Datasets for a Chatbot for Clinical Trial Search

Yumeng Yang, MS[1], **Ethan B Ludmir, MD**[2], **Kirk Roberts, PhD**[1]
[1]McWilliams School of Biomedical Informatics
The University of Texas Health Science Center at Houston, Houston, TX, USA,
[2]Department of Radiation Oncology
The University of Texas MD Anderson Cancer Center, Houston, TX, USA
Yumeng.Yang@uth.tmc.edu

Abstract

Matching patients to clinical trials is a critical bottleneck hindered by complex eligibility criteria. While conversational AI offers a promising solution, its safe deployment depends on high-quality, domain specific data. This paper introduces three benchmark datasets designed to support the development and evaluation of conversational agents for clinical trial pre-screening. First, a manually-annotated paired-criterion dataset provides a gold standard for structuring raw criteria, which we used to objectively group 12,596 criteria. Second, we curated a human-authored question benchmark to validate the clinical fidelity and patient-centric clarity of questions generated by a medical LLM, ensuring the AI's dialogue is accurate and understandable. Third, we constructed a human-validated assessment corpus of criterion-question-answer tuples with human-labeled outcomes to evaluate criterion classification based on a patient's answer to a generated question. The primary contribution of this work is a foundational set of benchmark datasets, designed to support and evaluate key components for a chatbot for clinical trial search.

Keywords: Eligibility Criteria, Natural Language Processing, Large Language Models, Conservational AI

1. Introduction

The matching of eligible patients to suitable clinical trials is a critical challenge in finding treatments for diseases such as cancer (Unger et al., 2016; Murthy et al., 2004). While clinical trials offer access to potentially life-saving innovative treatments, patient enrollment rates remain persistently low, with estimates suggesting that fewer than 5% of adult cancer patients participate (Unger et al., 2016; Murthy et al., 2004). A barrier to enrollment is the complexity of clinical trial eligibility criteria (Kim et al., 2017; Idnay et al., 2021), which are often articulated in dense, technical prose across disparate trial documents. This information overload makes it exceedingly difficult for patients and even clinicians to efficiently identify appropriate trials, leading to missed opportunities and delays in care.

The conventional approach to trial recruitment relies on manual eligibility screening by clinical staff, a process that is notoriously laborious and time-consuming (Cowie et al., 2017; Köpcke and Prokosch, 2014). Additionally, patients often lack effective tools to actively find potentially suitable clinical trials (Borno et al., 2022). Recent advances in natural language processing (NLP) and large language models (LLMs) present a promising avenue for mitigating this challenge. Conversational AI systems, in particular, have the potential to democratize access to trial information by transforming the complex pre-screening process into an intuitive, interactive dialogue. Such systems can guide patients through a series of clarifying questions to determine their potential eligibility. However, the efficacy, safety, and reliability of these clinical AI tools are not only a function of model architecture; they are fundamentally dependent on the quality of the underlying data resources used for their development and validation.

This paper addresses a gap between the potential of conversational AI in clinical trial matching and the practical, resource-intensive work required to build a trustworthy system. We argue that the development of robust language resources, accompanied by a rigorous, multi-stage evaluation protocol, is a prerequisite for a successful clinical application and constitutes a significant research contribution in itself. We present a detailed case study of our data preparation, annotation, and evaluation pipeline designed to support a conversational AI for cancer trial pre-screening. We specifically focus on cancer clinical trials, but none of our methods are limited to cancer; it is just the most prominent use case for clinical trials. Our work is organized around three data-centric stages: (1) clustering of semantically similar eligibility criteria, (2) generation of natural language questions for each criteria cluster, and (3) assessment of patient answers against the source criterion.

Further, we evaluate several state-of-the-art methods as baselines for creating and validating the linguistic assets at each stage. For criteria clustering, we systematically compare the performance of a density-based algorithm, HDBSCAN (Campello et al., 2013), against the centroid-based K-Means, establishing a more effective method for grouping nuanced clinical statements. For question generation, we develop a gold-standard dataset

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 139–148
12 May 2026.

of human-authored questions and use it to benchmark and iteratively refine questions generated by a specialized medical LLM, MedGemma (Sellergren et al., 2025). This iterative refinement process, guided by metrics such as BERTScore (Zhang et al., 2020) and ROUGE-L (Lin, 2004), resulted in a highly optimized prompt engineering strategy. Finally, we detail the creation of a human-validated assessment corpus to evaluate the LLM's ability to accurately map a user's answers to the underlying eligibility criterion, ensuring the final system's reasoning is clinically sound.

The primary contribution of this work is not the conversational agent itself, but the datasets that enable its construction. We envision a variety of conversational architectures are possible. However, the three sub-tasks studied how are fundamental to developing interactive systems for clinical trials. These tasks represent the basic building blocks of (i) reducing the dimensionality of clinical trial eligibility criteria, (ii) converting the criteria to a conversational prompt, and (iii) filtering trials based on the user's responses. In addition, this paper provides the initial prompts used to generate questions and assess patient responses against eligibility criteria, which can serve as a baseline for future refinement.

2. Related Work

Early efforts to automate patient-trial matching primarily relied on structured data and rule-based systems. Researchers developed systems that converted free-text eligibility criterion into structured representations using medical ontologies like UMLS and SNOMED-CT, enabling queries against patient data in Electronic Health Records (EHRs) (Weng et al., 2011; Yuan et al.). These approaches were often labor-intensive, requiring extensive manual effort to create and maintain rules, and struggled to capture the semantic nuance inherent in clinical prose (Weng et al., 2011). The emergence of machine learning has led to the development of methods include information extraction and text classification. These systems used classic Natural Language Processing (NLP) pipelines, including Named Entity Recognition (NER) and relation extraction, to identify clinical concepts within the trial criteria (Kang et al., 2017). However, these models were dependent on feature engineering and large, domain-specific annotated datasets, which are expensive and time-consuming to create. The paradigm shifted significantly with the development of deep learning-based contextual embeddings. Domain-specific pre-trained models such as ClinicalBERT (Alsentzer et al., 2019) and BioBERT (Lee et al., 2020), which are fine-tuned on clinical and biomedical corpora, demonstrated a superior ability to understand the semantics of medical text. These models have become the backbone for numerous downstream clinical tasks, including cohort identification and criteria analysis.

More recently, generative LLMs like Med-PaLM (Singhal et al., 2023) and Med-Gemini (Saab et al., 2024) have shown remarkable capabilities in complex clinical reasoning and natural language generation. Research has showcased their potential for tasks like medical question answering and clinical note summarization (Nori et al., 2023). However, a significant portion of this research focuses on end-task performance, often overlooking the critical, resource-intensive data preparation pipeline required for safe and reliable deployment. Our previous work includes manually annotating whether a cancer trial includes or excludes a key criterion in its protocol, which is important for downstream tasks (Yang et al., 2024a,b). This highlights a crucial gap between demonstrating a model's potential and engineering a trustworthy clinical application.

Our work's primary contribution is the methodology for building the data foundation for a clinical trial conversational agent. This involves three key NLP tasks: structuring the source criteria, generating patient-facing questions, and classifying the user's response. Clustering has been explored for discovering topics in clinical text, such as grouping patient notes or research articles (Arnold et al., 2016). However, to our knowledge, an evaluation of different embedding strategies and clustering algorithms specifically for the nuanced text of eligibility criteria has been lacking. Our analysis of SBERT vs. other embeddings and K-Means vs. HDBSCAN provides a methodological contribution to this niche area. Furthermore, while question generation is an established NLP task, its application to high-stakes clinical scenarios necessitates a focus on quality and fidelity. The standard practice of relying solely on automated metrics like ROUGE or BERTScore (Zhang et al., 2020) is insufficient where patient safety is concerned. This paper, therefore, outlines a replicable, human-in-loop methodology for generating and assessing conversational datasets to better support the clinical trial screening process.

3. Experimental Setup

This paper presents a multi-stage pipeline to construct and validate the linguistic resources for a conversational AI system for trial pre-screening. The methodology is partitioned into three stages: (1) the unsupervised clustering of eligibility criteria to distill recurring clinical concepts; (2) the generation and evaluation of natural language questions for each cluster; and (3) the development of a validated corpus to assess a large language model's ability to accurately interpret patient responses.

Medical Concepts (Cluster)	Raw Eligibility Criteria
Dihydropyrimidine Dehydrogenase (DPD) Deficiency	1. Known dihydropyrimidine dehydrogenase (DPD) deficiency. 2. Known DPD deficiency tested by measuring uracil level in blood or by genetic testing for DPD mutations.
Eastern Cooperative Oncology Group (ECOG) Performance Status ≤ 1	1. Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1. 2. Patients must have an ECOG performance status score of ≤ 1.

Table 1: Example eligibility criteria for Dihydropyrimidine Dehydrogenase (DPD) Deficiency and Eastern Cooperative Oncology Group (ECOG) Performance Status ≤ 1.

3.1. Clustering of Eligibility Criteria

The initial stage focused on structuring the raw criteria text into semantically coherent groups. To begin, a balanced dataset was curated by randomly sampling 100 clinical trials for each of five major cancer types—breast, lung, prostate, colorectal, and pancreatic cancer—resulting in a corpus of 500 trials. All inclusion and exclusion criteria were extracted from these trials, yielding 12,596 individual criteria. To prepare this text for clustering, each criterion was transformed into a high-dimensional vector using four distinct embedding strategies: the contextually-aware Sentence-BERT (SBERT) and Universal Sentence Encoder (USE), the domain-adapted ClinicalTrialBERT (Yang et al., 2024b), and a classical lexical baseline using Term Frequency–Inverse Document Frequency (TF-IDF) with dimensionality reduction via Singular Value Decomposition (SVD). These embedded representations were subsequently grouped using two complementary clustering algorithms: the centroid-based K-Means, configured with k=50 and k=300 to capture varying granularities, and the density-based HDBSCAN, tested under tight, moderate, and relaxed density parameterizations. This approach produced an ensemble of 20 unique clustering configurations for evaluation.

To identify the optimal clustering configuration, we manually annotated a set of 500 criterion pairs selected from the most frequent disagreements among our 30 model variants. These pairs were identified by calculating their co-clustering frequency and sampling from pairs that were grouped together in 5 to 15 of the 20 runs. Two annotators labeled these pairs based on whether they represented the same clinical concept and could be addressed by a single question, achieving a strong inter-annotator agreement with a Cohen's Kappa of 0.742, and agreement of 0.915.

Each of the 20 configurations was then benchmarked against this gold standard, with performance assessed using precision, recall, and F1. The results indicated that SBERT embeddings provided the most effective representations, achieving the highest average F1 (0.71) across all clustering methods. While HDBSCAN configurations consistently yielded high precision at the cost of lower recall, K-Means offered a more balanced performance. The best configuration, SBERT combined with K-Means (k=200), achieved a superior F1 of 0.84 (precision=0.94, recall=0.76). Consequently, this model was selected to generate the definitive criteria clusters for all subsequent stages of the pipeline. The detailed results for each embedding method are listed in Table 2.

While K-means was used to help build the gold standard, it is not included in our selection of clustering algorithms because it requires the number of clusters to be specified a priori, which depends heavily on the specific disease and number of trials. Instead, our clustering experiments for an algorithm to actually include in the chatbot relies on HDBSCAN, which determines the optimal number of clusters adaptively based on the intrinsic structure of the data. To determine the optimal hyperparameters for HDBSCAN clustering on SBERT-encoded eligibility criteria, we conducted a systematic grid search over a broad parameter space. The four key parameters controlling cluster granularity were: minimum cluster size (min_cluster_size $\in$ {5, 20, 50, 100}), minimum samples (min_samples $\in$ {1, 2, 3, 4, 5}), cluster selection epsilon ($\varepsilon \in$ {0.3, 0.5, 0.6, 0.7, 0.8, 0.9}), and alpha ($\alpha \in$ {0.5, 0.7, 0.9, 1.0}). For each configuration, clustering performance was quantified in terms of precision, recall, F1, and accuracy, where a pair was considered correctly predicted if both sentences appeared in the same cluster.

The grid search results demonstrated that clustering performance was relatively stable across a narrow range of high-performing configurations (F1 = 0.91). Configurations with smaller min_cluster_size and min_samples values tended to yield higher recall and finer-grained clusters, whereas larger values improved precision but produced coarser groupings. Reducing ε below 0.8 increased the total number of clusters but did not substantially affect F1 performance, suggesting that clustering quality was robust to moderate density threshold variations.

Based on this analysis, we selected the config-

Embedding	Avg Accuracy	Avg Precision	Avg Recall	Avg F1
ClinicalTrialBERT	0.5096	0.9182	0.4652	0.5668
SBERT	**0.6508**	**0.9436**	**0.6326**	**0.7131**
TFIDF	0.5384	0.9013	0.5132	0.5949
USE	0.5892	0.9182	0.5727	0.6505

Table 2: Evaluation of Embedding Models

uration of HDBSCAN with min_cluster_size = 5, min_samples = 1, $\varepsilon = 0.7$, and $\alpha = 0.9$. This configuration achieved a balanced trade-off (precision = 0.841, recall = 0.995, F1 = 0.911), while producing a reasonable number of clusters (n = 425). This balance provided both high intra-cluster homogeneity and sufficient granularity for downstream question generation and patient–trial matching tasks.

3.2. Question Generation and Evaluation

Following the identification of semantically coherent eligibility criteria clusters, the next stage focused on generating a representative, patient-facing question for each cluster. This benchmark dataset serves as a standardized foundation for evaluating conversational agents in clinical trial screening, ensuring consistency and reproducibility across different models and cancer types. By creating representative questions for each cluster, we capture common semantic patterns in eligibility criteria and translate them into language that patients can easily understand, thereby bridging the gap between clinical and layperson terminology.

Asking the right questions is essential for efficiently filtering ineligible trials and improving screening precision. For example, instead of vague, model-generated questions like “How would you describe your ability to carry out daily activities?”, it is more effective to ask directly about the patient’s ECOG performance status, which preserves clinical specificity and avoids ambiguity. For criteria such as “ECOG between 0–2,” the model was instructed to generate open-ended rather than binary questions, enabling downstream assessment modules to interpret responses more flexibly and accurately. This benchmark data set supports the evaluation of question generation models and provides a reusable framework to evaluate conversational AI systems in healthcare contexts.

We utilized a domain-specific large language model, MedGemma 27b, within a prompt engineering framework to generate these questions. The prompt’s design (shown in Table 3) was guided by several core principles established through iterative refinement. The primary goal was to generate a single, specific question whose answer would be directly classsifiable (e.g., as a number or a yes/no response). Key instructions emphasized preserving all medical specificity, directing the model to retain all technical terms, units, and test names (like ECOG, ALT, or g/dL) rather than generalizing them. Furthermore, the prompt explicitly guided the model to prioritize quantifiable answers—asking for a patient’s specific lab value or score instead of a simple binary confirmation—and to translate complex terms by adding brief, parenthetical clarifications (e.g., "ECOG (Eastern Cooperative Oncology Group)"). This structured approach, which also included rules for handling logical conditions (e.g., ’OR’ statements), was engineered to produce questions that were clinically precise, unambiguous for a layperson, and machine-readable for downstream assessment. To validate their quality, the machine-generated questions were benchmarked against a gold-standard set of human-made questions. The evaluation relied on BERTScore for measuring semantic similarity, thereby quantifying the fidelity of the generated questions to the human references. We manually selected 100 high-quality clusters for annotation, which were defined by two criteria: first, the cluster had a clear medical theme, with most criteria pertaining to the same concept (e.g., Eastern Cooperative Oncology Group (ECOG)); and second, the majority of criteria within a cluster could be consolidated and addressed by a single question, despite being expressed differently.

The benchmark dataset was created by two annotators who made questions for each of the 100 clusters. We observed that some clusters contain multiple themes; in these cases, we generated multiple questions to ensure complete coverage. Following a consolidation phase between the two annotators, the final corpus served as our gold-standard benchmark. For the experiment, the model was constrained to generate a single question per cluster, which was then evaluated for semantic fidelity against the human-made references for that cluster using BERTScore. The evaluation yielded a mean precision of 0.3256, mean recall of 0.5278, and a mean F1 of 0.4234, indicating a moderate level of semantic alignment between the model-generated and human-authored questions. These results suggest that while the model captures key aspects of the reference questions, there remains room for improvement in fully mirroring human intent and phrasing. To better illustrate the effectiveness of our approach, several representative examples from the benchmark are shown below. For instance, within the functional status cluster, the criterion “ECOG between 0–2” led to the question

Prompt Type	Prompt Content
Question Generation Prompt	GOAL: Analyze each clinical trial eligibility criterion and generate a single, specific question whose answer can be directly used to determine eligibility (classifiable as yes/no or numeric). CORE PRINCIPLES: 1. Preserve all medical specificity – keep every named medical entity, test, score, or unit (e.g., ECOG, KPS, WHO, albumin, hemoglobin, ALT, AST, creatinine). Do not generalize or replace these with vague phrases. 2. Prioritize quantifiable or discrete answers – if the criterion includes a number, grade, or threshold (e.g., ECOG ≤ 1, Albumin ≥ 3.0 g/dL), ask for that number explicitly and encourage patients to find it in their medical record. 3. Translate, don't simplify away meaning – keep technical terms and add short clarifications when needed (e.g., ECOG performance status $\rightarrow$ ECOG (Eastern Cooperative Oncology Group) performance status). 4. Handle logical conditions explicitly – include both sides of logical conditions (e.g., bleeding disorders OR taking anticoagulants) in the question. 5. Be clinically actionable – phrase questions so clinicians can use answers directly for eligibility screening. EXAMPLES: • ECOG performance status of 0 or 1 $\rightarrow$ Q: What is your most recent ECOG (Eastern Cooperative Oncology Group) performance status score? It is a number from 0 to 5 in your oncology record. • Serum albumin ≥ 3.0 g/dL $\rightarrow$ Q: What is your most recent serum albumin level, as listed in your blood test results? • History of bleeding disorders OR current use of blood thinners $\rightarrow$ Q: Do you have a bleeding disorder or are you currently taking blood-thinning medication such as warfarin or apixaban? OUTPUT FORMAT: Generated Question: "[your question here]" Criterion: {rep_criteria}
Assessment Prompt	You are assessing patient eligibility against a clinical trial criterion. Inputs: Patient Answer: "{patient_answer}" Question: "{question}" Criterion Type: {criterion_type} Criterion: {criterion_text} Instructions: • If INCLUSION criterion: patient must meet this requirement. • If EXCLUSION criterion: patient must not meet this requirement. Decision options (respond with exactly one): • INCLUDE – patient meets inclusion or does not meet exclusion. • EXCLUDE – patient does not meet inclusion or meets exclusion. • UNKNOWN – insufficient data to determine. Output: provide only one decision word (INCLUDE / EXCLUDE / UNKNOWN), nothing else.

Table 3: Prompts for question generation and assessment

"What is your ECOG score?", which promotes more flexible downstream assessment than a binary version. Similarly, in the hematologic function cluster (e.g., "Absolute neutrophil count $\geq$ 1500/mm^3"), the generated question "Do you know your most recent blood test results, including neutrophil or platelet counts?" successfully balances clinical specificity and patient readability. Finally, for organ function and cancer stage clusters, questions such as "Have you been told that your kidney function is normal?" and "What type and stage of cancer were you diagnosed with?" demonstrate how technical eligibility language can be converted into patient-centered prompts without loss of medical meaning.

3.3. Patient Answer Assessment and Validation

The third stage established a framework to evaluate the conversational agent's ability to assess patient answers and make eligibility decisions. This benchmark dataset enables end-to-end assessment of the clinical trial screening pipeline—from question generation to response interpretation and eligibility assessment.

To establish a framework for evaluating conversational eligibility assessment, we developed a corpus of 13,344 labeled tuples, each comprising a criterion, a corresponding patient-facing question, and responses (answers). The dataset integrates two distinct components: a machine-generated set of 6,660 examples, where the model was prompted to produce two opposing synthetic answers per criterion, and a human-generated answers with set of 6,684 examples, where real users provided up to three nuanced responses per question. To ensure these answers reflected real-world screening scenarios, we intentionally included a spectrum of response types. Beyond straightforward "yes" and "no" answers, the dataset incorporates conditional responses (e.g., "I can stop my medication if required") and "incomplete" entries where patients cannot recall specific dates or lab values. This design forces the model to move beyond binary logic and handle cases requiring follow-up or manual review.

Evaluating the MedGemma 27B model against this human-labeled gold standard revealed a three-class accuracy of 74.1% and a binary (non-exclusion vs. exclusion) accuracy of 81.3%. The model demonstrated reliable recognition of the INCLUDE class (precision = 0.83, recall = 0.72, F1 = 0.77) and the EXCLUDE class (precision = 0.69, recall = 0.82, F1 = 0.75). However, the UNKNOWN class exhibited a strong dependence on label prevalence; in the machine-generated subset where UNKNOWN frequency was low (5.7%), the model struggled significantly (F1 = 0.32), whereas performance improved in the human-annotated subset (19.8% prevalence) to an F1 of 0.68. Across the aggregate dataset, the UNKNOWN class achieved an overall F1 of 0.61. These results indicate that while the model effectively identifies explicit affirmations and denials, its ability to defer for clarification is sensitive to data distribution and the inherent ambiguity of natural patient language.

The failure cases listed in Table 5 underscore the critical need for our assessment dataset. While large language models demonstrate powerful general reasoning capabilities, their assessments varied when faced with domain-specific challenges such as unit conversion and contextual interpretation. This highlights that careful, targeted evaluation is essential for specialized applications. Our dataset, therefore, provides a crucial benchmark for identifying these gaps and driving the development of more robust and reliable conversational AI for assessing clinical trial eligibility.

4. Results

Our multi-stage data generation pipeline produced three validated benchmark datasets designed to support and evaluate the development of conversational AI for clinical trial screening. The first dataset is a manually annotated gold standard of paired eligibility criteria, which played a critical role in objectively determining the optimal clustering configuration. By evaluating 20 different configurations against this benchmark, we identified Sentence-BERT (SBERT) as the superior embedding model, achieving the highest average F1-score of 0.7131. The final optimized model, combining SBERT with HDBSCAN, demonstrated high performance with an F1 of 0.911 (precision = 0.841, recall = 0.995) on our benchmark. The application of this validated model successfully structured 12,596 raw eligibility criteria into a set of 425 clusters.

The second benchmark data is the human-made question dataset corresponding to 100 criteria clusters. To ensure comprehensive coverage, our annotators manually crafted multiple valid, patient-facing questions for clusters that include more than one clinical theme. This design provides a flexible evaluation target; an automated system is considered successful if it generates a question semantically similar to any of the human-made variants for a given cluster. We used this benchmark to evaluate the questions generated by the MedGemma model, using BERTScore to measure semantic fidelity. The evaluation yielded a mean precision of 0.3256, mean recall of 0.5278, and a mean F1 of 0.4234. This quantitative result indicates a moderate level of semantic alignment, confirming that while the model captures key aspects of the clinical intent (as shown in Table 4), there remains signifi-

Cluster Theme	Human-generated Question	Machine-generated Question
Functional Status (ECOG)	What is your Eastern Cooperative Oncology Group (ECOG) score?	What is your most recent Eastern Cooperative Oncology Group (ECOG) score?
Liver Function (AST)	Any recent blood tests or scans your doctor or nurse mentioned?	What are your most recent AST (aspartate aminotransferase) or ALT (alanine aminotransferase) levels?
Renal Function (Urine Protein)	Could you describe your most recent urine test results?	What is your most recent urine protein level or urinalysis result?
Infectious Disease Status (HIV)	Can you tell me if you have HIV, and whether you are currently receiving treatment?	Do you have a diagnosis of Human Immunodeficiency Virus (HIV) infection, and if so, are you on treatment?

Table 4: Example comparison of human-generated and machine-generated patient-facing questions.

cant room for improvement in fully mirroring human phrasing and nuance. This benchmark, therefore, serves its purpose as a critical tool for both validating and iteratively refining generative models for patient-facing communication.

The final benchmark is an assessment corpus designed to evaluate the end-to-end reasoning and classification capabilities of a large language model. This dataset is structured as a collection of tuples: criterion, criterion type, question, user's answer, comprising 13,344 labeled examples drawn from 1,366 unique clinical trials. Using this benchmark, we quantitatively evaluated the model's performance against a human-labeled gold standard (Table 6). Across the full evaluation set, the base MedGemma 27B model achieved a three-class accuracy of 74.1% and a binary (non-exclusion vs. exclusion) accuracy of 81.3%, the latter reflecting the model's tendency to err on the side of caution when a definitive exclusion decision cannot be reliably supported.

The results revealed several systematic patterns. The model performed most reliably on explicit, unambiguous eligibility cases: the INCLUDE class achieved precision 0.83, recall 0.72, and F1 0.77, while EXCLUDE reached precision 0.69, recall 0.82, and F1 0.75. However, performance degraded substantially when faced with ambiguity. The UNKNOWN class represented the key failure point, with precision 0.58, recall 0.64, and F1 0.61 overall—and dropping as low as F1 0.32 on the original dataset where such cases are underrepresented (5.7% of labels), indicating that the model frequently defaulted to a premature binary decision rather than deferring for clarification. This failure mode aligns closely with the qualitative limitations identified in Table 5, including an inability to perform unit conversions for laboratory values (e.g., mmol/L to mg/dL), difficulty interpreting non-clinical patient language (e.g., "within normal range"), and challenges in resolving semantic ambiguity in conditional or temporally qualified statements. This benchmark not only quantifies these predictable weaknesses but also establishes a validated resource for future research, enabling targeted fine-tuning to enhance the clinical assessment capabilities of advanced language models.

5. Discussion

In this paper, we introduce three benchmark datasets designed to advance clinical trial prescreening research. The paired-criterion benchmark establishes a structured and replicable framework for clustering selection and evaluation. The human-made question benchmark supports the assessment of patient-facing generative AI by providing diverse phrasings of clinical concepts, enabling more nuanced evaluation of question-generation models. The assessment corpus focuses on identifying reasoning challenges such as unit conversion, contextual understanding, and semantic ambiguity. Together, these datasets offer evaluation resources that can inform the development and coordination of future multi-agent chatbots for clinical trial search and recommendation.

Despite these contributions, there are some limitations in our study too. First, our datasets are derived exclusively from five major cancer types. While our data-centric method is designed to be generalizable, the resulting resources are domain-specific. Applying and validating the pipeline in other therapeutic areas will be important to confirm its generalizability. Second, the patient answers in our assessment corpus were synthetically generated by LLM. Although designed to cover diverse response types, they cannot fully capture the complexity, ambiguity, and unpredictability of real-world patient language. Future work include augmentation this corpus with authentic patient dialogues to ensure real-world robustness. These limitations underscore the necessity of expanding data collection

Type of Limitation	Example Question and Answer	Corresponding Criterion	Model Issue / Outcome
Interpretation and Context Understanding	**Q:** What is your recent blood test that checked how your blood clots, like INR, PT, or aPTT? Were the results normal, and are you currently taking any blood-thinning medications such as warfarin or heparin? **A:** My blood results are within normal range.	International normalized ratio (INR) and activated partial thromboplastin time (aPTT) or partial thromboplastin time (PTT) $>$ 1.5 × ULN (unless on therapeutic coagulation).	Model fails to infer that "within normal range" implies eligibility and should include all related trials.
Ambiguity in Expression	**Q:** What is your usual resting heart rate (in beats per minute)? **A:** My usual resting heart rate is around 72 bpm, though it can increase if I'm stressed or just woken up.	Mean resting heart rate 50–90 bpm (determined from ECG); indication of treatment with hormone therapy and hypofractionated radiotherapy.	Model cannot equate "usual resting heart rate" with "mean resting heart rate," marking it as "need more information."
Unit Conversion	**Q:** What were your most recent fasting cholesterol and triglyceride levels (NCT01301911)? **A:** Total cholesterol: 185 mg/dL; triglycerides: 110 mg/dL.	Cholesterol 7.75 mmol/L and triglycerides 2.5 × ULN.	Model unable to convert mmol/L to mg/dL (7.75 mmol/L 140 mg/dL), incorrectly marking as "need more information."
Incomplete Criterion	—	CNS inclusion criteria: [incomplete statement].	Model continues to assess instead of marking as "need more information."

Table 5: Examples of Model Limitations in Question–Answer–Criterion Assessment

Class	Precision	Recall	F1-score
INCLUDE	0.83	0.72	0.77
EXCLUDE	0.69	0.82	0.75
UNKNOWN	0.58	0.64	0.61

Table 6: Evaluation Performance on Eligibility Assessment

across different diseases and validating our data against real-world interactions.

6. Conclusion

In this paper, we introduced three benchmark datasets that provide the evaluation resources for developing and validating conversational AI in the clinical trial domain. We presented: (1) a paired-criterion dataset for objectively optimizing the clustering of complex eligibility criteria; (2) a human labeled question dataset for the evaluation of patient-facing language generation; and (3) an assessment corpus that quantifies critical reasoning gaps in current LLMs and enables targeted model improvement, also for further fine-tuning. Ultimately, this work demonstrates that progress in clinical AI is dependent not only on the advancement of the model architecture, but also on the careful curation of high-quality data. These datasets represent a tangible step toward building safer and more reliable patient-centric tools to navigate the complexities of clinical trial recruitment.

7. Acknowledgments

Research reported in this publication was supported by the U.S. National Library of Medicine of the National Institutes of Health under award numbers R01LM014508 and R01LM011934. Yumeng Yang is a predoctoral fellow supported by the Cancer Prevention and Research Institute of Texas (CPRIT).

8. Bibliographical References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Corey W Arnold, Andrea Oh, Shawn Chen, and William Speier. 2016. Evaluating topic model interpretability from a primary care physician perspective. *Computer methods and programs in biomedicine*, 124:67–75.

Hala T. Borno, Li Zhang, Sylvia Zhang, Celia Kaplan, Alexander Bell, Brian Bakke, Amy Lin, Rahul Aggarwal, and Eric J. Small. 2022. Mobile Clinical Trial Matching Technology in Medical Oncology Clinic: A Pilot Feasibility Study. *JCO Clinical Cancer Informatics*, 6:e2100182.

Ricardo J.G.B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer.

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

Betina Idnay, Caitlin Dreisbach, Chunhua Weng, and Rebecca Schnall. 2021. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *Journal of the American Medical Informatics Association*, 29(1):197–206.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. 2017. EliIE: An open-source information extraction system for clinical trial eligibility criteria. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Edward S. Kim, Suanna S. Bruinooge, Samantha Roberts, Gwynn Ison, Nancy U. Lin, Lia Gore, Thomas S. Uldrick, Stuart M. Lichtman, Nancy Roach, and Julia A. Beaver, et al. 2017. Broadening Eligibility Criteria to Make Clinical Trials More Representative: American Society of Clinical Oncology and Friends of Cancer Research Joint Research Statement. *Journal of Clinical Oncology*, 35(33):3737–3744.

Felix Köpcke and Hans-Ulrich Prokosch. 2014. Employing Computers for the Recruitment into Clinical Trials: A Comprehensive Systematic Review. *Journal of Medical Internet Research*, 16(7):e161.

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

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81.

Vivek H Murthy, Harlan M Krumholz, and Cary P Gross. 2004. Participation in Cancer Clinical Trials: Race-, Sex-, and Age-Based Disparities. *JAMA*, 291(22):2720–2726.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicoló Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, and Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, and Elahe Vedadi, et al. 2024. Capabilities of Gemini Models in Medicine. *arXiv preprint arXiv:2404.18416*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, and Charles Lau, et al. 2025. MedGemma Technical Report. *arXiv*, 2507.05201.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, and Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620.

Joseph M Unger, Elise Cook, Eric Tai, and Archie Bleyer. 2016. *American Society of Clinical Oncology Educational Book*, volume 36, chapter The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. ASCO.

Chunhua Weng, Xiaoying Wu, Zhihui Luo, Mary Regina Boland, Dimitri Theodoratos, and Stephen B Johnson. 2011. EliXR: a system for extracting and encoding eligibility criteria in clinical trials. *Journal of the American Medical Informatics Association*, 18(Supplement 1).

Yumeng Yang, Ashley Gilliam, Ethan B Ludmir, and Kirk Roberts. 2024a. Exploring the Generalization of Cancer Clinical Trial Eligibility Classifiers across Diseases. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Yumeng Yang, Soumya Jayaraj, Ethan Ludmir, and Kirk Roberts. 2024b. Text Classification of Cancer Clinical Trial Eligibility Criteria. In *Proceedings of the AMIA Annual Symposium*.

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, and Chunhua Weng. Criteria2Query: a natural language interface to clinical databases for cohort definition.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations*.