Beyond the Registry: A Protocol-Derived Dataset and Embedding-Guided Method for Complete Clinical Trial Eligibility Criteria Generation

Anonymous ACL submission

Abstract

Eligibility criteria (EC) define who can participate in a clinical trial and are central to trial design and recruitment. Existing EC generation studies rely on ClinicalTrials.gov registry entries, which are compressed summaries that fail to capture the EC for real trial practice. We introduce PROTOCOLEC, a protocol-derived dataset of 4,302 Phase III trials pairing registry metadata with complete EC extracted from trial protocols spanning 22 therapeutic areas. PROTOCOLEC contains roughly twice as many criteria and words as registry EC, and a blinded three-judge audit confirms that they fully cover registry EC in 91.3% of trials and add criteria in 66.5%. We further benchmark zero-shot and fine-tuned LLMs, biomedical models, and a prior state-of-the-art EC generation method, and propose Q-ECGEN, a simple yet effective adaptation of a Q-Former-style connector that maps BioBERT representations into soft prefix tokens for a LoRA-tuned decoder. Q-ECGEN achieves the best performance across all evaluated metrics, improving over the prior SOTA by 0.08 ROUGE-L precision, 0.04 ROUGE-L F1, 11.4 BLEU, and 5.6 chrF.

1 Introduction

Clinical trials are essential for evaluating new therapies, yet many fail to meet their enrollment targets. Accrual failures are substantial as they increase trial cost, delay therapeutic advances, waste clinical research resources, and limit the generalizability of findings. Eligibility criteria (EC) are a key determinant of patient accrual, as they specify which patients can participate. In an analysis of closed Phase 2 and 3 interventional trials on ClinicalTrials.gov, 19% either terminated due to unsuccessful accrual or completed with fewer than 85% of their target enrollment, and a greater number of EC was significantly associated with unsuccessful accrual (Carlisle et al., 2015). More recent work likewise reports that trials with the most complex EC have substantially higher accrual-failure rates than those with the least complex (Peterson et al., 2023).

Designing EC is challenging because it requires balancing competing goals: criteria must be specific enough to ensure participant safety and scientific validity, but not so restrictive that they slow recruitment, exclude representative patients, or undermine generalizability (Kim et al., 2017, 2021; Peterson et al., 2023). EC design is also time-consuming and expertise-intensive, demanding clinical judgment, knowledge of precedent trials, regulatory awareness, and disease-specific expertise. These challenges have motivated sustained efforts to modernize eligibility across domains such as prior therapies, laboratory thresholds, and performance status, and they underscore the need for computational tools that can support EC design (Jin et al., 2017; Kim et al., 2021; Harvey et al., 2021; Osarogiagbon et al., 2021; Spira et al., 2021; Magnuson et al., 2021).

Computational support for EC design has therefore become an important NLP problem. Earlier work largely focused on analyzing, extracting, and normalizing existing criteria for cohort identification, trial matching, and EC recommendation to inform EC design (Yuan et al., 2019; Datta et al., 2024; Wang and Sun, 2022; Kim et al., 2024), while more recent systems generate EC directly from trial descriptions or structured trial features (Wang et al., 2023; Lekuthai et al., 2025; Chen et al., 2025). Our study focuses on the latter. While EC analysis studies draw on complete EC from trial protocols to enable faithful clinical analysis (Liu et al., 2021), EC recommendation, ranking, and generation studies are built from ClinicalTrials.gov registry EC, which are concise public summaries of the protocol-level criteria. As a result, those systems built on registry EC cannot reflect the EC that actually govern enrollment in real trials, undermining the fundamental purpose of this line of work: to support trial

design as it is practiced.

To our knowledge, no large-scale protocol-derived dataset of complete EC exists. To close this gap, we introduce PROTOCOLEC, a protocol-derived benchmark for complete clinical trial EC generation. Starting from all completed Phase III trials with available protocol PDFs, we construct 4,302 paired trials spanning 22 therapeutic areas that link registry metadata and registry EC to protocol-extracted complete EC. Quantitative analysis shows that protocol-derived EC contain approximately twice as many criteria per trial as registry EC. A blinded three-judge audit further confirms that protocol-extracted EC fully cover registry EC in 91.3% of trials and add criteria in 66.5%.

Building on this benchmark, we further propose Q-ECGEN, an embedding-guided generation method for complete EC generation. Q-ECGEN adopts a Q-Former-style query connector (Li et al., 2023) to transform frozen BioBERT representations (Lee et al., 2020) of the target trial features and a retrieved reference trial EC into soft prefix tokens for a LoRA-tuned decoder (Hu et al., 2022). We benchmark Q-ECGEN against zero-shot LLMs, supervised fine-tuned models, biomedical models, and EC-RAFT (Lekuthai et al., 2025), a current state-of-the-art retrieval-augmented fine-tuning method for EC generation. As a result, Q-ECGEN achieves the best performance across all evaluated lexical and semantic metrics and surpasses the current best method. An ablation study confirms that both the Q-Former connector and retrieved reference EC contribute to this gain, with the full model exceeding either component in isolation. Gains are also consistent across the ten largest therapeutic areas, and sensitivity analysis shows stable performance across connector configurations.

Our contributions are threefold:

- We introduce PROTOCOLEC, to our knowledge the first large-scale protocol-derived benchmark for complete clinical trial EC generation.
- We provide quantitative and qualitative validation showing that PROTOCOLEC is substantially more complete than registry EC while largely preserving registry information.
- We propose Q-ECGEN, a simple yet effective adaptation of a Q-Former-style connector for complete EC generation, and demonstrate consistent improvements over all baseline methods.

2 Related Work

Clinical-trial EC datasets and benchmarks. Several recent efforts curate clinical-trial datasets to support data-driven trial research. HINT and CTO benchmark trial-outcome prediction from registry features (Fu et al., 2022; Gao et al., 2024); TrialBench covers eight clinical-trial prediction tasks, including registry-derived EC design (Chen et al., 2025); and Panacea's TrialAlign and TrialInstruct provide instruction data for trial search, summarization, design, and matching (Lin et al., 2024). These resources are primarily built from ClinicalTrials.gov registry records, which limits EC supervision to compressed public summaries. In contrast, Liu et al. (2021) use complete EC from full trial protocols, manually curating protocol-level criteria from a small set of oncology trials for real-world-data trial emulation. This work demonstrates the clinical value of protocol-level EC, but does not construct a large-scale generation benchmark. PROTOCOLEC addresses this gap by pairing registry artifacts with protocol-extracted complete EC across 4,302 Phase III trials.

EC analysis, structuring, ranking, recommendation, and patient–trial matching. Beyond generation, several adjacent research threads inform EC design without producing complete EC as output. EC extraction and structuring methods transform free-text criteria into computable representations for cohort identification and downstream querying (Kury et al., 2020; Dobbins et al., 2022; Yuan et al., 2019; Park et al., 2024; Datta et al., 2024). EC analysis studies characterize design patterns and complexity at scale, including clustering-based summarization of registry EC (Bornet et al., 2025) and quantitative analyses linking EC complexity to accrual (Peterson et al., 2023). EC recommendation and ranking methods retrieve or rank candidate criteria for a target trial (Kim et al., 2024; Wang and Sun, 2022), while patient–trial matching systems use EC as constraints to assess patient eligibility and rank relevant trials (Jin et al., 2024; Wornow et al., 2025; Nievas et al., 2024). These lines of work share infrastructure with our setting, including registry-derived EC text, biomedical encoders, and retrieval, but treat EC as inputs to extraction,

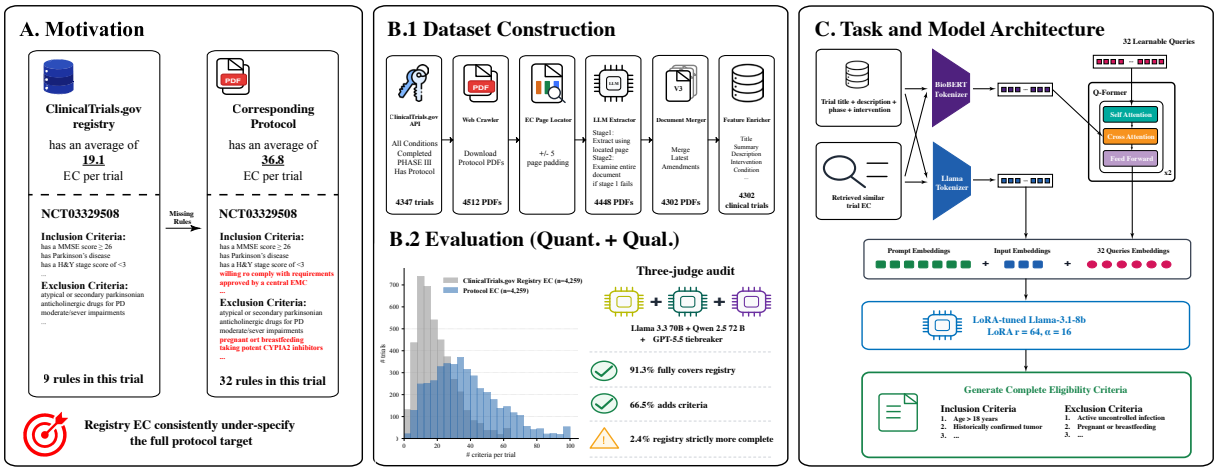

Figure 1: Overview of PROTOCOLEC and Q-ECGEN. **A.** ClinicalTrials.gov registry EC are concise summaries that can omit protocol-level eligibility rules. **B.** PROTOCOLEC is constructed by retrieving Phase III trial records, downloading protocol PDFs, localizing EC sections, extracting protocol-derived EC, merging protocol versions, and validating the resulting trial pairs. **C.** Q-ECGEN combines textual prompting with BioBERT-guided soft prefix tokens produced by a Q-Former-style connector and generates complete EC with a LoRA-tuned decoder.

ranking, or matching rather than as the target output of generation.

Eligibility-criteria generation. This work focuses on EC generation, where the target output is a complete inclusion and exclusion criteria set for a new trial. AutoTrial introduces controllable EC generation through discrete and neural prompting, retrieval over precedent trials, and explicit reasoning chains, demonstrating that LLMs can generate fluent and clinically meaningful EC when guided by instructions and similar trials (Wang et al., 2023). EC-RAFT advances this direction with retrieval-augmented fine-tuning: it retrieves related trials, synthesizes intermediate reasoning traces, and fine-tunes Llama-3.1-8B-Instruct to generate structured EC from trial titles and descriptions (Lekuthai et al., 2025). In our experiments, EC-RAFT is the strongest published EC generation baseline. Beyond task-specific systems, recent multi-task benchmarks and foundation models also include EC generation: TrialBench evaluates generic LLMs zero-shot on registry-derived EC design (Chen et al., 2025), and Panacea is fine-tuned on registry trial data and evaluated across multiple clinical-trial tasks, including EC design (Lin et al., 2024). These prior systems are trained or evaluated against registry EC and are therefore not designed to recover the protocol-level requirements that distinguish complete EC from public registry summaries. Our work targets this gap by introducing a protocol-derived benchmark and an embedding-guided generation framework, Q-ECGEN, which adapts a Q-Former-style query connector to project frozen biomedical encoder representations into soft prefix tokens for a LoRA-tuned decoder.

3 Methods

3.1 ProtocolEC Dataset Construction

Figure 1B summarizes the ProtocolEC construction pipeline. We construct ProtocolEC from completed Phase III clinical trials across all conditions on ClinicalTrials.gov that have at least one attached protocol PDF. Each example pairs a trial's registry metadata and registry EC with a more complete set of EC extracted from the corresponding protocol document. Full prompt templates and pipeline details are provided in Appendix A.

Trial selection and protocol acquisition. We retrieved candidate trials via the ClinicalTrials.gov v2 REST API, restricted to Phase III trials with overall status *Completed* and at least one attached protocol document (U.S. National Library of Medicine, 2024). For each qualifying trial we fetched the registry record (title, descriptions, conditions, interventions, outcomes, and registry EC, among other fields) and downloaded all attached protocol PDFs. The raw corpus contains 4,347 trials and 4,512 protocol PDFs, with some trials contributing multiple documents (i.e. original protocol and amendments).

Eligibility-section localization. Protocol PDFs vary substantially in length, formatting, and section

Table 1: Dataset statistics across train/validation/test splits. Each split contains paired ClinicalTrials.gov registry EC and protocol-extracted complete EC. Values are reported as mean $\pm$ standard deviation per trial. Criteria counts denote top-level bulleted eligibility criteria; words are whitespace-separated tokens in the raw EC text.

Statistic	Train (N=3,443)		Validation (N=429)		Test (N=430)	
	Registry	Protocol	Registry	Protocol	Registry	Protocol
Mean inclusion criteria / trial	7.3 ± 6.2	13.5 ± 10.2	7.0 ± 5.8	13.5 ± 9.7	6.9 ± 5.8	13.1 ± 9.6
Mean exclusion criteria / trial	11.9 ± 9.4	23.3 ± 14.3	11.8 ± 9.5	24.3 ± 14.1	11.4 ± 8.5	22.6 ± 13.4
Mean total criteria / trial	19.2 ± 12.7	36.8 ± 21.1	18.8 ± 12.5	37.8 ± 21.0	18.3 ± 11.3	35.8 ± 19.5
Mean words / trial	407.2 ± 321.1	834.4 ± 552.0	389.9 ± 312.9	872.1 ± 559.5	389.3 ± 303.8	820.2 ± 505.2

Table 2: Therapeutic-area distribution of the final PROTOCOLEC dataset (N=4,302 trials).

Therapeutic area	N	%
Oncology	592	13.76
Infectious Disease	571	13.27
Neurology	391	9.09
Dermatology	335	7.79
Endocrinology	331	7.69
Hematology	266	6.18
Pulmonology	254	5.90
Psychiatry	234	5.44
Ophthalmology	217	5.04
Cardiology	204	4.74
Gastroenterology	168	3.91
Rheumatology	165	3.84
Obstetrics and Gynecology	120	2.79
Anesthesiology	91	2.12
Nephrology	83	1.93
Allergy and Immunology	82	1.91
Urology	61	1.42
Orthopedics	50	1.16
Pediatrics	32	0.74
Dentistry	28	0.65
Otolaryngology	26	0.60
Radiology	1	0.02
Total	**4,302**	**100.00**

organization, so we first localize candidate EC sections using a deterministic three-stage locator. The locator searches, in order, (i) PDF bookmarks and outline entries matching inclusion/exclusion patterns, (ii) rendered-text headings such as "Inclusion Criteria" paired with nearby exclusion headings, and (iii) table-of-contents pages mapping printed page numbers to PDF page indices. Each candidate slice is verified using an EC-signal heuristic that checks for criterion-like cues (bullet lists, eligibility modal verbs, age expressions, laboratory thresholds, inclusion/exclusion terminology). The locator falls through to the next strategy if a candidate slice fails the heuristic. Full stage specifications and the EC-signal validation rules are in Appendix A.1.

LLM-based EC extraction. For each localized protocol document, we construct a padded EC slice by taking the detected EC pages plus five pages on each side, accounting for criteria that begin before the detected heading or continue into adjacent pages. The padded slice is rendered as a fresh PDF and passed to a GPT-5.4-based LLM extractor with instructions to produce a verbatim two-section transcription ("Inclusion Criteria," "Exclusion Criteria"); where paraphrasing and summarization are prohibited. If slice-based extraction fails, for example, malformed output, context overflow, or missing structure, we fall back to full-PDF extraction. The full extraction prompt is provided in Appendix A.2.

Multi-document merging and metadata enrichment. For trials with multiple protocol-related PDFs (e.g., amendments), we run the extractor independently on each document and merge the outputs at the trial level where the latest available amendment is retained as the primary protocol-derived target. Each paired example is enriched with registry metadata used for generation and stratified analysis, and assigned a primary therapeutic area via an LLM classifier over the trial's condition list, title, and brief summary (full classifier prompt in Appendix A.3).

Train, validation, and test splits. After extraction and merging, ProtocolEC contains 4,302 paired trials. We split 80/10/10 with a fixed random seed, stratified by primary therapeutic area, yielding 3,443 training, 429 validation, and 430 test examples.

3.2 Task Formulation

Given a clinical trial input x (title, summary, phase, intervention, and condition), the task is to generate complete protocol-level eligibility criteria $\mathbf{y}$, structured as inclusion and exclusion criteria. Given input, we retrieve a similar trial from the training set and use its EC r as reference context. The model is trained to estimate

$$p(\mathbf{y} \mid x, r). \tag{1}$$

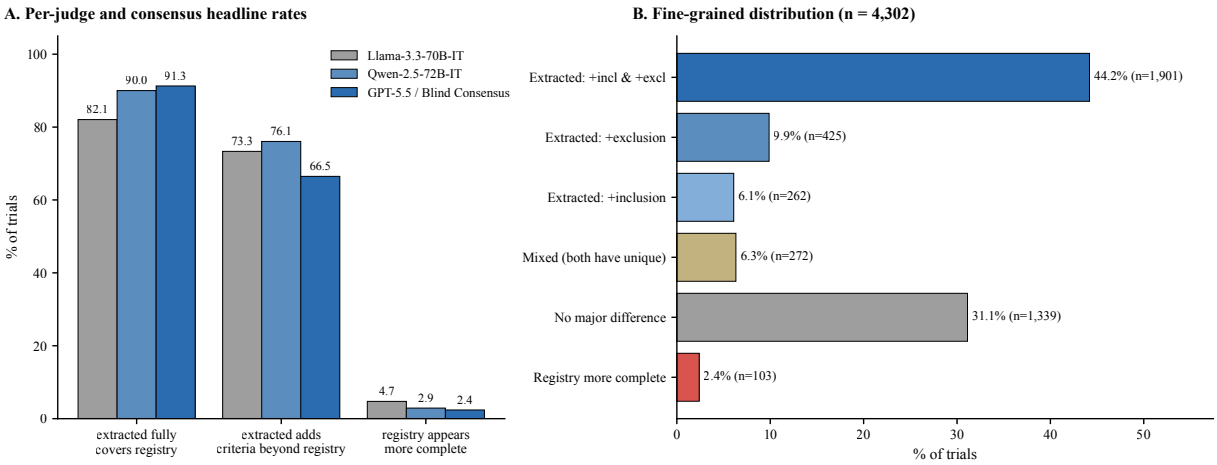

Figure 2: Three-judge qualitative audit comparing protocol-extracted and ClinicalTrials.gov registry eligibility criteria (EC). Two independent LLM judges (Llama-3.3-70B-Instruct, Qwen-2.5-72B-Instruct) compared blinded EC pairs, with GPT-5.5 adjudicating disagreements. (A) Per-judge and consensus rates on three headline outcomes: protocol-extracted EC fully covers the registry EC; protocol-extracted EC adds criteria beyond the registry; and registry EC appears strictly more complete. (B) Fine-grained consensus distribution across 4,302 trials, decomposing the three headline outcomes by location of enrichment or omission.

3.3 Q-ECGen

Figure 1C illustrates Q-ECGen, which combines hard textual prompting with biomedical embedding guidance. A frozen biomedical encoder produces token-level representations that a Q-Former-style connector converts into soft prefix tokens. These are then interleaved with the textual prompt and input to a LoRA-tuned decoder.

Cross-modal connector. A frozen BioBERT encoder E_{b} encodes the trial text x and a retrieved reference trial r. We concatenate their token-level representations into a single memory and project it to the connector width:

$$\begin{aligned} \mathbf{H} &= \big[E_{\mathrm{b}}(x);\, E_{\mathrm{b}}(r)\big] \in \mathbb{R}^{T\times d_e}, \\ \mathbf{M} &= \mathbf{H}\mathbf{W}_{\mathrm{in}}^{\top} \in \mathbb{R}^{T\times d}, \end{aligned} \tag{2}$$

where $d_e = 768$, $d = 256$, and $\mathbf{W}_{\mathrm{in}} \in \mathbb{R}^{d\times d_e}$ maps BioBERT features into the connector space.

The connector maintains $N = 32$ learnable query vectors $\mathbf{Q}^{(0)} \in \mathbb{R}^{N\times d}$, refined by $K = 2$ pre-norm query-transformer blocks. For each block $\ell = 1, \ldots, K$, we apply query self-attention, cross-attention to the BioBERT memory, and a feed-forward network:

$$\tilde{\mathbf{Q}} = \mathbf{Q}^{(\ell-1)} + \mathrm{SA}_\ell\Big(\mathrm{LN}\Big(\mathbf{Q}^{(\ell-1)}\Big)\Big), \tag{3}$$

$$\hat{\mathbf{Q}} = \tilde{\mathbf{Q}} + \mathrm{CA}_\ell\Big(\mathrm{LN}\Big(\tilde{\mathbf{Q}}\Big), \mathbf{M}\Big), \tag{4}$$

$$\mathbf{Q}^{(\ell)} = \hat{\mathbf{Q}} + \mathrm{FFN}_\ell\Big(\mathrm{LN}\Big(\hat{\mathbf{Q}}\Big)\Big). \tag{5}$$

We apply pre-LayerNorm (LN) before each sublayer to ensure training stability. SA_ℓ is h=8-head self-attention over the query tokens, CA_ℓ is cross-attention from queries to the BioBERT memory, and FFN_ℓ has hidden width $4d$. The memory is fixed across blocks; only the query states are updated, so the learnable queries iteratively read task-relevant biomedical information from the encoder memory.

The final query states are projected into the decoder embedding space to form N soft prefix tokens:

$$\mathbf{S} = \gamma\ \mathrm{LN}\Big(\mathbf{Q}^{(K)}\mathbf{W}_{\mathrm{out}}^{\top} + \mathbf{b}_{\mathrm{out}}\Big) \in \mathbb{R}^{N\times d_{\mathrm{dec}}}, \tag{6}$$

where $\mathbf{W}_{\mathrm{out}} \in \mathbb{R}^{d_{\mathrm{dec}}\times d}$, $d_{\mathrm{dec}} = 4096$, and $\gamma = 0.01$ matches the magnitude of the decoder's input embeddings.

Decoding. We insert the soft prefix tokens between the embedded textual prompt and a fixed generation launchpad:

$$\mathbf{Z} = \big[E_{\mathrm{d}}(x_{\mathrm{p}});\, \mathbf{S};\, E_{\mathrm{d}}(x_{\mathrm{lp}})\big], \tag{7}$$

where E_{d} is the decoder input-embedding layer, x_{p} contains the instruction, trial features, and retrieved reference criteria, and x_{lp} is the launchpad "Inclusion Criteria:". A LoRA-tuned decoder D_ϕ then generates autoregressively, minimizing

$$\mathcal{L}_{\mathrm{gen}} = -\sum_{t=1}^{|\mathbf{y}|} \log p_{D_\phi}(y_t \mid \mathbf{y}_{<t}, \mathbf{Z}). \tag{8}$$

The BioBERT encoder and base decoder remain frozen; only the cross-modal connector and LoRA parameters are updated.

Implementation details. Q-ECGen uses BioBERT hidden size $d_e = 768$, connector width $d = 256$, $N = 32$ query tokens, $K = 2$ connector blocks, $h = 8$ attention heads, decoder hidden size $d_{\text{dec}} = 4096$, and soft-token scale $\gamma = 0.01$. The decoder is Llama-3.1-8B with LoRA rank $r = 64$ and scaling factor $\alpha = 16$. For comparability with prior work, training and generation hyperparameters (epochs, learning rate, batch size, decoding temperature, and sampling parameters) follow the EC-RAFT setup (Lekuthai et al., 2025). Full values are listed in Appendix B.

3.4 Evaluation

We evaluate generated EC using both lexical and semantic metrics. Lexical metrics include ROUGE-L precision and ROUGE-L F1 (Lin, 2004), BLEU (Papineni et al., 2002), and chrF (Popović, 2015); semantic metrics include BERTScore F1 (Zhang et al., 2020) and BioBERT cosine similarity (Lee et al., 2020) between generated and reference EC. All benchmark results are averaged across seeds 1, 2, and 3 and reported as mean ± standard deviation.

4 Results

Dataset statistics and quantitative validation. PROTOCOLEC contains 4,302 paired clinical trials drawn from 22 therapeutic-area categories; the largest are Oncology (13.8%), Infectious Disease (13.3%), Neurology (9.1%), Dermatology (7.8%), Endocrinology (7.7%), and Hematology (6.2%) (Table 2). Our pipeline extracts 262,773 atomic eligibility rules in total (88,816 inclusion, 173,957 exclusion). Table 1 reports per-trial statistics by split. In every split, protocol-derived EC contain approximately twice as many criteria and twice as many words as the corresponding ClinicalTrials.gov registry entries. This confirms that registry EC are systematically compressed relative to protocol-level criteria.

Qualitative validation. We further evaluate target quality with a blinded three-judge audit (judging prompt, questions, and label definitions in Appendix C). For every trial, two independent LLM judges, Llama-3.3-70B-Instruct and Qwen-2.5-72B-Instruct, first decide whether the protocol-extracted EC fully covers the registry EC, then assign one of six fine-grained labels describing how the two sets differ (e.g., extracted adds inclusion criteria, extracted adds exclusion criteria, mixed, no major difference, registry more complete). Disagreements between the two judges are adjudicated by GPT-5.5.

Consensus finds that protocol-extracted EC fully cover the registry EC in 91.3% of trials, and add criteria beyond the registry in 66.5% of trials (Figure 2A). The fine-grained distribution in Figure 2B further characterizes how the two sets differ. Among the 91.3% fully covered cases, 44.2% add both inclusion and exclusion criteria, 9.9% add only exclusion, 6.1% add only inclusion, and 31.1% show no major difference. The remaining 8.7% of trials are not fully covered: 6.3% are mixed cases where each set contains unique criteria, and only 2.4% have a registry EC that is strictly more complete. Combining the four categories where the extracted EC adds something beyond the registry (44.2 + 9.9 + 6.1 + 6.3) recovers the 66.5% headline above.

To understand the residual 2.4% where the registry is judged more complete, we manually traced all 103 trials' missing items: 64.6% are locator under-coverage issues, 27.8% reflect content genuinely absent from the protocol PDF, 7.6% are judge false positives, and 0.5% are unparseable documents (full failure-mode trace procedure in Appendix C.5). Even among the strictest disagreement cases, the majority of errors arise from tractable extraction issues rather than fundamental information loss in protocol-derived EC.

Benchmark performance. Table 3 reports benchmark results on the PROTOCOLEC test set. We organize baselines into four families. *General-purpose open-weight LLMs* include Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen-2.5-7B-Instruct (Qwen Team et al., 2025), and Gemma-2-9B-Instruct (Gemma Team et al., 2024), evaluated both zero-shot and after supervised fine-tuning on PROTOCOLEC. *Biomedical LLMs* comprise Meditron-7B (Chen et al., 2023), a Llama-2-based model continued-pretrained on PubMed and clinical guidelines, and Panacea-7B (Lin et al., 2024), a clinical-trial foundation model fine-tuned on trial search, summarization, design, and matching tasks. *Proprietary GPT models* include GPT-4o-mini, GPT-5.4-mini, and GPT-5.4 in the zero-shot setting (Achiam et al., 2023). The strongest

Table 3: Benchmark results on the ProtocolEC test set. Results are averaged across seeds 1, 2, and 3 and reported as mean ± standard deviation. ↑ indicates higher is better. **Best** results in each metric are bolded and second-best results are underlined.

Method	**BERTScore F1** ↑	**BioBERT Cos.** ↑	**ROUGE-L P** ↑	**ROUGE-L F1** ↑	**BLEU** ↑	**chrF** ↑
GPT-4o-mini (zero-shot)	0.835 ± 0.00	0.961 ± 0.00	0.333 ± 0.00	0.156 ± 0.00	0.938 ± 0.01	17.501 ± 0.07
GPT-5.4-mini (zero-shot)	0.828 ± 0.00	0.958 ± 0.00	0.298 ± 0.00	0.147 ± 0.00	0.856 ± 0.02	19.200 ± 0.06
GPT-5.4 (zero-shot)	0.828 ± 0.00	0.956 ± 0.00	0.247 ± 0.00	0.161 ± 0.00	2.606 ± 0.02	26.778 ± 0.15
Llama-3.1-8B-Instruct (zero-shot)	0.823 ± 0.00	0.941 ± 0.00	0.228 ± 0.00	0.155 ± 0.00	3.504 ± 0.08	25.206 ± 0.17
Qwen-2.5-7B-Instruct (zero-shot)	0.827 ± 0.00	0.954 ± 0.00	0.247 ± 0.00	0.151 ± 0.00	2.679 ± 0.03	23.171 ± 0.16
Gemma-2-9B-Instruct (zero-shot)	0.823 ± 0.00	0.947 ± 0.00	0.273 ± 0.00	0.143 ± 0.00	1.736 ± 0.09	19.344 ± 0.29
Meditron-7B (zero-shot)	0.830 ± 0.00	0.959 ± 0.00	0.321 ± 0.00	0.149 ± 0.00	0.884 ± 0.05	17.410 ± 0.25
Panacea-7B (zero-shot)	0.828 ± 0.00	0.952 ± 0.00	0.215 ± 0.00	0.157 ± 0.00	5.198 ± 0.16	28.258 ± 0.39
Llama-3.1-8B-Instruct (fine-tuned)	0.839 ± 0.00	0.964 ± 0.00	0.256 ± 0.00	0.163 ± 0.00	5.344 ± 0.29	28.020 ± 0.49
Qwen-2.5-7B-Instruct (fine-tuned)	0.835 ± 0.00	0.960 ± 0.00	0.221 ± 0.00	0.148 ± 0.00	5.589 ± 0.13	28.765 ± 0.68
Gemma-2-9B-Instruct (fine-tuned)	0.835 ± 0.00	0.966 ± 0.00	0.193 ± 0.01	0.145 ± 0.00	6.102 ± 0.44	32.439 ± 1.39
Meditron-7B (fine-tuned)	0.840 ± 0.00	0.965 ± 0.00	0.250 ± 0.00	0.165 ± 0.00	6.051 ± 0.22	29.805 ± 0.38
Panacea-7B (fine-tuned)	0.839 ± 0.00	0.959 ± 0.00	0.257 ± 0.00	0.163 ± 0.00	6.146 ± 0.11	28.630 ± 0.38
EC-RAFT (fine-tuned)	0.856 ± 0.00	0.958 ± 0.00	0.353 ± 0.01	0.315 ± 0.01	18.505 ± 0.68	43.544 ± 0.77
Q-ECGen (ours)	$\mathbf{0.869 \pm 0.00}$	$\mathbf{0.973 \pm 0.00}$	$\mathbf{0.435 \pm 0.00}$	$\mathbf{0.353 \pm 0.00}$	$\mathbf{29.884 \pm 0.35}$	$\mathbf{49.190 \pm 0.57}$

existing baseline is *EC-RAFT* (Lekuthai et al., 2025), which fine-tunes Llama-3.1-8B-Instruct with retrieval-augmented training and synthesized reasoning traces. We note that AutoTrial (Wang et al., 2023), an earlier EC generation system, is not included as a baseline because its code has not been publicly released and we were unable to faithfully reproduce its setup from the paper alone.

Across baselines, two patterns emerge. First, supervised fine-tuning consistently improves over zero-shot generation for every open-weight model: Llama-3.1-8B-Instruct's BERTScore F1 rises from 0.823 to 0.839, Meditron-7B from 0.830 to 0.840, and similar gains hold for Qwen, Gemma, and Panacea. Yet even the best fine-tuned non-retrieval model (Meditron-7B at BERTScore F1 = 0.840, ROUGE-L F1 = 0.165) trails EC-RAFT by a wide margin on the lexical metrics that most directly capture criterion-level overlap (EC-RAFT: 0.856 BERTScore F1, 0.315 ROUGE-L F1, 18.5 BLEU). Second, within the proprietary GPT family, the older and smaller GPT-4o-mini is competitive with the newer GPT-5.4 variants on semantic metrics (BERTScore F1 = 0.835 vs. 0.828; ROUGE-L precision = 0.333 vs. 0.247), suggesting that capacity alone does not determine zero-shot performance on this task.

Q-ECGEN improves over EC-RAFT across all six metrics: BERTScore F1 increases from 0.856 to 0.869, BioBERT cosine similarity from 0.958 to 0.973, ROUGE-L precision from 0.353 to 0.435, ROUGE-L F1 from 0.315 to 0.353, BLEU from 18.5 to 29.9, and chrF from 43.5 to 49.2. Because

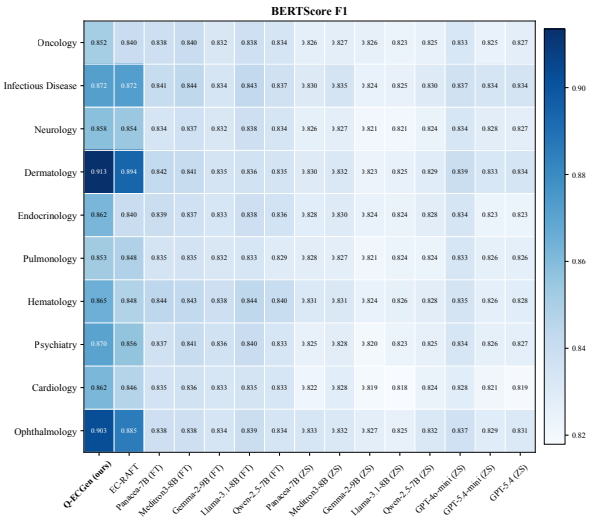

Figure 3: Per-therapeutic-area BERTScore F1 on the ProtocolEC test set. We report results for the ten largest therapeutic areas by trial count, averaged across seeds 1, 2, and 3. Darker blue indicates higher performance. Q-ECGen achieves the highest BERTScore F1 across all therapeutic areas, showing consistent generalization beyond the aggregate test-set results.

EC-RAFT already combines retrieval, fine-tuning, and reasoning, these gains indicate that biomedical embedding guidance via a Q-Former-style connector contributes signal beyond what retrieval-augmented fine-tuning alone provides.

Per-therapeutic-area performance. To assess whether the aggregate gains are driven by a small number of disease areas, Figure 3 reports BERTScore F1 across the ten largest therapeutic areas. Q-ECGEN achieves the highest BERTScore

Table 4: Ablation study of Q-ECGEN on the PROTOCOLEC test set. All variants use LoRA supervised fine-tuning on Llama-3.1-8B-Instruct.

Model Variant	BERTScore F1	ROUGE-L F1	BLEU	chrF
LoRA SFT	0.839	0.163	5.379	28.263
LoRA SFT + Q-Former	0.851	0.240	12.703	36.471
LoRA SFT + Retrieval	0.853	0.314	15.762	40.856
Q-ECGen (Q-Former + Retrieval)	**0.867**	**0.350**	**29.478**	**48.534**

F1 in every area, with the largest gains in Dermatology (0.913 vs. 0.894 for EC-RAFT) and Ophthalmology (0.903 vs. 0.885). Gains generalize across diverse clinical domains rather than concentrating in a single dominant area.

Ablation study. To isolate the contribution of each component of Q-ECGEN, we ablate the Q-Former connector and the retrieved reference context (Table 4). Starting from the LoRA-SFT baseline, the Q-Former alone improves all four metrics, and retrieval alone yields larger gains. The *LoRA SFT + Retrieval* variant achieves performance comparable to EC-RAFT, with residual gap that we attribute to EC-RAFT's use of synthesized reasoning steps during fine-tuning. Combining both components in Q-ECGEN yields further gains across all four metrics, indicating that the BioBERT-guided soft prefix and retrieved reference text are complementary rather than redundant.

Sensitivity analysis. We examine Q-ECGEN's sensitivity to connector capacity by varying connector depth L and the number of learnable query tokens q (Table 7, Appendix D). Performance is stable. Larger configurations yield marginal and inconsistent gains relative to their increased capacity, so we use the smallest setting in the main experiments.

5 Conclusion

We present PROTOCOLEC, a protocol-derived benchmark for complete clinical trial eligibility criteria generation. Unlike prior resources that use ClinicalTrials.gov registry EC as generation targets, PROTOCOLEC is constructed from full protocol documents and captures substantially more detailed inclusion and exclusion requirements. Our quantitative analysis shows that protocol-derived EC contain approximately twice as many criteria and twice as many words as registry EC, and a blinded three-judge audit confirms that they preserve the registry content in over 91% of trials and add new criteria in over 66%.

Using PROTOCOLEC, we benchmark zero-shot LLMs, fine-tuned LLMs, biomedical models, and retrieval-augmented EC generation, and introduce Q-ECGEN, an embedding-guided generation framework that adopts a Q-Former-style connector to condition a LoRA-tuned decoder on BioBERT-derived soft prefix tokens. Q-ECGEN consistently improves over all baselines, including EC-RAFT, the strongest published retrieval-augmented method, across both aggregate and per-therapeutic-area evaluation. An ablation study shows that the Q-Former connector and retrieved reference EC contribute complementary signal, with the full model exceeding either component in isolation. Sensitivity analysis shows that the framework is stable across connector configurations, supporting the use of a lightweight adapter.

Together, PROTOCOLEC and Q-ECGEN reposition EC generation from matching ClinicalTrials.gov summaries to recovering the protocol-level criteria that actually govern enrollment. Model weights, code, and dataset will be released upon acceptance.

Limitations

PROTOCOLEC is built from completed Phase III trials with available protocol PDFs on ClinicalTrials.gov, and is overwhelmingly English-language. We focus on Phase III because it is the pivotal stage where EC design is most consequential and most complex. Earlier-phase trials and non-English protocols are out of scope. The dataset is constructed with LLM-based extraction, which may introduce hallucinations or omissions. Further validation could strengthen the reliability of the extracted criteria. Our evaluation uses lexical and semantic similarity metrics, which measure text-level resemblance to reference EC rather than clinical correctness where a criterion can be lexically close but clinically wrong. Finally, the qualitative audit uses LLM judges, which can carry biases that human evaluators do not. We mitigate these through blinding and the use of two independent open-weight judges from different model families before adjudication, but LLM judging cannot verify clinical accuracy and should be read as a coverage and enrichment audit rather than a correctness audit. Expert evaluation by clinicians is necessary to assess whether generated criteria would be safe and acceptable in a real protocol-design workflow.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alban Bornet, Philipp Khlebnikov, Florian Meer, Quentin Haas, Anthony Yazdani, Boya Zhang, Poorya Amini, and Douglas Teodoro. 2025. Analysis of eligibility criteria clusters based on large language models for clinical trial design. *Journal of the American Medical Informatics Association*, 32(3):447–458.

Benjamin Carlisle, Jonathan Kimmelman, Tim Ramsay, and Nathalie MacKinnon. 2015. Unsuccessful trial accrual and human subjects protections: An empirical analysis of recently closed trials. *Clinical Trials*, 12(1):77–83.

Jintai Chen, Yaojun Hu, Mingchen Cai, and 1 others. 2025. Trialbench: Multi-modal ai-ready datasets for clinical trial prediction. *Scientific Data*, 12(1):1564.

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

Surabhi Datta, Kyeryoung Lee, Hunki Paek, and 1 others. 2024. AutoCriteria: A generalizable clinical trial eligibility criteria extraction system powered by large language models. *Journal of the American Medical Informatics Association*, 31(2):375–385.

Nicholas J. Dobbins, Tony Mullen, Özlem Uzuner, and Meliha Yetisgen. 2022. The Leaf clinical trials corpus: A new resource for query generation from clinical trial eligibility criteria. *Scientific Data*, 9(1):490.

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. 2022. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4):100445.

Chufan Gao, Jathurshan Pradeepkumar, Trisha Das, Shivashankar Thati, and Jimeng Sun. 2024. Automatically labeling clinical trial outcomes: A large-scale benchmark for drug development. *arXiv preprint arXiv:2406.10292*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, pages 1–16.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, pages 1–92.

R. Donald Harvey, Kathryn F. Mileham, Vishal Bhatnagar, and 1 others. 2021. Modernizing clinical trial eligibility criteria: Recommendations of the asco-friends of cancer research washout period and concomitant medication work group. *Clinical Cancer Research*, 27(9):2400–2407.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Qiao Jin, Zifeng Wang, Charalampos S. Floudas, and 1 others. 2024. Matching patients to clinical trials with large language models. *Nature Communications*, 15(1):9074.

Susan Jin, Richard Pazdur, and Rajeshwari Sridhara. 2017. Re-evaluating eligibility criteria for oncology clinical trials: Analysis of investigational new drug applications in 2015. *Journal of Clinical Oncology*, 35(33):3745–3752.

Edward S. Kim, Suanna S. Bruinooge, Samantha Roberts, and 1 others. 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.

Edward S. Kim, Thomas S. Uldrick, Caroline Schenkel, and 1 others. 2021. Continuing to broaden eligibility criteria to make clinical trials more representative and inclusive: Asco-friends of cancer research joint research statement. *Clinical Cancer Research*, 27(9):2394–2399.

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

Fabrício Kury, Alex Butler, Chi Yuan, Li-heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. 2020. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific Data*, 7(1):281.

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

Nopporn Lekuthai, Nattawit Pewngam, Supitcha Sokrai, and Titipat Achakulvisut. 2025. EC-RAFT: Automated generation of clinical trial eligibility criteria through retrieval-augmented fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9432–9444, Vienna, Austria. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiacheng Lin, Hanwen Xu, Zifeng Wang, Sheng Wang, and Jimeng Sun. 2024. Panacea: A foundation model for clinical trial search, summarization, design, and recruitment. *arXiv preprint arXiv:2407.11007*.

Ruishan Liu, Shemra Rizzo, Samuel Whipple, and 1 others. 2021. Evaluating eligibility criteria of oncology trials using real-world data and ai. *Nature*, 592(7855):629–633.

Allison Magnuson, Suanna S. Bruinooge, Harpreet Singh, and 1 others. 2021. Modernizing clinical trial eligibility criteria: Recommendations of the asco-friends of cancer research performance status work group. *Clinical Cancer Research*, 27(9):2424–2429.

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.

Raymond U. Osarogiagbon, Diana Merino Vega, Lola Fashoyin-Aje, and 1 others. 2021. Modernizing clinical trial eligibility criteria: Recommendations of the ASCO–Friends of Cancer Research prior therapies work group. *Clinical Cancer Research*, 27(9):2408–2415.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jimyung Park, Yilu Fang, Casey Ta, and 1 others. 2024. Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *Journal of Biomedical Informatics*, 154:104649.

John S. Peterson, Deborah Plana, Danielle S. Bitterman, Skyler Bryce Johnson, Hugo J. W. L. Aerts, and Benjamin Harris Kann. 2023. Growth in eligibility criteria content and failure to accrue among National Cancer Institute (NCI)-affiliated clinical trials. *Cancer Medicine*, 12(4):4715–4724.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Qwen Team, An Yang, Baosong Yang, and 1 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, pages 1–26.

Alexander I. Spira, Mark D. Stewart, Suzanne Jones, and 1 others. 2021. Modernizing clinical trial eligibility criteria: Recommendations of the asco-friends of cancer research laboratory reference ranges and testing intervals work group. *Clinical Cancer Research*, 27(9):2416–2423.

U.S. National Library of Medicine. 2024. Clinicaltrials.gov. `https://clinicaltrials.gov/`. Accessed: 2024.

Zifeng Wang and Jimeng Sun. 2022. Trial2Vec: Zero-shot clinical trial document similarity search using self-supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6377–6390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Chi Yuan, Patrick B. Ryan, Casey Ta, and 1 others. 2019. Criteria2query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4):294–305.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Appendix

A Dataset Construction Details

This appendix provides additional details on the PROTOCOLEC construction pipeline summarized in Section 3.

A.1 Eligibility-Section Locator

The locator described in Section 3 proceeds through three deterministic stages, falling through to the next stage whenever a candidate slice fails an EC-signal validation check.

Stage 1: PDF bookmarks. We read the PDF outline (`get_toc`) and search bookmark titles for inclusion-, exclusion-, or eligibility-related entries. Inclusion bookmarks contain "inclusion" and "criteria"; exclusion bookmarks contain "exclusion" and "criteria"; eligibility-parent bookmarks contain phrases such as "eligibility," "patient/subject selection," "selection of patients/subjects," or "study/trial population." Bookmarks with malformed or unresolvable page targets (e.g., back-references reporting page 0) are skipped. If an inclusion bookmark is found, the next bookmark after the EC region bounds the end of the slice; otherwise we extend eight pages beyond the inclusion or exclusion bookmark.

Stage 2: Body-text heading scan. We scan every page of the PDF for heading-like text using anchored regular expressions on whole lines. The patterns accept stacked qualifiers (e.g., "Main Patient Inclusion Criteria," "Subject Inclusion Criteria"), section numbering (e.g., "4.1.1 Inclusion Criteria"), and alternative phrasings (e.g., "Criteria for Inclusion," "Eligibility for Inclusion"). The same approach yields three pattern families:

- **Inclusion** and **Exclusion criteria** headings, possibly preceded by section numbers and up to two qualifier words.
- **Eligibility-parent** headings (e.g., "Eligibility Criteria," "Patient Selection," "Study Population," "Entry/Enrollment Criteria," "Inclusion and Exclusion Criteria," "I/E Criteria").
- **End-of-section** headings used to bound the EC slice from below (e.g., "Withdrawal Criteria," "Study Procedures," "Randomization," "Concomitant Medications," "Safety Assessments," "Statistical Analysis," "Schedule of Assessments," "Study Endpoints").

We collect *all* heading hits across the document, then prefer the latest inclusion heading that is followed by an exclusion heading within five pages—this skips the abbreviated EC summary that many protocols include in their front matter. Amendment-style prefixes (e.g., "Changes to Inclusion Criteria," "Updated Inclusion Criteria") are explicitly excluded from matching, since they describe revisions rather than the criteria themselves. Pages that resemble tables of contents (detected by dot-leader patterns, "Table of Contents" headers, or stacked page-reference lines) are skipped during the heading scan to avoid false positives from TOC entries that pymupdf splits across lines.

Stage 3: Table-of-contents parser. We search the first 25% of the document (with a floor of 25 and ceiling of 60 pages) for pages containing at least three dot-leader patterns (e.g., `".....127"`), which identifies rendered TOC pages. Regular expressions then extract the printed page numbers for "Inclusion Criteria," "Exclusion Criteria," and eligibility-parent entries. Printed page numbers are mapped to PDF page indices using the PDF's own page-label metadata when available; otherwise a 1:1 fallback is applied (printed page 1 = PDF index 0). This stage rescues hybrid PDFs whose EC pages are scanned but whose TOC is digital.

EC-signal validation. Each candidate slice is validated using a signal-counting heuristic before being accepted. We count list-marker patterns (numbered, lettered, bulleted, and parenthesized list items) and EC-phrase patterns ("inclusion criteria," "exclusion criteria," "eligibility criteria," "subjects/patients must have/be/meet…"), weighting phrase hits at $2\times$. A slice is accepted if the total signal is at least 4. Slices with fewer than 200 characters of extractable text are accepted unconditionally, since the locator may be correct but the text unreadable due to scanning. Two counter-signals can reject an otherwise plausible slice: (i) lines containing both an amendment-tracking word ("updated," "revised," "modified," "removed," "added," "deletion of," "addition of," "changed") and an EC phrase are treated as amendment-change descriptions rather than EC content, and a slice with three or more such lines is rejected; (ii) if amendment-style language outweighs EC signal anywhere in the slice, the slice is rejected. Failed validation triggers fall-through to the next locator strategy.

A.2 LLM-Based Eligibility-Criteria Extraction Prompt

For each localized protocol slice, we render the padded page range as a fresh PDF and pass it to GPT-5.4 with the following prompt:

```
You are reading selected pages of a clinical trial protocol PDF. These pages should contain the inclusion and exclusion criteria for the trial.

Extract the eligibility criteria verbatim and output them in EXACTLY this format (no preamble, no markdown, no headings other than the two below):

Inclusion Criteria:
- <criterion 1 text>
- <criterion 2 text>
. . .

Exclusion Criteria:
- <criterion 1 text>
- <criterion 2 text>
. . .

Rules:

  • Preserve EXACT thresholds, units, and time windows. Do not paraphrase.
  • One criterion per dash-prefixed line.
  • Expand sub-bullets and tabular criteria into separate lines.
  • Drop page headers, footers, and "see Section X" cross-references with no criterion text.
  • If both Inclusion and Exclusion sections exist, output both. If only one exists, still emit the other header followed by an empty line.

If these pages do NOT contain the eligibility criteria, output exactly the single line: CANNOT_FIND_EC

Output ONLY one of the two formats above. No JSON, no markdown fences, no commentary.
```

If slice-based extraction returns `CANNOT_FIND_EC` or produces malformed output, we fall back to passing the full PDF to GPT-5.4 with a variant prompt that asks the model to locate and extract the eligibility section anywhere in the document, returning `NO_EC_FOUND` if absent.

A.3 Therapeutic-Area Classifier Prompt

Each trial is assigned a primary therapeutic area using GPT-5.4 conditioned on the registry condition list, brief title, and brief summary. The classifier also emits a SNOMED CT preferred term for the trial's primary clinical concept, which we use during dataset analysis. The system prompt is:

```
You are a clinical-trial taxonomy classifier. For each trial you receive the registry's free-text Conditions, Brief Title, and Brief Summary. You must:

  1. Emit snomed_term: the single SNOMED CT preferred term that best captures the trial's PRIMARY clinical concept. Use exact SNOMED CT wording (e.g., "Type 2 diabetes mellitus," "Non-small cell lung carcinoma," "Atrial fibrillation," "Major depressive disorder"). If multiple conditions apply, pick the one the trial is principally studying based on the title and summary.
  2. Emit primary_therapeutic_area: the parent bucket. You will be given the current parent list—strongly prefer an existing parent if any reasonably fits. Only propose a new parent if none of the existing ones cover this trial's area; name it as a broad therapeutic specialty in Title Case (e.g., "Pulmonology," not "Severe Asthma").
  3. Emit is_new_parent: true if proposing a new parent, false otherwise.
  4. Emit rationale: one short sentence explaining the choice.

Reply with strict JSON only.
```

The final therapeutic-area taxonomy consists of the 22 categories listed in Table 2, which emerged from incremental classification across all trials with `is_new_parent=true` responses reviewed and merged into the taxonomy when justified.

B Hyperparameters

Table 5 reports the full set of training, model, and generation hyperparameters used for Q-ECGEN. Training-side values follow the EC-RAFT setup (Lekuthai et al., 2025) for comparability.

Hardware and software. Training and evaluation run on 8 × NVIDIA RTX PRO 6000 Blackwell Max-Q GPUs (96 GB HBM each), via `torchrun -nproc_per_node=8` with DDP. Each rank loads the full Llama-3.1-8B with LoRA adapters and the frozen BioBERT encoder. Evaluation peak memory is approximately 70 GB per GPU at `gen_batch_size=32` with `max_new_tokens=4096`.

Post-processing. Generated text is post-processed before metric computation as follows:

Table 5: Training, model, and generation hyperparameters for Q-ECGEN.

Hyperparameter	Value
Training	
Optimizer	AdamW (β_1=0.9, β_2=0.999)
Learning rate	2×10^{-4}
LR schedule	Linear warmup, then linear decay
Warmup ratio	0.03 (of total optimizer steps)
Weight decay	0.0
Gradient clipping	1.0 (global ℓ_2 norm)
Mixed precision	bf16 (`torch.amp.autocast`)
Per-device batch size	2
Gradient accumulation	4 steps
GPUs (DDP)	8
Effective batch size	64 ($2 \times 4 \times 8$)
Epochs	3
Gradient checkpointing	enabled
Loss	generation cross-entropy
Seeds	{1, 2, 3} for main results; 1 for sensitivity sweep
Architecture	
Encoder (frozen)	BioBERT (`dmis-lab/biobert-base-cased-v1.2`), 768-d, max length 512
Decoder	Llama-3.1-8B-Instruct, base frozen, LoRA adapters
LoRA rank r	64
LoRA α	16
LoRA dropout	0.1
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Connector queries N	32 (default; sensitivity sweep $\in \{32, 64, 128\}$)
Connector depth K	2 blocks (default; sensitivity sweep $\in \{2, 4, 8\}$)
Connector hidden d	256
Attention heads h	8
FFN expansion	$4d$
Query initialization	$\mathcal{N}(0, 0.02^2)$
Output projection ($d \to d_{\text{dec}}$)	Linear $\to$ LayerNorm $\times\gamma$, $\gamma = 0.01$
Generation	
Decoding	stochastic sampling
Temperature	0.3
min-p	0.03 (drop tokens with prob $< 0.03 \times \max$ prob)
max_new_tokens	4096
Attention backend	SDPA flash kernel
Generation batch size	32

(i) truncate at the first `[SEP]` token; (ii) truncate at the first occurrence of any junk marker (HTML entities, URL query fragments, Q&A continuations, disclaimer or sponsor boilerplate, essay-prose phrases). The cleaned text is then split into inclusion and exclusion sections at the "Exclusion Criteria:" header.

C Three-Judge Qualitative Audit

This appendix provides the full prompts, label definitions, and adjudication procedure used in the three-judge audit reported in Section 4 (Figure 2).

C.1 Audit Setup

For every trial in PROTOCOLEC we presented two independent LLM judges with a blinded pair of EC sets and elicited a symmetric structured comparison. To prevent the judge from inferring which set was the registry and which was the protocol-extracted output, each pair was shown as "Source A" and "Source B" without identifying labels; the trial NCT ID was withheld from the prompt for the same reason. The mapping (Source A = registry, Source B = extracted) was fixed in the analysis pipeline but never disclosed to the judge.

The two primary judges are Llama-3.3-70B-Instruct and Qwen-2.5-72B-Instruct, served locally with vLLM at temperature 0. When the two judges disagreed on either the binary coverage question or the coarse three-way comparison bucket (extracted-wins / registry-wins / tie), or when either judge returned an error, the same prompt was re-run with GPT-5.5 as a blind adjudicator. GPT-5.5 received only the trial text and EC pair, not the prior judges' verdicts.

C.2 Judging Prompt

All three judges receive the same system prompt:

```
You are a clinical-research reviewer
comparing TWO versions of the
eligibility criteria (EC) for the
SAME clinical trial. You will see
Source A and Source B—you do not know
which is which and you should judge
them on their content alone, without
assuming either is the ground truth.

IMPORTANT—what counts as a difference:

  • Allow reasonable paraphrasing,
    reformatting, and re-grouping.
    Wording differences DO NOT count
    as missing information when
    the same clinical constraint is
    expressed.
      – example: A says "Age ≥18 years"
        and B says "Adults aged 18
        or older" → SAME content, no
        difference.
      – example: A says "Female, age
        50–75" and B says "Female" +
        "Age 50 to 75 years" as two
        bullets → SAME content, no
        difference.
  • Only count SUBSTANTIVE
    differences—a clinical constraint
    that appears in one source and
    has no analogue (even loosely
    paraphrased) in the other.

Answer five boolean flags + rationale.
Reply with strict JSON only.

  • a_has_extra_inclusion: true if
    Source A contains at least one
    substantive INCLUSION criterion
    that has no analogue anywhere in
    Source B.
  • a_has_extra_exclusion: true if
    Source A contains at least one
    substantive EXCLUSION criterion
    that has no analogue anywhere in
    Source B.
  • b_has_extra_inclusion: true if
    Source B contains at least one
    substantive INCLUSION criterion
    that has no analogue anywhere in
    Source A.
  • b_has_extra_exclusion: true if
    Source B contains at least one
    substantive EXCLUSION criterion
    that has no analogue anywhere in
    Source A.
  • noise: true if either source is
    too garbled, irrelevant, or non-EC
    content to compare meaningfully.
  • rationale: 2–3 sentences citing
    the concrete differences (refer to
    A/B only—no claims about which is
    "better" or "official").
```

The user prompt presents the trial's brief title and brief summary (truncated to 600 characters) for clinical context, followed by the Source A and Source B EC blocks and a final instruction to reply with the JSON schema above.

C.3 Derived Labels

The five boolean flags returned by the judge are deterministically mapped to the binary coverage outcome (**Q1**) and the fine-grained comparison label (**Q2**) reported in the main paper. Because the judge is blinded to the Source A / Source B mapping, the four directional "extras" flags are interpreted under the fixed convention Source A = registry, Source B = extracted.

Q1 — fully covered. The trial is labeled "yes" if the registry has no extras over the extracted (neither extra inclusion nor extra exclusion criteria), and "no" otherwise.

Q2 — comparison label. The six labels are assigned as follows:

No major difference. Neither source has any extras over the other. The two EC sets express the same clinical content modulo paraphrasing.

Extracted: +inclusion. The extracted EC has extra inclusion criteria, no extra exclusion criteria, and the registry has no extras.

Extracted: +exclusion. The extracted EC has extra exclusion criteria, no extra inclusion criteria, and the registry has no extras.

Extracted: +incl & +excl. The extracted EC has extras on both axes (inclusion and exclusion) and the registry has no extras.

Registry more complete. The registry has extras (inclusion, exclusion, or both) and the extracted EC has none.

Mixed (both have unique). Each set contains substantive criteria that the other lacks.

When the judge sets the `noise` flag, the trial is excluded from headline statistics and reported separately. After adjudication, the noise rate was negligible and these trials are omitted from Figure 2.

C.4 Adjudication Procedure

GPT-5.5 is invoked as a blind third judge when Llama-3.3 and Qwen-2.5 disagree on either of two coarsened views of their verdicts:

1. **Q1 disagreement:** the two judges produce different binary coverage answers (one "yes," one "no").

Table 6: Per-rule failure-mode distribution across the 370 missing rules identified in the 103 registry-more-complete trials.

Failure mode	Rules	%
Locator under-coverage	121	32.7
Extractor omission	118	31.9
Not in protocol PDF	101	27.3
Judge false positive	28	7.6
PDF unparseable (scanned image)	2	0.5
Total	**370**	**100.0**

2. **Coarse Q2 disagreement:** the two judges' fine-grained labels fall into different coarse buckets, where the buckets are *EXTR_MORE* (any of the three Extracted-adds categories), *REG_MORE* (Registry more complete), and *TIE* (No major difference, Mixed, or Noise).

GPT-5.5 receives the same system and user prompts as the primary judges—no knowledge of either prior verdict—and produces an independent set of five flags. The adjudicator's verdict, deterministically mapped through the same Q1/Q2 derivation, becomes the trial's consensus label.

C.5 Failure-Mode Trace

To understand the 103 trials where the consensus judged the registry EC strictly more complete, we conducted a manual root-cause analysis of every missing rule flagged by the judges. For each trial, the two judges' free-text rationales (and the adjudicator's rationale when invoked) were parsed into a list of specific rules that the registry contained but the extracted EC lacked, yielding 1–6 missing rules per trial (mean 3.6) and 370 rules in total across the 103 trials.

Each rule was independently traced against four sources in order of evidence strength: (i) the extracted EC produced by our pipeline; (ii) the EC slice that was actually shown to the extraction LLM (located EC pages padded by five pages on each side); (iii) the full protocol PDF text; and (iv) no source, if the rule is genuinely absent from the PDF. Every label assignment required a verbatim substring match in the indicated source to prevent inferred or paraphrased evidence. Rules were assigned to one of four labels:

- **Extractor omission.** The rule is present in the EC slice text but absent from the extracted EC—the extraction model dropped content it could see.
- **Locator under-coverage.** The rule is present somewhere in the full PDF but outside the EC slice—the locator failed to pull the right pages, typically because the rule lives in appendices, cohort-specific subsections, or extended-arm eligibility blocks.
- **Not in protocol.** The rule is genuinely absent from the entire protocol PDF; the registry contains content the protocol does not.
- **Judge false positive.** The rule is actually present in the extracted EC; the judge over-flagged.

Two trials whose protocol PDFs were scanned images with fewer than 5,000 extractable characters were identified separately and labeled **PDF unparseable**, since neither the locator nor the extractor could see the eligibility section in these cases.

Table 6 reports the per-rule label distribution. Recoverable pipeline issues (extractor omission + locator under-coverage) account for 64.6% of missing rules, content genuinely absent from the protocol PDF accounts for 27.8%, and judge false positives account for 7.6%. At the trial level, 55.3% of disagreement cases have a primary failure mode that is recoverable in future pipeline iterations, while 34.0% reflect trials where the registry summary is intrinsically the more complete source.

Table 7: Sensitivity analysis of Q-ECGEN to connector depth L and number of learnable query tokens q on the PROTOCOLEC test set.

L	q	BERTScore F1 ↑	BioBERT Cos. ↑	ROUGE-L F1 ↑	BLEU ↑	chrF ↑
2	32	0.867	0.972	0.350	29.48	48.53
2	64	0.869	0.973	0.355	30.51	49.81
2	128	0.870	0.974	0.357	29.83	49.50
4	32	0.870	0.974	0.355	29.68	49.09
4	64	0.869	0.973	0.352	29.77	49.03
4	128	0.869	0.974	0.353	30.00	49.37
8	32	0.869	0.974	0.354	29.78	49.06
8	64	0.869	0.974	0.355	30.73	50.09
8	128	0.870	0.974	0.357	30.33	49.69

D Sensitivity Analysis

Table 7 reports the full sensitivity sweep referenced in Section 4, covering all nine configurations of connector depth $L \in \{2, 4, 8\}$ and learnable query tokens $q \in \{32, 64, 128\}$ on the PROTOCOLEC test set.