AMEND++: Benchmarking Eligibility Criteria Amendments in Clinical Trials

Anonymous ACL submission

Abstract

Clinical trial amendments frequently introduce delays, increased costs, and administrative burden, with eligibility criteria being the most commonly amended component. We introduce *eligibility criteria amendment prediction*, a novel NLP task that aims to forecast whether the eligibility criteria of an initial trial protocol will undergo future amendments. To support this task, we release AMEND++, a benchmark suite comprising two datasets: AMEND, which captures eligibility-criteria version histories and amendment labels from public clinical trials, and AMEND_LLM, a refined subset curated using an LLM-based denoising pipeline to isolate substantive changes. We further propose *Change-Aware Masked Language Modeling* (CAMLM), a revision-aware pretraining strategy that leverages historical edits to learn amendment-sensitive representations. Experiments across diverse baselines show that CAMLM consistently improves amendment prediction, enabling more robust and cost-effective clinical trial design. [1]

1 Introduction

Clinical trial protocols define the scientific and operational blueprint of a study and must be followed exactly as approved by regulatory authorities. In practice, however, many trials undergo post-approval modifications (i.e., protocol amendments) to address emerging scientific questions, operational challenges, or regulatory feedback. While often necessary, amendments introduce substantial burdens by delaying timelines, requiring extensive coordination across investigators and sites, and generating high additional costs for sponsors and regulatory bodies.

A prior study from the Tufts Center for the Study of Drug Development has shown that amendments are both frequent and increasingly complex, with a growing proportion triggered by design deficiencies or changes in eligibility criteria (Getz et al., 2016). More recent analyses indicate a rise in amendment prevalence and longer implementation cycles (Getz et al., 2024), reflecting the intensification of regulatory expectations and operational demands. Regulatory agencies face similar pressures, processing large volumes of amendments each year, often with multi-week review periods. These findings underscore that many amendments could be anticipated earlier if potential issues in the initial protocol were detected proactively.

Eligibility criteria (EC)[2] are a particularly common driver of protocol amendments (Getz et al., 2024). Changes to inclusion and exclusion criteria can arise for multiple reasons, as noted in ICH E9 and prior clinical research (Lewis, 1999; Cleophas et al., 2009). New medical insights may necessitate amendments, while persistent violations of entry criteria or inadequate recruitment can also prompt revisions. Because modifications to EC directly affect who can enroll, changes introduced mid-trial can yield meaningful differences between participant populations before and after an amendment, with implications for study integrity and interpretability. These factors make EC amendments both impactful and important to anticipate early in the design process.

Despite the central role of protocol design in clinical trial efficiency, the computational literature lacks large-scale resources and models to predict whether EC in an initial protocol will be modified later (Table 1). EC define

[1] We will publicly release code and data once the paper is accepted.

[2] Throughout this paper, we use EC to refer to the eligibility criteria section as a single textual unit.

Table 1: Summary of existing studies on clinical trial protocol amendments and dataset availability. None of the prior studies released a public dataset suitable for machine learning tasks.

Study	Data Source	# of Trials	Focus / Contribution	Public Dataset
(Getz et al., 2016)	15 pharmaceutical companies and CROs (Tufts CSDD collaboration)	836	Quantified prevalence, causes, and costs of protocol amendments across commercial trials.	✗
(Getz et al., 2024)	16 pharmaceutical companies and CROs	950	Updated benchmark analysis on amendment frequency and impact on study performance.	✗
(Botto et al., 2024)	(Getz et al., 2024)	950	Compared amendment prevalence and completion outcomes in oncology vs. non-oncology trials.	✗
(Joshi, 2023)	53 NHS-sponsored trials (UK)	53	Mixed-methods study identifying common amendment reasons and avoidability patterns.	✗
AMEND (ours)	https://clinicaltrials.gov	161970	Prepared an ML-ready benchmark dataset containing all versions of eligibility criteria and amendment labels.	✓
AMEND_LLM (ours)	https://clinicaltrials.gov	64641	Subset of AMEND dataset with denoised amendment labels using LLMs.	✓

the study population and are among the most frequently amended components of trial protocols, making the initial EC text a strong signal for predicting future eligibility-criteria amendments. Automatically identifying EC at elevated risk of future amendments could help sponsors improve protocol quality, reduce avoidable revisions, and streamline trial execution.

To support research on clinical trial design, we introduce AMEND++, a benchmark suite for studying eligibility-criteria (EC) amendment prediction from initial trial protocols. AMEND++ includes two datasets capturing different trade-offs between scale and label fidelity: AMEND, a large-scale collection derived from public protocol version histories, and AMEND_LLM, a refined subset emphasizing substantive eligibility changes. Using this benchmark, we show that the information of initial eligibility criteria contains predictive signals of future amendment risk.

We make three primary contributions:

1. We formulate *eligibility criteria amendment prediction* as a new NLP task grounded in real-world clinical trial design, where the objective is to forecast whether an initial protocol will later undergo eligibility-related amendments.

2. We release AMEND++, a benchmark suite consisting of two large-scale datasets: AMEND, which provides eligibility criteria version histories with amendment labels derived from public trial records, and AMEND_LLM, a refined subset constructed using an LLM-based label denoising algorithm. This algorithm decomposes EC amendments into added, removed, and modified components to isolate substantive eligibility changes, resulting in high-quality amendment labels with substantially higher agreement with human annotations than alternative labeling approaches.

3. We propose *Change-Aware Masked Language Modeling* (CAMLM), a revision-aware pretraining strategy that leverages historical eligibility-criteria edits to learn amendment-sensitive representations. Through extensive benchmarks and ablations, we show that CAMLM consistently improves amendment prediction across multiple datasets, backbone encoders, and downstream classifiers.

2 Related Work

In recent years, a variety of machine learning and deep learning approaches have been

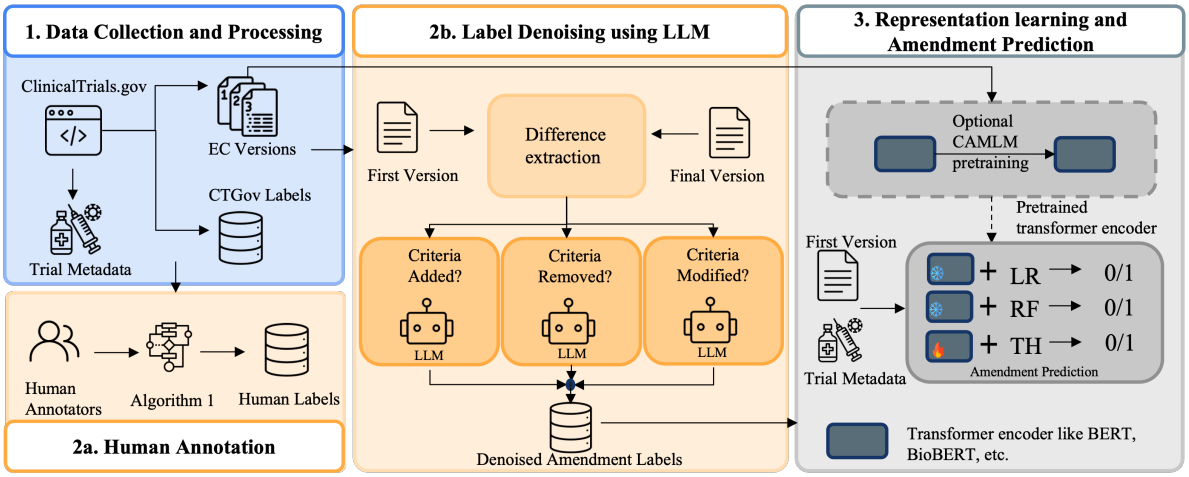

Figure 1: Overview of the proposed framework for clinical trial eligibility criteria (EC) amendment prediction, from data collection to downstream task application. LR: logistic regression, RF: random forest, TH: trial head used for predicting amendment. Human annotation is performed only on the test data.

applied throughout the clinical-trial lifecycle, including outcome prediction (Gao et al., 2025; Fu et al., 2022), generation of digital twins or simulated trial participants (Das et al., 2023, 2025a), and trial-document similarity search (Das et al., 2025b; Luo et al., 2024). These studies demonstrate the growing potential of data-driven methods to support trial design, analysis, and operational decision-making.

Few NLP systems have focused on extracting and normalizing eligibility criteria into structured representations, enabling cohort querying and downstream analysis but not modeling protocol evolution or amendment risk (Tseo et al., 2020; Yuan et al., 2019). More recently, multi-task benchmarks such as TrialBench have released standardized datasets spanning trial design and prediction tasks, including eligibility-criteria design, but do not consider revision-aware pretraining or forecasting protocol amendments from initial drafts (Chen et al., 2024).

In contrast, research on protocol amendments has been primarily descriptive and statistical. Prior work has characterized the frequency, causes, and operational burden of amendments (Getz et al., 2016; Botto et al., 2024; Joshi, 2023), but none have framed amendment forecasting as a predictive modeling problem or released ML-ready datasets. To our knowledge, no machine learning study has examined whether the text of initial protocols can predict future amendments. A closely related line of research examines the downstream statistical consequences of EC amendments. One study proposes conducting separate hypothesis tests for participants enrolled before and after an EC amendment and combining them using Fisher's combination test, highlighting that population shifts induced by amendments can meaningfully affect inference (Lösch and Neuhäuser, 2008). While this work focuses on post hoc analysis rather than prediction, it reinforces the broader importance of EC amendments and the need for methodological tools that account for changes in study populations.

3 Methods

Our pipeline (Figure 1) builds the AMEND++ benchmark from longitudinal eligibility-criteria histories, combines human annotation and LLM-based denoising for label construction, formulates amendment prediction as a supervised task, and introduces CAMLM for revision-aware pretraining.

3.1 Data Collection and Processing

We constructed AMEND++, a benchmark suite of interventional drug trials curated from *ClinicalTrials.gov*. We first filtered for pharmacological interventional studies and programmatically scraped the *Record History* tab using `BeautifulSoup` to reconstruct longitudinal EC

version histories, which are otherwise unavailable via standard API or direct downloads.

The resulting suite comprises two datasets: `AMEND`, a large-scale collection capturing raw EC version histories and amendment labels directly from public records; and `AMEND_LLM`, a refined subset where labels are denoised via an LLM-based pipeline to isolate substantive changes from minor administrative edits (e.g., formatting or typos). For each trial, we also extracted baseline metadata (e.g., titles, descriptions, disease indications, investigational drugs, MeSH terms, and study phases, etc.) to provide a comprehensive feature set for amendment prediction.

After scraping, we saved each version separately, preserving the full content and order of the original ECs. Disease names were standardized using MeSH terms to allow consistent comparisons across trials. For label denoising (see Section 3.3), we only used the first and final EC versions. However, all versions are released for each clinical trial with EC amendments.

3.2 Task Definition: Eligibility Criteria Amendment Prediction

We formulate *eligibility criteria amendment prediction* as a supervised binary classification task. For each trial i, the model takes as input the initial eligibility criteria text $E_i^{(0)}$ and trial-level metadata M_i available at trial initiation, and predicts $\hat{y}_i \in \{0, 1\}$, where $\hat{y}_i = 1$ indicates that the eligibility criteria are amended in later protocol versions and $\hat{y}_i = 0$ otherwise.

3.3 Label Denoising

Human Annotation

To generate high-quality labels for EC amendments, we relied on human annotation only for the test split. For these trials, human annotators examined the first and final versions of the ECs and applied a set of systematic rules (Algorithm 1) to determine whether a substantive amendment had occurred. This split serves as a gold-standard evaluation set.

Label Denoising using LLM

For the training and validation splits, large-scale manual annotation is prohibitively costly. Instead, we generate amendment labels using large language models (LLMs), resulting in the `AMEND_LLM` dataset, a refined subset of the `AMEND` dataset.

We decompose amendment detection into three complementary, criterion-level checks, each handled by a separate LLM instance:

1. **Criteria Added**: identifying whether any new inclusion or exclusion criteria are introduced in the final version.

2. **Criteria Removed**: detecting whether any criteria present in the initial version are absent from the final version.

3. **Criteria Modified**: identifying modifications to existing criteria that alter the eligible population, such as changes in numeric thresholds, medical entities, logical operators, or temporal constraints.

This decomposition allows each LLM to focus on a single, well-defined type of change, reducing ambiguity and improving robustness compared to a single, monolithic amendment judgment.

For each check, we use a 1-shot prompting strategy, where the LLM is provided with a single illustrative example followed by the target input. To reduce context length and focus the LLMs on semantically meaningful changes, we compute textual differences between the first and final EC versions and provide only the extracted differences as input to the LLM. Instructions provided to the LLM also abide by the rules in Algorithm 1. The prompts for LLMs are available in the Appendix (see prompts 4, 5, 6).

Formally, for a given trial i, let $E_i^{(0)}$ and $E_i^{(T)}$ denote the initial and final EC texts, respectively. We compute the differences:

$$\Delta E_i = \textsc{Diff}(E_i^{(0)}, E_i^{(T)}), \tag{1}$$

which serves as the LLM input for all amendment checks. Each LLM check produces a binary output:

$$y_i^{(k)} = f_k(\Delta E_i), \quad k \in \{\text{add}, \text{remove}, \text{modify}\}, \tag{2}$$

where $f_k(\cdot)$ denotes the LLM corresponding to the k-th amendment type.

The final amendment label is obtained by aggregating the outputs of the three checks

using a logical OR:

$$y_i = \mathbb{I}\left(\bigvee_k y_i^{(k)} = 1\right), \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Based on validation against human-annotated test data, the LLM-generated labels show strong agreement with human judgments and are therefore used for the training and validation splits of AMEND_LLM. This strategy enables scalable and consistent denoising of amendment labels while reserving human annotation exclusively for evaluation.

3.4 Change-Aware Masked Language Modeling (CAMLM)

Standard masked language modeling (MLM) treats all tokens in a document as equally informative during pretraining. In contrast, EC amendments in clinical trial protocols are typically sparse and localized: only specific criteria are added, removed, or modified across versions. To exploit this structure, we propose a pretraining strategy, *Change-Aware Masked Language Modeling* (CAMLM), that biases representation learning toward historically unstable regions of EC text. CAMLM does not introduce a new model architecture; instead, it modifies the masking policy used during MLM pretraining by leveraging historical versions of EC. The objective is to encourage the encoder to learn representations that are sensitive to content that has been revised in prior protocols, which may signal fragility or ambiguity in the initial trial design.

For each clinical trial i with recorded EC version history, we construct ordered version pairs $\left(E_i^{(t)}, E_i^{(T)}\right)$, where $E_i^{(t)}$ denotes an earlier EC version ($t < T$) and $E_i^{(T)}$ denotes the final EC version. If multiple historical versions are available, each earlier version is paired with the final version, yielding multiple training instances that capture long-horizon protocol evolution. For trials without intermediate versions, we construct a single pair $\left(E_i^{(0)}, E_i^{(T)}\right)$, allowing both amended and non-amended trials to contribute to pretraining.

Given a version pair $\left(E_i^{(t)}, E_i^{(T)}\right)$, we compute a token-level diff

$$\Delta E_i^{(t)} = \textsc{Diff}\left(E_i^{(t)}, E_i^{(T)}\right), \tag{4}$$

which identifies EC content that is deleted or replaced in the final version. Tokens in $E_i^{(t)}$ that correspond to deletions or substitutions in $\Delta E_i^{(t)}$ are treated as *unstable spans*, representing eligibility criteria that were ultimately modified or removed.

Let $x_i^{(t)} = (x_{i,1}^{(t)}, \ldots, x_{i,T_i}^{(t)})$ denote the token sequence of $E_i^{(t)}$. During pretraining, CAMLM applies a span-aware masking strategy in which tokens within unstable spans are masked with a high probability p_{span}, while tokens outside these spans are masked with a lower background probability p_{low}. The model is trained using the standard MLM objective:

$$\mathcal{L}_{\text{CAMLM}} = -\sum_{(i,t)} \sum_{m \in \mathcal{M}_i^{(t)}} \log p\left(x_{i,m}^{(t)} \mid x_{i,\setminus m}^{(t)}\right), \tag{5}$$

where $\mathcal{M}_i^{(t)}$ denotes the set of masked token positions for trial i and version t.

After CAMLM pretraining, the resulting encoder can be used either as a frozen feature extractor or fine-tuned end-to-end for eligibility-criteria amendment prediction as defined in Section 3.2. Because CAMLM introduces no additional parameters and makes no architectural changes, it can be applied directly to existing transformer encoder models like BERT, BioBERT, etc. In Section 4.3, we show that CAMLM consistently improves amendment prediction performance across multiple model classes and datasets, with gains that persist under extensive ablation studies.

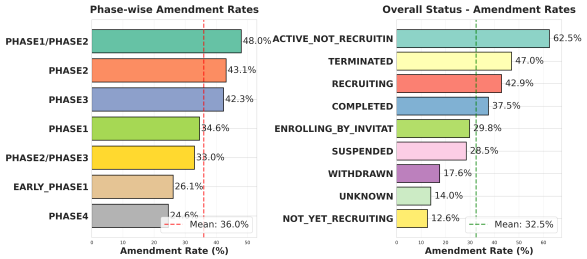

Figure 2: Clinical trial eligibility criteria amendment rates across phases and trial status in the AMEND_LLM dataset (n=64,641) dataset. Dashed lines indicate mean amendment rates for each dimension.

Table 2: Dataset statistics for AMEND and AMEND_LLM.

Property	AMEND	AMEND_LLM
Train/Validation/Test Sizes		
Train trials (1 / 0)	140,312 (52,071 / 88,241)	49,678 (17,178 / 32,500)
Validation trials (1 / 0)	15,591 (5,846 / 9,745)	8,896 (3,240 / 5,656)
Test trials (1 / 0)	6,067 (2,786 / 3,281)	6,067 (2,786 / 3,281) (same as AMEND)
Label Sources		
Train + validation labels	CTGov labels	LLM-denoised amendment labels
Test labels	Human annotations	Human annotations
Amendment Count Statistics (amended trials only)		
Min / Avg / Median / Max	1 / 2.92 / 2 / 38	1 / 3.07 / 2 / 38

Table 3: Performance comparison of different models on the AMEND dataset for eligibility criteria amendment prediction. Results are reported as mean ± standard deviation over multiple runs. **Bold** indicates the best score for each metric, and underlining marks the better-performing model between CAMLM and its corresponding non-CAMLM variant.

Model	AUROC	AUPRC	Accuracy
BioBERT + LR	0.676 ± 0.007	0.609 ± 0.010	0.629 ± 0.006
BioBERT_CAMLM + LR	0.689 ± 0.007	0.637 ± 0.010	0.643 ± 0.006
BioBERT + RF	0.683 ± 0.007	0.616 ± 0.011	0.624 ± 0.006
BioBERT_CAMLM + RF	0.684 ± 0.007	0.633 ± 0.010	0.633 ± 0.006
Finetuned BioBERT	0.704 ± 0.006	0.656 ± 0.009	0.631 ± 0.005
Finetuned BioBERT_CAMLM	$\mathbf{0.714 \pm 0.007}$	$\mathbf{0.670 \pm 0.009}$	$\mathbf{0.659 \pm 0.006}$

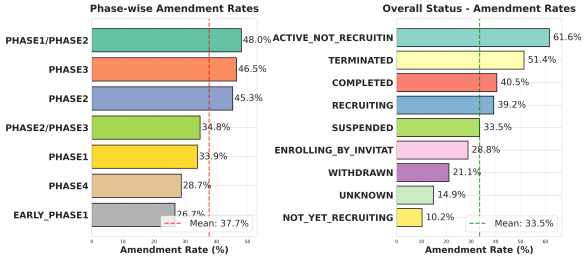

Figure 3: Clinical trial eligibility criteria amendment rates across phases and trial status in the AMEND (n=161,970) dataset. Dashed lines indicate mean amendment rates for each dimension.

4 Experiments and Results

4.1 Summary of Datasets

For this work, we focused exclusively on interventional drug trials. After filtering out trials with missing values, the AMEND_LLM dataset contains 49,678 training, 8,896 validation, and 6,067 test trials, while the larger AMEND dataset includes 140,312 training, 15,591 validation, and the same 6,067 test trials. Both datasets share the test split to enable fair evaluation against human-labeled annotations. In addition to multiple versions of EC, both datasets provide rich trial-level metadata (i.e., disease area, intervention drugs, study phase, trial titles) that can serve as additional input features for amendment prediction models, helping to improve performance and contextual understanding. Detailed statistics of the datasets are shown in Table 2. Figure 2 and Figure 3 show EC amendment rates across phases and trial status in AMEND_LLM and AMEND datasets respectively.

4.2 Implementation Details

For LLM-based denoising, we utilized the Qwen/Qwen2.5-7B-Instruct-1M model. All generations were performed with a temperature of 0.1 to ensure deterministic and stable outputs. The exact prompts used for denoising are provided in Appendix 5. For extracting differences between two versions of EC, we utilized difflib python package (Foundation, 2025). For benchmarking experiments, we employed BioBERT using the dmis-lab/biobert-base-cased-v1.1 checkpoint as the backbone encoder. We also used BERT and longformer

Table 4: Performance comparison of different models on the AMEND_LLM dataset for eligibility criteria amendment prediction. Results are reported as mean ± standard deviation over multiple runs. **Bold** indicates the best score for each metric, and underlining marks the better-performing model between CAMLM and its corresponding non-CAMLM variant.

Model	AUROC	AUPRC	Accuracy
BioBERT + LR	0.681 ± 0.007	0.610 ± 0.011	0.633 ± 0.006
BioBERT_CAMLM + LR	0.691 ± 0.007	0.632 ± 0.010	0.639 ± 0.006
BioBERT + RF	0.676 ± 0.007	0.603 ± 0.011	0.617 ± 0.006
BioBERT_CAMLM + RF	0.686 ± 0.007	0.628 ± 0.010	0.633 ± 0.006
Finetuned BioBERT	0.687 ± 0.007	0.621 ± 0.010	0.619 ± 0.006
Finetuned BioBERT_CAMLM	$\mathbf{0.697 \pm 0.006}$	$\mathbf{0.634 \pm 0.010}$	$\mathbf{0.644 \pm 0.006}$

Table 5: Ablations on pretraining strategies on AMEND_LLM. Results shown for BioBERT + LR model.

Method	AUROC	AUPRC	Accuracy
No pretraining	0.681 ± 0.007	0.610 ± 0.011	0.633 ± 0.006
MLM	0.687 ± 0.007	0.618 ± 0.010	0.635 ± 0.006
Span MLM	0.688 ± 0.007	0.628 ± 0.010	0.636 ± 0.006
CAMLM	$\mathbf{0.691 \pm 0.007}$	$\mathbf{0.632 \pm 0.010}$	$\mathbf{0.639 \pm 0.006}$

for showing results using general and long context encoders respectively in the Appendix. All evaluation metrics are reported using non-parametric bootstrap resampling with 1,000 iterations over the test set, where we report the mean and standard deviation. Additional implementation details, including training configurations, hyperparameters and feature importance analysis are provided in Appendix.

4.3 Eligibility Criteria Amendment Prediction

We evaluate *eligibility criteria amendment prediction*, where models use the initial EC and other trial metadata available at trial initiation to predict whether EC will be amended later.

As shown in Table 3 and Table 4, frozen-embedding baselines achieve moderate performance on both datasets, indicating that static representations of initial eligibility criteria already capture some amendment-relevant signals. Among these baselines, BioBERT + Logistic Regression (LR) and BioBERT + Random Forest (RF) exhibit comparable performance when representations are fixed.

End-to-end fine-tuning with a BioBERT classifier consistently outperforms frozen-embedding baselines across both datasets, highlighting the importance of task-specific representation learning. While absolute performance is higher on the full AMEND dataset in most of the cases, incorporating CAMLM yields consistent gains across all model classes and metrics. For the fine-tuned models, CAMLM improves AUROC by approximately 1.4–1.5%, AUPRC by about 2.1%, and accuracy by 4.0–4.4% relative on both datasets. [3] Notably, these improvements are observed both on AMEND and AMEND_LLM.

We evaluated the statistical significance of the AUROC differences using the DeLong's test. As shown in Table 9, CAMLM pretraining yields statistically significant improvements over standard BioBERT for logistic regression and fine-tuned BioBERT models on both AMEND and AMEND_LLM ($p < 0.05$). No significant difference is observed for Random Forest on AMEND, while a significant gain is observed in AMEND_LLM. As shown in Appendix Table 10, CAMLM pretraining consistently improves AUROC, AUPRC, and accuracy across different backbones (BERT, Longformer) and classifiers (LR, RF), demonstrating its generalizability beyond BioBERT. We focus on BioBERT-based baselines in the main paper because BioBERT is pretrained on biomedical text and consistently outperforms general-

[3] Relative improvements are computed as (CAMLM − baseline)/baseline × 100.

domain BERT and long-context Longformer models (see Appendix).

4.4 Ablations

To evaluate the impact of revision-aware pretraining, we compare four settings. (1) No pretraining: BioBERT is the off-the-shelf encoder. (2) MLM: further pretrains BioBERT with standard masked language modeling on the first EC version, capturing in-domain terminology but ignoring protocol evolution. (3) Span-aware MLM: masks spans corresponding to textual changes between consecutive EC versions while lightly masking unchanged tokens, capturing short-range revision patterns. (4) CAMLM: extends this to long-horizon modeling by pairing intermediate EC versions with the final protocol, focusing on regions predictive of eventual amendments. Table 5 shows that MLM on the first EC provides modest gains, span-aware MLM improves further by leveraging local changes, and CAMLM consistently outperforms all baselines, demonstrating that modeling EC evolution yields richer representations for amendment prediction than static text alone.

4.5 Validity and Impact of LLM-Denoised Amendment Labels

Given the substantially higher agreement between our LLM-generated labels and human annotations compared to both raw CTGov-derived labels and a baseline chain-of-thought (CoT) approach (Table 6), we directly use our LLM-generated amendment labels for the training and validation splits of the `AMEND_LLM` dataset. Unlike the baseline CoT method, which treats amendment detection as a single-step prediction problem, our approach decomposes eligibility-criteria amendments into added, removed, and modified components, enabling more precise isolation of substantive eligibility changes. This structured decomposition yields markedly improved label fidelity while avoiding the prohibitive cost of large-scale human annotation, allowing us to scale reliable supervision beyond what manual labeling alone could support. To ensure a high-quality evaluation benchmark, the test split was independently annotated by two clinical trial protocol experts (Algorithm 1), with disagreements resolved through consensus discussion.

Table 6: Agreement with Human-Annotated EC Amendment Labels (N=6067)

Method	Mismatches	Match Rate
CTGov Label	230	96.21%
Baseline CoT	257	95.76%
Our Approach	**40**	**99.34%**

Table 7: Performance comparison of models trained on raw CTGov labels vs. LLM-denoised amendment labels, evaluated on human-labeled test data. These are BioBERT+LR results for showing the utility of the denoising step. Models are trained on `AMEND_LLM` train split.

Trained On	AUROC	AUPRC
CTGov labels	0.672 ± 0.007	0.606 ± 0.011
LLM labels	$\mathbf{0.681 \pm 0.007}$	$\mathbf{0.610 \pm 0.011}$

Table 7 shows that replacing noisy CTGov[4] labels with LLM-denoised labels consistently improves all evaluation metrics when tested against human annotations. The training and validation trials are from `AMEND_LLM` as those have both CTGov labels and LLM-denoised labels. The test is the same test set annotated by human. This shows that denoising substantially enhances the quality of the label and yields better prediction performance.

5 Conclusion

We introduce *eligibility criteria amendment prediction*, a novel NLP task addressing the frequent and impactful problem of clinical trial protocol amendments. To support this task, we release `AMEND++`, a benchmark suite comprising two datasets: `AMEND`, which captures eligibility-criteria version histories and amendment labels from public clinical trials, and `AMEND_LLM`, a refined subset curated using an LLM-based denoising pipeline to isolate substantive changes. In addition, we propose CAMLM, a revision-aware pretraining strategy that takes advantage of historical edits to learn amendment-sensitive representations. Experiments demonstrate that CAMLM consistently improves prediction performance across datasets and architectures. Together, our datasets and pretraining approach offer a scalable framework for anticipating protocol amendments, enabling more efficient, cost-effective, and robust clinical trial design.

[4]CTGov is short for *ClinicalTrials.gov*.

Limitations

While we are the first to construct a machine-learning–ready dataset for protocol amendments, our work focuses exclusively on EC. EC is one of the most frequently amended and conceptually complex protocol sections, making it a natural starting point. However, other sections (such as outcomes, study design elements, and interventions) are also commonly amended, and extending our dataset and methods to these areas is an important direction for future work. A second limitation is our assumption that *ClinicalTrials.gov* labels indicating no eligibility-criteria amendment (label 0) are correct. We do not explicitly account for potential false negatives, which are expected to be rare in curated trial registries but could introduce residual label noise.

References

Emily Botto, Zachary Smith, and Kenneth Getz. 2024. New benchmarks on protocol amendment experience in oncology clinical trials. *Therapeutic Innovation & Regulatory Science*, 58(4):645–654.

Jintai Chen, Yaojun Hu, Mingchen Cai, Yingzhou Lu, Yue Wang, Xu Cao, Miao Lin, Hongxia Xu, Jian Wu, Cao Xiao, and 1 others. 2024. Trialbench: Multi-modal artificial intelligence-ready clinical trial datasets. *arXiv preprint arXiv:2407.00631*.

Ton J Cleophas, Aeilko H Zwinderman, Toine F Cleophas, and Eugene P Cleophas. 2009. *Statistics applied to clinical trials*. Springer.

Trisha Das, Afrah Shafquat, Mandis Beigi, Jacob Aptekar, Jason Mezey, and Jimeng Sun. 2025a. Seqtrial: Utility preserving sequential clinical trial data generator. In *AMIA Annual Symposium Proceedings*, volume 2024, page 329.

Trisha Das, Afrah Shafquat, Mandis Beigi, Jacob Aptekar, and Jimeng Sun. 2025b. *secret*: Semi-supervised clinical trial document similarity search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5278–5291, Vienna, Austria. Association for Computational Linguistics.

Trisha Das, Zifeng Wang, and Jimeng Sun. 2023. Twin: Personalized clinical trial digital twin generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 402–413.

Python Software Foundation. 2025. The 'difflib' module — 'difflib' — python 3.12.0 documentation. https://docs.python.org/3/library/difflib.html. Accessed: 2025-11-25.

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).

Chufan Gao, Jathurshan Pradeepkumar, Trisha Das, Shivashankar Thati, and Jimeng Sun. 2025. Automatically labeling clinical trial outcomes: A large-scale benchmark for drug development. *Preprint*, arXiv:2406.10292.

Kenneth Getz, Zachary Smith, Emily Botto, Elisabeth Murphy, and Arnaud Dauchy. 2024. New benchmarks on protocol amendment practices, trends and their impact on clinical trial performance. *Therapeutic Innovation & Regulatory Science*, 58(3):539–548.

Kenneth A Getz, Stella Stergiopoulos, Mary Short, Leon Surgeon, Randy Krauss, Sybrand Pretorius, Julian Desmond, and Derek Dunn. 2016. The impact of protocol amendments on clinical trial performance and cost. *Therapeutic innovation & regulatory science*, 50(4):436–441.

Shivam Joshi. 2023. Common clinical trial amendments, why they are submitted and how they can be avoided: a mixed methods study on nhs uk sponsored research (amendments assemble). *Trials*, 24(1):10.

John A. Lewis. 1999. Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in medicine*, 18 15:1903–42.

Christian Lösch and Markus Neuhäuser. 2008. The statistical analysis of a clinical trial when a protocol amendment changed the inclusion criteria. *BMC medical research methodology*, 8(1):16.

Junyu Luo, Cheng Qian, Lucas Glass, and Fenglong Ma. 2024. Clinical trial retrieval via multi-grained similarity learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2950–2954.

Yitong Tseo, MI Salkola, Ahmed Mohamed, Anuj Kumar, and Freddy Abnousi. 2020. Information extraction of clinical trial eligibility criteria. *arXiv preprint arXiv:2006.07296*.

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, 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.

Appendix A

Hyperparameter Tuning

For the BioBERT + LR baseline, we tuned C (inverse of the regularization strength) from [0.01, 0.1, 1, 10, 100]. For the BioBERT + RF baseline, we tuned n_estimators from [100, 200, 300] and max_depth from [5, 10]. We did not let any weight updates of BioBERT for these baselines. For finetuning the BioBERT model, we used 10 epochs for training. We selected a model based on the validation AUROC evaluated after each epoch.

For CAMLM, MLM and Span MLM, we trained BioBERT for 3 epochs. For CAMLM pretraining, we use BioBERT's WordPiece tokenizer. Change spans are computed at the token level using `difflib.SequenceMatcher` between the tokenized initial and final EC versions, where `replace` and `delete` operations define unstable spans. We use a *change-aware* masked language modeling (MLM) objective that emphasizes amendment-prone regions, masking tokens within detected change spans with probability $p_{\text{span}} = 0.8$ and masking remaining tokens with a lower background probability $p_{\text{low}} = 0.05$. All masked tokens are replaced with the standard `[MASK]` token, and supervision is provided only at masked positions following standard MLM training.

Feature Importance

We tried to understand feature importance on the BioBERT + LR model. We assume that the findings from the results will be similar for other models too. Table 8 shows that using the first version of EC, disease/condition, intervention and phase combined gives the highest scores.

LLM Prompts

Figure 6, figure 5 and figure 4 are prompts used for denoising amendment labels.

Usage of AI Assistants

We utilized chatgpt for paper drafting (e.g., summarizing, paraphrasing, etc.) assistance.

Prompt for "Criteria Removed" Task

```
Below is an example diff between two
versions of eligibility criteria (EC), with
correct reasoning.

{example}

Now apply the same reasoning to the new case
below.
Here are the differences between two
versions of eligibility criteria:

{difference}

Your thoughts:
You are a clinical trial analyst. Think
step-by-step about the differences between
two versions of eligibility criteria (EC)
and determine if any existing criteria have
been removed.

Step-by-step Instructions:
- Read all lines starting with '-' as a
  single block. These represent the original
  eligibility criteria.
- Read all lines starting with '+' as a
  single block. These represent the updated
  eligibility criteria.
- Compare the meaning of the '-' block with
  the '+' block as a whole.
- Ignore formatting changes (e.g., line
  breaks, indentation) and rewordings that
  preserve the original meaning.
- Focus only on whether the '-' block
  contains any criteria that are completely
  missing in the '+' block.
- If any requirement, rule, or constraint
  was present in the '-' block but is no
  longer found in the '+' block, then a
  criterion has been removed.

If at least one criterion has been removed:
Final output: 1
If no criteria have been removed:
Final output: 0
Think step-by-step.
Your final answer must be a single line at
the end in the following format
(case-sensitive):
Final output: 0
or
Final output: 1
```

Figure 4: Prompt given to the LLM for the "Criteria Removed" task.

Prompt for "Criteria Modified" Task

```
Below is an example diff between two versions of eligibility criteria (EC), with correct
reasoning.

{example}

Now apply the same reasoning to the new case below.
Here are the differences between two versions of eligibility criteria:

{difference}

Your thoughts:
You are a clinical trial analyst. Your task is to detect whether any existing eligibility
criteria have been *modified* in a way that changes who is eligible. You are not responsible for
detecting new criteria or removed criteria - those are handled separately.

Instructions:
- Treat '-' lines as the original version and '+' lines as the updated version of the *same*
  criterion.
- Compare carefully for semantic differences.

Count as meaningful changes (Final output: 1):
- Numeric changes (age limits, lab thresholds, dosage values).
- Severity qualifiers (e.g., "liver disease" -> "severe liver disease").
- Logical operator changes (e.g., "and" <-> "or").
- Added or removed biomedical entities within the same criterion:
   * Drugs (e.g., "warfarin" -> "warfarin or heparin").
   * Diseases/conditions (e.g., "HIV" -> "HIV or HBV").
   * Enzymes/transporters/biomarkers (e.g., "CYP3A4" -> "CYP3A4 or P-gp").
- Time window changes (e.g., "within 6 months" -> "within 3 months").
- Condition scope changes (e.g., "cancer" -> "breast cancer", "diabetes" -> "type 2 diabetes").
- Entity substitutions (e.g., "rituximab" -> "adalimumab").
Ignore as non-meaningful (Final output: 0):
- Rearrangement of criteria or bullet points.
- Capitalization differences (e.g., "BOTH" -> "both").
- Formatting or indentation changes.
- Punctuation only (commas, semicolons, periods).
- Abbreviation expansion or contraction (e.g., "ART" <-> "antiretroviral therapy (ART)", "CYP3A4"
  <-> "cytochrome P450 3A4").
- Synonyms with no scope change (e.g., "high blood pressure" <-> "hypertension").
- Spelling corrections or typos (e.g., "diabtes" -> "diabetes").
- Minor stylistic rephrasing without meaning change (e.g., "patients who have" -> "patients
  with").
- Splitting or merging sentences without altering meaning.
- Reordering of biomedical terms without additions/removals (e.g., "HIV, HBV, HCV" <-> "HBV, HCV,
  HIV").

Output format (case sensitive):
- If at least one meaningful modification is found:
  Final output: 1
- If no meaningful modification is found:
  Final output: 0
```

Figure 5: Prompt given to the LLM for the "Criteria Modified" task.

Prompt for "Criteria Added" Task

```
Below is an example diff between two versions of eligibility criteria (EC), with correct
reasoning.

{example}

Now apply the same reasoning to the new case below. Here are the differences between two versions
of eligibility criteria:

{difference}

Your thoughts:
You are a clinical trial analyst. Think step-by-step about the differences between two versions
of eligibility criteria (EC) and determine if any new criteria have been added.

Step-by-step Instructions:
- Read all lines starting with '-' as a single block. These represent the original eligibility
  criteria.
- Read all lines starting with '+' as a single block. These represent the updated eligibility
  criteria.
- Compare the meaning of the '+' block with the '-' block as a whole.
- Ignore formatting changes (e.g., line breaks, indentation) and rewordings that preserve the
  original meaning.
- Focus only on whether the '+' block introduces any new criterion that is not already present in
  the '-' block.
- If any new requirement, rule, or constraint appears in the '+' block that is absent from the '-'
  block, it is a new criterion and constitutes an amendment.

If at least one new criterion is added:
Final output: 1
If no new criteria are added:
Final output: 0
Think step-by-step.
Your final answer must be a single line at the end in the following format (case-sensitive):
Final output: 0
or
Final output: 1
```

Figure 6: Prompt given to the LLM for the "Criteria Added" task.

Table 8: Feature Importance Exploration

Features Used	AUROC	AUPRC	Accuracy	F1
First_EC	0.675 ± 0.007	0.607 ± 0.011	0.630 ± 0.006	0.563 ± 0.008
First_EC, disease	0.676 ± 0.007	0.609 ± 0.011	0.630 ± 0.006	0.563 ± 0.008
First_EC, disease, intervention	0.680 ± 0.007	0.606 ± 0.011	0.633 ± 0.006	0.567 ± 0.008
First_EC, disease, intervention, phase	0.681 ± 0.007	0.610 ± 0.011	0.633 ± 0.006	0.567 ± 0.008

Algorithm 1: Human Annotation Procedure for Determining EC Amendments

Input: First version of eligibility criteria EC_{first},
Final version of eligibility criteria EC_{final},
`ctgov_label` $\in \{0, 1\}$

Output: `amendment` $\in \{0, 1\}$

if `ctgov_label` $= 0$ **then**
  return 0 `// Use ctgov label as ground truth`

Initialize: `amendment` $\leftarrow 0$

Rule 1: Addition Check
if *there exists a criterion in* EC_{final} *not present in* EC_{first} **then**
  `amendment` $\leftarrow 1$

Rule 2: Removal Check
if *there exists a criterion in* EC_{first} *not present in* EC_{final} **then**
  `amendment` $\leftarrow 1$

Rule 3: Significant Modification Check
if *any criterion is modified such that:* **then**
  numeric thresholds change **or** medical entities are added/removed **or**
  the meaning of the criterion changes or the affected population differs
`amendment` $\leftarrow 1$

Ignored non-amendments:
Formatting differences (bullets, numbering, indentation)
Minor spelling or grammar corrections
Synonyms or wording changes with preserved meaning
Abbreviation expansion or contraction (e.g., ECG → electrocardiogram)
Reordering of criteria without semantic change
Text added or removed that does not affect eligibility or population
etc

return `amendment`

Comparison	**AMEND**		**AMEND_LLM**	
	p-value	Sig.	p-value	Sig.
BioBERT (LR) vs. BioBERT-CAMLM (LR)	3.24×10^{-3}	✓	3.03×10^{-2}	✓
BioBERT (RF) vs. BioBERT-CAMLM (RF)	9.24×10^{-1}	–	4.87×10^{-3}	✓
Finetuned BioBERT vs. Finetuned BioBERT-CAMLM	1.23×10^{-3}	✓	5.05×10^{-3}	✓

Table 9: Statistical significance of AUROC differences between BioBERT and BioBERT-CAMLM variants, evaluated using DeLong's test.

Table 10: Performance comparison of all baseline models with and without CAMLM on the AMEND_LLM dataset. Results are reported as mean ± standard deviation over multiple runs. **Bold** indicates the best score for each metric, and underlining marks the better-performing model between CAMLM and its corresponding non-CAMLM variant.

Model	**AUROC**	**AUPRC**	**Accuracy**
BERT + LR	0.669 ± 0.007	0.602 ± 0.010	0.616 ± 0.006
BERT_CAMLM + LR	0.675 ± 0.007	0.620 ± 0.010	0.624 ± 0.006
BERT + RF	0.660 ± 0.007	0.598 ± 0.010	0.605 ± 0.006
BERT_CAMLM + RF	0.674 ± 0.007	0.616 ± 0.010	0.618 ± 0.006
Longformer + LR	0.670 ± 0.007	0.611 ± 0.010	0.622 ± 0.006
Longformer_CAMLM + LR	0.674 ± 0.007	0.618 ± 0.010	0.626 ± 0.006
Longformer + RF	0.666 ± 0.007	0.602 ± 0.010	0.597 ± 0.006
Longformer_CAMLM + RF	0.673 ± 0.007	0.627 ± 0.010	0.619 ± 0.006
BioBERT + LR	0.681 ± 0.007	0.610 ± 0.011	0.633 ± 0.006
BioBERT_CAMLM + LR	0.691 ± 0.007	0.632 ± 0.010	0.639 ± 0.006
BioBERT + RF	0.676 ± 0.007	0.603 ± 0.011	0.617 ± 0.006
BioBERT_CAMLM + RF	0.686 ± 0.007	0.628 ± 0.010	0.633 ± 0.006
Finetuned BioBERT	0.687 ± 0.007	0.621 ± 0.010	0.619 ± 0.006
Finetuned BioBERT_CAMLM	$\mathbf{0.697 \pm 0.006}$	$\mathbf{0.634 \pm 0.010}$	$\mathbf{0.644 \pm 0.006}$