accumulation diseases, muscular dystrophies, cardiomyopathy with reversible causes, hypertrophic obstructive cardiomyopathy, complex congenital heart disease, or pericardial constriction
	• Probable alternative cause of the participant's heart failure
	• Concomitant systemic therapy with potent cytochrome P450 isoenzyme 3A4 (CYP3A4) inhibitors or moderate to potent CYP3A4 inducers
	• Any other condition or therapy that would make the participant unsuitable for the study
	• Concurrent participation in another interventional clinical study using an investigational agent