ClinicalReTrial: Clinical Trial Redesign with Self-Evolving Agents

Anonymous ACL submission

Abstract

Clinical trials constitute a critical yet exceptionally challenging and costly stage of drug development ($2.6B per drug), where protocols are encoded as complex natural language documents, motivating the use of AI systems beyond manual analysis. Existing AI methods predict trial failure accurately but offer no actionable remedies. To fill this gap, this paper proposes ClinicalReTrial, a multi-agent framework that formulates clinical trial optimization as an iterative redesign problem on textural protocols. Our method integrates failure diagnosis, safety-aware modifications, and candidate evaluation in a closed-loop, reward-driven optimization framework. Serving the outcome prediction model as a simulation environment, ClinicalReTrial enables low-cost evaluation and dense reward signals for continuous self-improvement. We further propose hierarchical memory capturing iteration-level feedback within trials and distill transferable redesign patterns across trials. Empirically, ClinicalReTrial improves 83.3% of trial protocols with a mean success probability gain of 5.7% with negligible cost ($0.12 per trial). Retrospective case studies demonstrate alignment between the discovered redesign strategies and real-world clinical trial modifications. The code is anonymously available at: https://anonymous.4open.science/r/ClinicalFailureReasonReTrial-9632.

1 Introduction

Clinical trials represent the most critical and expensive phase in drug discovery, with an estimated cost of $2.6 billion (DiMasi et al., 2016) per approved drug, and low success rates of approximately 10-20% (Yamaguchi et al., 2021). Serving as documented plan that specifies the study's objectives, clinical trial protocols involve complex, interdependent design choices (Getz and Campo, 2017) expressed in natural language documents, such as eligibility criteria, dosing strategies, and endpoint definitions, where small design flaws can propagate into irreversible failure. These challenges motivate the use of AI systems (Zhang et al., 2023) that can reason over high-dimensional trial designs, leverage historical evidence, and systematically assess failure risks at scale.

Recent advances in AI have enabled increasingly accurate prediction of clinical trial outcomes. For example, Lo et al. (2019) uses structured metadata to model success likelihood; Fu et al. (2022); Chen et al. (2024c, 2025) integrate heterogeneous data sources using architectures including graph neural networks and hierarchical attention mechanisms to predict trial approval rate; Yue et al. (2024); Liu et al. (2025) incorporate Large Language Models (LLMs) and external knowledge bases to enhance reasoning and explainability in trial outcome prediction through natural advanced language understandings of complex medical texts.

Despite their success, existing approaches are inherently reactive in nature: they operate on a fixed clinical trial protocol and produce a prediction or post-hoc explanation of trial success or failure. However, these methods do not address a more practically consequential problem: they are unable to respond to a determined trial failure due to the lack of actionable interventions. In real-world drug discovery, stakeholders require not only assessments of failure risk, but also actionable guidance on protocol redesign, including principled modifications or augmentations informed by the identified sources of risk (Dagenais et al., 2022; Baumfeld Andre et al., 2020).

To bridge this gap, in this work, we propose ClinicalReTrial, a self-evolving AI agent that moves beyond static prediction toward actionable intervention via end-to-end trial protocol optimization, while continuously improving its redesign policies. The inherently language-rich nature of clinical trial protocols makes this an ideal domain for LLM-based optimization. Our framework in-

stantiates a coordinated multi-agent pipeline that performs failure diagnosis, protocol redesign, and candidate evaluation, with domain knowledge and safety awareness embedded at each decision stage. Beyond a single optimization run, we adopt the prediction model as a simulation environment to provide rewards for continuous self-improvement. Specifically, `ClinicalReTrial` maintains local memory to accumulate iteration-level feedback and reward attributed modification outcomes for within-trial adaptation, while a global memory distills transferable redesign patterns across trials to enable warm start initialization and exploration calibration. Through this hierarchical learning structure and reward-driven closed-loop optimization, `ClinicalReTrial` systematically explores the protocol modification space and learns to identify high-impact interventions that improve clinical trial success probability.

Experimentally, our prediction model demonstrated the strongest performance (PR-AUC>0.75), allowing it to serve as a reliable simulation environment for evaluation and agent optimization. In the trial redesign experiments, `ClinicalReTrial` successfully improved 83.3% of trial protocols with mean probability gain $\Delta p = 5.7\%$, achieved at negligible cost ($\$0.12$/trial). We further conduct multiple real-world retrospective case studies. Impressively, the redesigns generated by `ClinicalReTrial` exhibit strategic alignment with independently derived real-world trial modifications, highlighting the potential of self-evolving AI agents to support principled trial redesign.

Main contributions are listed as follows: (1) (to the best of our knowledge) We are the first to formulate clinical trial optimization as an AI-solvable and *in silico*-verifiable problem. (2) We propose a multi-agent pipeline with domain knowledge that decomposes clinical trial protocol optimization into diagnosis, modification, and evaluation. (3) We develop a simulation-driven clinical trial optimization framework with *In-Context Learning* and multi-level memory for continuous self-improvement through reward attributed prompt optimization, dynamic redesign pool curation, and cross-trial knowledge distillation.

2 Related Work

Early efforts employed classical machine learning (logistic regression (LR), random forests) on expert-curated features (Gayvert et al., 2016; Lo et al., 2019), establishing feasibility but lacking multi-modal data integration. Deep learning approaches addressed this: Fu et al. (2022) proposed HINT, integrating drug molecules, ICD-10 codes, and eligibility criteria; Chen et al. (2024c) added uncertainty quantification; Wang et al. (2024) designed LLM-based patient-level digital twins; Chen et al. (2025) released a standardized TrialBench with multi-modal baselines. while maintaining competitive performance. Recent LLM approaches demonstrate medical reasoning (Singhal et al., 2023), enhanced via retrieval-augmented generation (Lewis et al., 2020) with databases like DrugBank (Wishart et al., 2018), Hetionet (Himmelstein et al., 2017), and domain-adapted encoders like BioBERT (Lee et al., 2020). Building on this, Yue et al. (2024) introduced ClinicalAgent, decomposing prediction into specialized sub-task agents with ReAct reasoning (Yao et al., 2023). Liu et al. (2025) proposed AutoCT for autonomous feature engineering via Monte Carlo Tree Search (Chi et al., 2024).

However, these methods are essentially predictive models without explaining *why* failures occur or *how* to modify protocols. First to formulate generative optimization, our multi-agent architecture leverages chain-of-thought (Wei et al., 2022), and least-to-most prompting (Zhou et al., 2023) for hierarchical problem decomposition.

3 Methodology

Overview. `ClinicalReTrial` is a self-improving multi-agent system that redesigns failed clinical trials through reward-driven iterative optimization. Specifically, we first formulate the clinical trial optimization problem in Section 3.1. Then, we build a multi-agent framework to address it in Section 3.2. To further improve our Agent performance, we integrate a knowledge retrieval system, and present In-Context Learning with multi-level memory in Section 3.3. For ease of exposition, Figure 1 illustrates the whole process and Algorithm 1 formalizes the iterative optimization procedure.

3.1 Problem: Clinical Trial Optimization

The goal of clinical trial is to evaluate the safety and efficacy of drug on patients. Clinical trial protocol, *e.g.,* eligibility criteria (to recruit patients), drug dosage, is designed to conduct clinical trial. Formally, let $T_0 = \{e_1, e_2, \ldots, e_K\}$ denote a clinical trial protocol decomposed into K modifiable elements (*e.g.*, eligibility criteria,

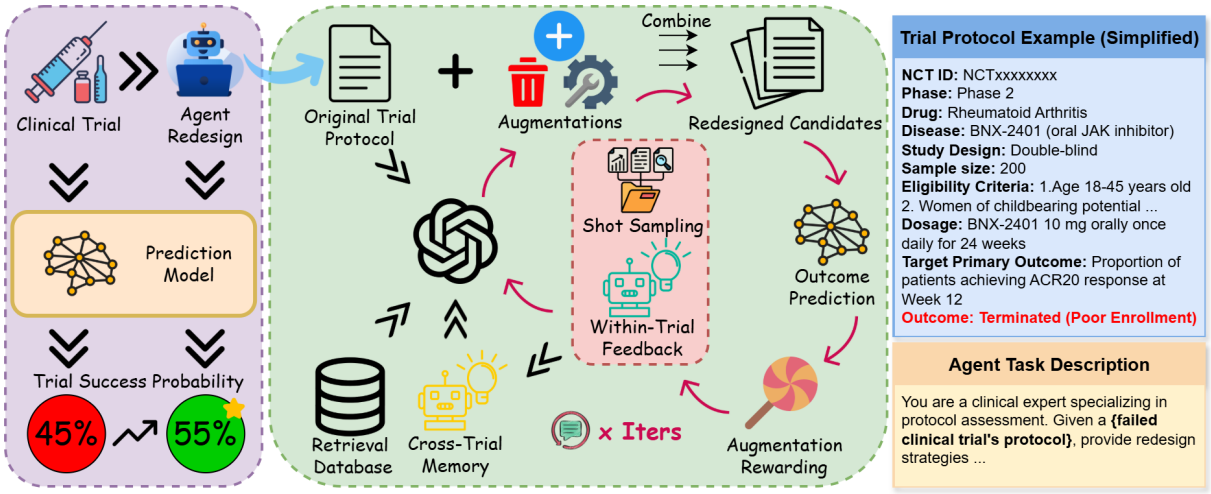

Figure 1: `ClinicalReTrial` Agent architecture. The system operates through iterative refinement: agents analyze failures, generate modifications, and receive rewards from the simulation environment. Historical explorations are extracted into structured knowledge that guides subsequent iterations, enabling progressive improvement.

Algorithm 1 Clinical trial protocol optimization with In-Context Learning and Multi-level Memory.

Require: Failed trial T_0, failure mode $y \in \{\text{enrollment, safety, efficacy}\}$, global memory $\mathcal{M}^{\text{global}}$
Ensure: Optimized protocol T^*, best reward r_{best}

Initialize: $r_{\text{best}} \leftarrow 0$, $T^* \leftarrow T_0$, $\mathcal{M}_0^{\text{local}} \leftarrow \emptyset$, $\mathcal{H}_0 \leftarrow \emptyset$
for $t = 1$ to $N_{\max}$ **do**
  $\mathcal{K}_t^s, \mathcal{K}_t^t \leftarrow$ **LoadMemory**$(\mathcal{M}^{\text{global}}[y], t)$
  $\mathcal{S}_t \leftarrow$ **AnalyzerAgent**$(T_{t-1}, y, \mathcal{M}_{t-1}^{\text{local}}, \mathcal{K}_t^s)$; $\mathcal{A}_t \leftarrow$ **GeneratorAgent**$(\mathcal{S}_t, T_{t-1}, \mathcal{M}_{t-1}^{\text{local}}, \mathcal{K}_t^t)$
  $\mathcal{R}_t, r_{\max} \leftarrow$ **ExploreSearch**$(\mathcal{A}_t, \mathcal{H}_{t-1}, T_{t-1})$
  if $r_{\max} > r_{\text{best}}$ **then** $r_{\text{best}} \leftarrow r_{\max}$, $T^* \leftarrow \arg\max_{T' \in \mathbb{T}} f_\theta(T')$
  $\mathcal{K}_t \leftarrow$ **DistillKnowledge**$(\mathcal{R}_t)$; $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup$ **ExtractPool**$(\mathcal{R}_t)$; $\mathcal{M}_t^{\text{local}} \leftarrow \mathcal{M}_{t-1}^{\text{local}} \cup \{\mathcal{K}_t, \mathcal{R}_t, \mathcal{H}_t\}$
end for
$\mathcal{M}^{\text{global}} \leftarrow \mathcal{M}^{\text{global}} \cup$ **TransferMemory**$(T^*, \mathcal{M}_{N_{\max}}^{\text{local}})$;
return T^*, r_{best}

dosage regimens, endpoint definitions), where a prediction model $f_\theta : \mathcal{T} \to [0, 1]$ assigns success probability $p_0 = f_\theta(T_0)$. Each element e_i, after redesign, admits a set of augmentations $\mathcal{A}_i = \{a_{i1}, a_{i2}, \ldots, a_{im_i}\}$ representing clinically valid modifications, where candidate protocols $\mathbb{T} = \{T'_1, T'_2, \ldots, T'_N\}$ are reconstructed by selectively replacing elements in T_0 with their augmented variants. Such an exploration set of redesigns $\mathbb{T}$ is evaluated through the prediction model to obtain success probabilities, with each $p'_j = f_\theta(T'_j)$. The optimization objective seeks the optimal protocol $T^* = \arg\max_{T' \in \mathbb{T}} f_\theta(T')$ with success probability p^*, where overall improvement is measured by $\Delta p = p^* - p_0$ with $p^* = f_\theta(T^*)$.

3.2 Multi-Agent Architecture

With heterogeneous necessities, clinical trial redesign requires distinct diagnostic expertise and remediation strategies (Fogel, 2018). We introduce a multi-agent architecture naturally align with this structure, comprising four coordinated components: multi-agent system (Chen et al., 2024b) for protocol optimization, `ClinicalReTrial` Agent comprises four coordinated components: Trial Failure Analyzer (§3.2.1) performs root cause diagnosis; the Protocol Refinement Generator (§3.2.2) synthesizes modifications; the Clinical Safety Validator (§3.2.3) prunes unsafe modifications; and the Protocol Candidate Evaluator (§3.2.4) provides simulation based feedback.

3.2.1 Protocol Diagnosis Analyzer

Given a failed protocol and prior failure modes: {POOR ENROLLMENT, SAFETY/ADVERSE EFFECT, DRUG LACK OF EFFICACY}, the Analyzer Agent produces a prioritized set of modifications based on Protocol Taxonomy and Action Determi-

nation, targeting the protocol feature under consideration; then specifies the action strategy {DELETE, MODIFY, ADD} with confidence score.

Protocol Taxonomy. Protocol features are classified based on medical terminology: eligibility criteria into participation barriers, safety exclusions, selection criteria, and enrichment criteria; dosage/outcomes by safety risk, failure contribution, and modification efficacy, guiding action selection given the observed failure reason and according analysis.

Action Determination. The agent extracts failure signatures via action category alignment and confidence scoring, calibrated using historical modification success patterns. At iteration $t = 1$, the agent receives warm start guidance from cross-trial memory (§3.3.3); from $t \geq 2$, it incorporates performance patterns from prior iterations (§3.3.2). These insights yield prioritized modification targets balancing domain knowledge with empirical feedback, which guide the Generator Agent (§3.2.2) to produce concrete modifications.

3.2.2 Protocol Refinement Generator

The Generator Agent translates diagnostic insights from the Analyzer Agent into diverse design refinements that address identified weaknesses while preserving clinical validity.

Action-specific Variant Generation. The agent employs action specific logic: DELETE critical failure factors while preserving safety; MODIFY adjusts thresholds or operationalizes vague terms; ADD introduces biomarker enrichment or contraindication criteria. Various modifications and augmentations first validated for clinical safety (§3.2.3) before proceeding to evaluation.

3.2.3 Clinical Safety Validator

To ensure the proposed augmentations satisfy clinical safety standards, the system employs *LLM-as-a-Judge* (Zheng et al., 2023) with domain database retrieval and autonomous validation.

Database Retrieval The multi-agent system enhances embedded parametric knowledge with targeted retrieval from biomedical databases: DrugBank (Wishart et al., 2018) with pharmacological profiles including toxicity, metabolism, contraindications; Disease Database (Chen et al., 2024a) that contains diagnostic criteria, symptomatology, risk factors; and PubMed Abstract spanning 1975-2025. Retrieval employs dense embeddings (BioBERT (Lee et al., 2020) for drugs/diseases, PubMedBERT (Gu et al., 2021) for literature) with FAISS indexing. Retrieved results have tangential content filtered, while enforcing strict temporal constraints that limit PubMed queries to prevent outcome leakage.

Autonomous Safety Validation. The agent checks and prunes unsafe modifications (dosage changes, population shifts, contraindications) triggering retrieval from external databases when identifying knowledge gaps medical assessments. Validated candidates that passed the validation proceed to the Evaluator (§3.2.4) for simulation-based evaluation and reward assignment.

3.2.4 Protocol Candidate Evaluator

The Evaluator combines validated augmentations with the original trial protocol into complete redesigned trial candidates, each evaluated through simulation-based assessment with outcome probability assigned that guides modification rewards and hierarchical learning (§3.3).

Search Strategy. Candidate trials are formed by combining validated augmentations across original protocols, where Beam Search (Freitag and Al-Onaizan, 2017) is used to reduce exponential complexity to approximately quadratic.

Simulation Environment. To provide reliable feedback for agent-generated modifications, we train model that predicts trial candidates' outcome probabilities from encoded trial features, serving as simulation environment that enables rapid evaluation for thousands of redesigns without conducting actual clinical trials, guiding the agent system toward promising protocol optimization. The improvement in predicted success probability, from the original trial to the redesigned candidate, serves as the reward signal.

3.3 In-Context Learning with Multi-level Memory

Without memory or iterative feedback, multi-agent systems repeat failed strategies and cannot leverage historical performance. We address this via In-Context Learning (Brown et al., 2020) from reward guided feedback (§3.3.1) with knowledge consolidation operating at two temporal scales: *within-trial learning* (§3.3.2) accumulates local memory across iterations for trial refinement, while *cross-trial learning* (§3.3.3) maintains global memory to transfer successful patterns across the trial corpus.

3.3.1 Redesign Reward

To identify specific modifications that drive improvement, we decompose protocol-level redesigns outcome probabilities $p(T')$ into augmentation-level rewards. The Evaluator first evaluates combined trial variants via prediction, then attributes credit to individual modifications. For each augmentation m, we compute its marginal contribution $r(m)$ across the explored combinatorial space:

$$r(m) = \mathbb{E}_{m \in T'}[p(T')] - \mathbb{E}_{m \notin T'}[p(T')]. \quad (1)$$

The complete reward distribution $\mathcal{R}_t$ encompasses all redesign augmentations with their validation status and rewards, enabling performance stratified knowledge extraction from both successful modifications and contraindicated patterns.

3.3.2 Within-trial Learning: Iterative Optimization

To refine LLM behavior by prompt optimization (Ramnath et al., 2025), we extract short-term memory from reward $\mathcal{R}_t$ and integrate with Agent.

Knowledge Extraction. We partition $\mathcal{R}_t$ into two types of knowledge: action-level patterns $\mathcal{K}_t^s$ aggregate modification type performance across aspects; and example-level demonstrations $\mathcal{K}_t^t$ comprise performance stratified modifications.

Agent Integration. The Analyzer's aspect prioritization logic integrates $\mathcal{K}_t^s$ via coverage-based confidence scoring: penalizing repeated patterns, rewarding unexplored spaces, and weighting by historical success rates. The Generator Agent samples performance stratified exemplars $\mathcal{K}_t^t$ for few-shot prompting (Brown et al., 2020). While the Evaluator maintains a redesign pool of high performing and positively rewarded modifications for combinatorial search reuse, enabling *test-time search space scaling* (Snell et al., 2024).

3.3.3 Cross-trial Learning: Global Memory

After each trial converges, global memory is extracted via LLM synthesis, where each trial benefits from and contributes to the evolving knowledge pool. Building on meta-learning frameworks (Parisi et al., 2019), we implement cross-trial knowledge transfer, maintaining generalizable patterns via two representations:

Qualitative Strategic Guidance. Aspect-level recommendations extracted from high performing redesign patterns provide warm start initialization for the Analyzer Agent at iteration $t = 1$, where iteration specific feedback is absent.

Quantitative Statistical Signatures. Recorded mean reward, variance, and modification success rates enable the Generator Agent to calibrate exploration intensity, scaling generation count inversely with historical success rates and proportionally to pattern variance.

4 Experiment

We evaluate `ClinicalReTrial` across two dimensions: (1) simulation environment performance, validating that simulation environment achieve sufficient accuracy to serve as reliable feedback oracles, and (2) `ClinicalReTrial` Agent optimization quality, demonstrating that our multi-agent system successfully redesigns failed trials through iterative learning.

4.1 Experimental Setup

Our system is built on GPT-4o-mini and evaluated on failed clinical trials from the TrialBench dataset (Chen et al., 2025). Using 20769 annotated Phase I-IV trials, we encode multi-modal features into 6,173-dimensional embeddings (details in Appendix A.2) and train LightGBM (Ke et al., 2017) classifiers to predict trial outcome $\hat{y} \in [0, 1]$. We follow TrialBench's train-test split, further splitting the training set 8:2 for training-validation. Due to computational constraints, we evaluate the agent on random selected sample of 60 failed trials from the test set (20 enrollment, 20 safety, 20 efficacy failures) representing diverse trial phases (Data detail in Appendix A.1). The agent operates with a 5-iteration budget. We use exhaustive search when the combinatorial space has $< 1{,}000$ candidates, otherwise beam search with width $k = 8$. We measure effectiveness through predicted probability improvement, threshold achievement rate, and convergence efficiency.

4.2 Simulation Environment Performance

Our Simulation Environment's model is compared against *baseline* approaches (TrialBench (Chen et al., 2025) and HINT (Fu et al., 2022), prior SOTA systems operating on original TrialBench features); and Logistic Regression also trained on the same encodes (Appendix A.2).

Failure-specific Prediction. Implemented in `ClinicalReTrial` Agent, the simulation environ-

ment must correctly predict specific failure outcomes. We train three independent models on our encoded features, each targeting one failure detection task against success. Table 1, 2, and 3 report comprehensive metrics across our models and baseline approaches trained for the task: Poor Enrollment, Safety/Adverse Effect and Lack of Efficacy prediction. Our model achieves PR-AUC > 0.75 across all failure modes, meeting the threshold for reliable discriminative feedback. All models achieve Failure Detection Rates of 70-74% with task specific prediction thresholds ($p \geq 0.6$ for enrollment, $p \geq 0.9$ for safety, $p \geq 0.85$ for efficacy). This ensures that predicted probability shifts $\Delta p > 0.03$ indicate improved trial designs (Appendix A.1).

Table 1: Performance on **Poor Enrollment** prediction.

Model	ROC-AUC	PR-AUC	Fail Det.
TrialBench	0.613 ± 0.007	0.626 ± 0.011	0.525 ± 0.013
HINT	0.534 ± 0.010	0.613 ± 0.012	0.580 ± 0.018
Logistic Reg.	0.622 ± 0.010	0.696 ± 0.012	0.669 ± 0.012
`ClinicalReTrial`	$\mathbf{0.676 \pm 0.009}$	$\mathbf{0.754 \pm 0.010}$	$\mathbf{0.740 \pm 0.012}$

Table 2: Performance on **Drug Adverse Effect** prediction.

Model	ROC-AUC	PR-AUC	Fail Det.
TrialBench	0.587 ± 0.017	0.892 ± 0.006	0.427 ± 0.035
HINT	0.513 ± 0.014	0.882 ± 0.009	0.459 ± 0.031
Logistic Reg.	0.612 ± 0.018	0.909 ± 0.008	0.422 ± 0.029
`ClinicalReTrial`	$\mathbf{0.656 \pm 0.018}$	$\mathbf{0.925 \pm 0.007}$	$\mathbf{0.695 \pm 0.028}$

Table 3: Performance on **Drug Efficacy** prediction.

Model	ROC-AUC	PR-AUC	Fail Det.
TrialBench	0.692 ± 0.012	0.862 ± 0.006	0.565 ± 0.020
HINT	0.559 ± 0.013	0.841 ± 0.008	0.525 ± 0.021
Logistic Reg.	0.665 ± 0.015	0.886 ± 0.009	0.549 ± 0.025
`ClinicalReTrial`	$\mathbf{0.746 \pm 0.013}$	$\mathbf{0.914 \pm 0.007}$	$\mathbf{0.725 \pm 0.021}$

TrialBench Benchmark. We further validate the same model architecture against existing benchmarks on the TrialBench 4-class classification task (predicting Success, Enrollment Failure, Safety Failure, or Efficacy Failure). Table 4 shows that the Simulation Environment's base model in our simulation environment performed best over all baselines, with a higher ROC-AUC of 0.06 to 0.19.

Feature Importance Analysis. We studied feature importance analysis with SHAP (Lundberg and Lee, 2017). Consistent with our hypothesis, eligibility, drug-disease interaction features and endpoint alignment features are most important for outcomes prediction (Appendix A.3).

4.3 Case Studies

We analyze clinical trial pairs, where investigators successfully redesigned and re-executed failed protocols spanning enrollment, safety, and efficacy failure modes against `ClinicalReTrial`'s redesign to validate its applicability, as well as to provides critical insight into clinical applicability. Such trial pair cases are chosen based on the mechanistic interpretability for systematic alignment measures.

We present a poor enrollment redesign case (safety and efficacy cases in Appendix E): NCT01298752, a Phase-III trial of *Mapracorat* (anti-inflammatory ophthalmic suspension) for post-cataract surgery inflammation that failed due to slow enrollment. Sponsored by *Bausch & Lomb*, the trial was subsequently redesigned and successfully executed as NCT01591161. Figure 2 illustrates `ClinicalReTrial`'s iterative refinement process across three optimization cycles. The Enrollment barrier (cataract surgery waiting requirement) is efficiently identified with a positive reward provided by simulation environment. The agent also progressively explores the modification space: baseline AC cell requirements are successfully added as an enrichment criterion; while agent also explores the safety enhancement, but end up failing to fully align with real-world redesign.

4.4 Agent's Protocol Optimization

Having confirmed the simulation environment's reliability, we evaluate `ClinicalReTrial` Agent's ability to redesign failed clinical trials.

4.4.1 Convergence Analysis

Table 5 reports comprehensive convergence statistics across all trials. The Agent had 83.3% of protocol designs improved (50/60 successfully processed trials showed positive Δp), with 4 trials (6.7%) encountering agent failures where the system identified zero opportunities of potential redesign, occurring in the efficacy failure mode.

The system demonstrated efficient convergence patterns, with 15% (9/60) of trials exhibiting natural termination before iteration 5 due to exhausted modification space. Most trials used all 5 iterations, suggesting adaptive stopping could improve efficiency.

Table 4: Performance of multi-class clinical trial outcome prediction across trial phases.

Model	Phase 1		Phase 2		Phase 3		Phase 4	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
TrialBench	0.475±0.027	0.255±0.006	0.569±0.010	0.295±0.008	0.550±0.012	0.279±0.008	0.477±0.022	0.256±0.007
HINT	0.540±0.022	0.272±0.009	0.535±0.009	0.267±0.005	0.474±0.019	0.251±0.006	0.548±0.021	0.273±0.014
Logistic Reg.	0.606±0.019	0.326±0.015	0.583±0.011	0.306±0.008	0.621±0.016	0.350±0.017	**0.550±0.022**	0.280±0.010
`ClinicalReTrial`	**0.633±0.016**	**0.344±0.016**	**0.662±0.011**	**0.382±0.011**	**0.669±0.017**	**0.412±0.019**	0.543±0.025	**0.282±0.012**

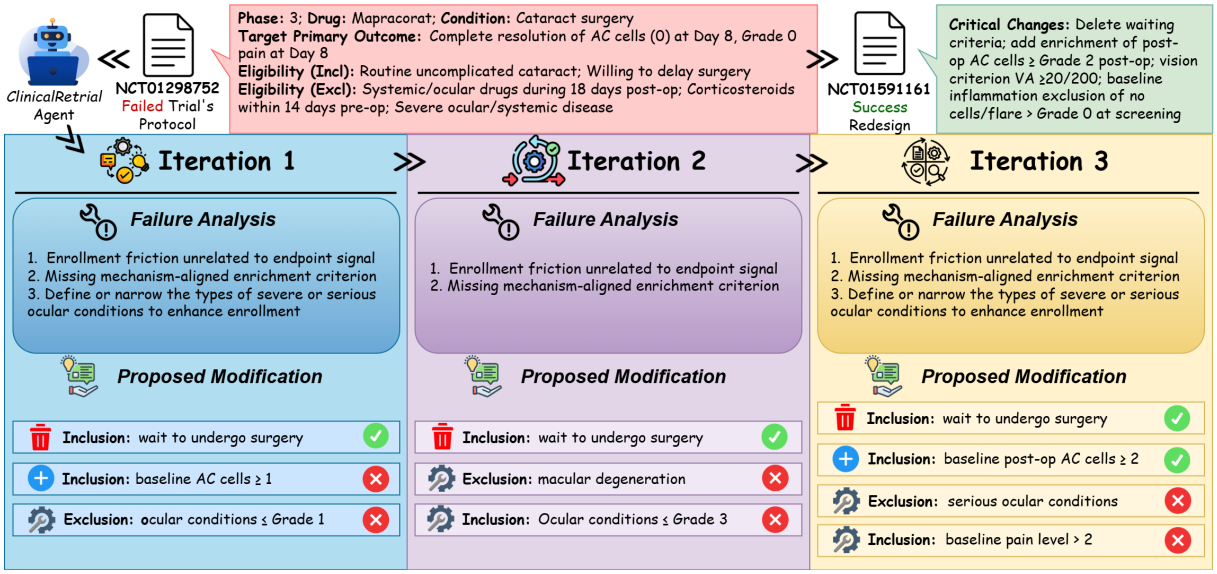

Figure 2: `ClinicalReTrial` Agent's flowchart on Poor Enrollment failed trial case study **(NCT01298752, 2011-02-16)**, together with the real-world redesign **(NCT01591161, 2012-05-02)**. The Agent's iterative refinement of failure analysis and according modifications are demonstrated.

Table 5: Convergence analysis across 60 test trials.

Mode	Trials	Failures	Improved Δp(%)	Threshold
Enrollment	20	0	20/20 (100%)	8/20 (40%)
Safety	20	0	18/20 (90%)	4/20 (20%)
Efficacy	20	4	12/20 (60%)	10/20 (50%)
Overall	**60**	**4 (6.7%)**	**50/60 (83.3%)**	**22/60 (36.7%)**

Table 6: Probability shift (Δp) analysis by failure mode across 56 trials. IQR is 25th-75th percentile range.

Mode	Trials	p_0 / p_{final}	Δp	IQR
Enrollment	20	0.506 / 0.563	+0.058	[+0.034, +0.068]
Safety	20	0.791 / 0.831	+0.070	[+0.032, +0.092]
Efficacy	16	0.813 / 0.859	+0.040	[+0.013, +0.039]
Overall	**56**	**0.695 / 0.730**	**+0.057**	**[+0.029, +0.073]**

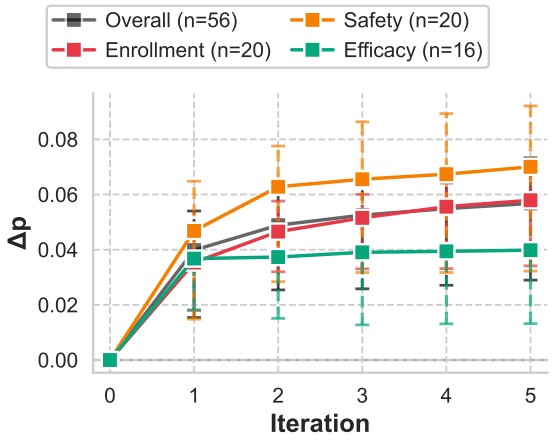

Figure 3: Trial redesign iterative improvement trajectories stratified by failure mode.

4.4.2 Iterative Improvement Trajectories

We examine the learning trajectory of successfully processed trials (56/60). The system achieved mean improvement of $\Delta p = +0.057$. Table 6 stratifies results by failure mode, while Figure 3 illustrates learning dynamics by iteration.

Performance heterogeneity across failure modes reflects the differential amenability of clinical trial design elements to protocol-level intervention. Safety failures exhibit the largest improvements (mean $\Delta p = +0.070$, IQR [+0.032, +0.092]), as adverse events often stem from identifiable contraindication patterns that can be systematically addressed through eligibility refinement and dosage

Table 7: Computational cost by failure mode.

Mode	N	Cost ($)	Δp	Cost/Δp
Enrollment	20	0.171	+0.055	4.76
Safety	20	0.100	+0.054	3.00
Efficacy	16	0.088	+0.038	3.05
Overall	**56**	**0.122**	**+0.053**	**3.68**

adjustment. Enrollment failures show substantial gains (mean $\Delta p = +0.058$), consistent with the observation that recruitment barriers frequently arise from overly restrictive or poorly specified inclusion criteria rather than fundamental feasibility constraints. Efficacy failures demonstrate the smallest yet statistically meaningful improvements (mean $\Delta p = +0.040$), as therapeutic effectiveness depends heavily on drug-disease compatibility, where sometimes protocol modifications alone cannot fix the essential drug failure.

As shown in Figure 3, the learning trajectory reveals major initial gains followed by decreasing returns. This diminishing returns pattern validates the knowledge distillation mechanism: high-quality modifications are identified early through rapid retrieval boosted analysis, while later iterations exploit narrower optimization opportunities by refining secondary parameters or addressing edge case contraindications.

4.4.3 Computational Cost and Efficiency

The system demonstrates practical feasibility with a mean cost of $0.12 per trial across 56 trials, which is a negligible fraction of typical $2.6 billion drug development costs (Lo et al., 2019). Table 7 shows consistent cost-effectiveness across failure modes (Cost/Δp: 3.0-4.8). Linear scaling enables industrial deployment: 1,000 trials cost $120, establishing `ClinicalReTrial` as practical for systematic optimization at scale.

4.5 Ablation Study on Self-improvement

To validate architectural components, we conducted paired ablation across 10 enrollment failure trials, systematically removing: (1) memory-guided iterative learning and (2) redesign pool optimization. Each trial was evaluated under full system and both conditions with identical initialization, enabling within-subjects comparison.

Both components contribute significantly and independently (Figure 4, Table 8). Memory removal degraded performance at iteration 1 ($\Delta p = +0.0131$) and iteration 5 ($\Delta p = +0.0190$), indicating immediate warm start benefits that strengthen over iterations. Removing the redesign pool yielded comparable degradation ($\Delta p = +0.0126$). These findings validate our design: memory transforms random exploration into intelligent exploration, while the redesign pool enables exploitation by reusing successful modifications.

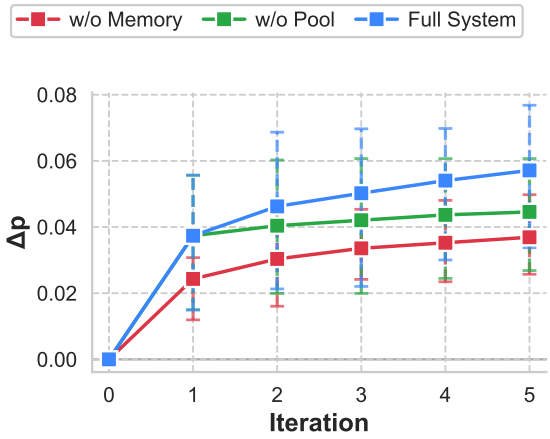

Figure 4: Self-Improving ablation study across 10 trials. Full system (blue) outperforms setups without memory (red) or redesign pool optimization (green). Memory provides early benefits, and pool effects compound over iterations.

Table 8: Self-Improving ablation study. Stats computed using paired t-tests with $n = 10$ trials. *p-value < 0.05, **p-value < 0.01.

Removed	Iteration	Ablation	Full System	Δp	p-value	Cohen's d_z
Memory	1	0.024 ± 0.016	0.037 ± 0.020	+0.013	0.042*	0.81
Memory	5	0.038 ± 0.016	0.057 ± 0.023	+0.019	0.007**	1.10
Pool	5	0.045 ± 0.022	0.057 ± 0.023	+0.013	0.005**	1.18

5 Conclusion

In this work, we present `ClinicalReTrial`, a novel self-evolving agent that moves beyond passive clinical trial outcome prediction to enable proactive optimization of clinical trial protocols. Evaluated on the standardized TrialBench benchmark, `ClinicalReTrial` achieves strong predictive performance while demonstrating the ability to discover significant protocol improvements with clinical best practices, highlighting the potential of agentic AI systems to serve as practical clinical decision support systems for more efficient autonomous trial design. Case study on tial pairs validate that `ClinicalReTrial` exhibit strategic alignment with independently derived real-world trial modifications. Future work will explore tighter integration with real-world constraints.

6 Limitations

Our framework has several limitations that suggest directions for future research. First, the simulation environment's predictive accuracy creates potential for false improvement signals and missed opportunities, though validation filtering and iterative refinement provide partial mitigation; this constraint could be addressed by integrating future state-of-the-art prediction models as drop-in replacements. Second, the system lacks adaptive convergence detection; 83.9% of trials exhausted the full 5-iteration budget rather than stopping when modification potential plateaus, suggesting the need for learned stopping criteria based on diminishing returns patterns. Third, retrospective case study analysis reveals tactical domain knowledge gaps: while the system excels at strategic-level reasoning, it may struggle with operational specifics such as selecting appropriate biomarkers, anticipating implementation constraints, and distinguishing when radical simplification outperforms incremental fortification, often over-relying on complexity multiplication where parsimony proves more effective.

Future work should prioritize prospective validation in collaboration with clinical trial sponsors, integration of specialized biomarker knowledge bases to address tactical gaps, development of adaptive stopping mechanisms to improve computational efficiency, and expansion to larger-scale evaluation encompassing broader disease areas and trial designs.

References

Elodie Baumfeld Andre, Robert Reynolds, Patrick Caubel, Laurent Azoulay, and Nancy A Dreyer. 2020. Trial designs using real-world data: the changing landscape of the regulatory approval process. *Pharmacoepidemiology and drug safety*, 29(10):1201–1212.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jintai Chen, Yaojun Hu, Mingchen Cai, Yingzhou Lu, Yue Wang, Xu Cao, Miao Lin, Hongxia Xu, Jian Wu, Cao Xiao, Jimeng Sun, Yuqiang Li, Lucas Glass, Kexin Huang, Marinka Zitnik, and Tianfan Fu. 2025. TrialBench: Multi-modal AI-ready datasets for clinical trial prediction. *Scientific Data*, 12(1):1564.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024a. CoD, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*.

Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. 2024b. A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint arXiv:2412.17481*.

Tianyi Chen, Nan Hao, Yingzhou Lu, and Capucine Van Rechem. 2024c. Uncertainty quantification on clinical trial outcome prediction. *Health Data Science*.

Yizhou Chi, Yizhang Lin, Sirui Hong, Duyi Pan, Yaying Fei, Guanghao Mei, Bangbang Liu, Tianqi Pang, Jacky Kwok, Ceyao Zhang, Bang Liu, and Chenglin Wu. 2024. SELA: Tree-search enhanced LLM agents for automated machine learning. *arXiv preprint arXiv:2410.17238*.

Simon Dagenais, Leo Russo, Ann Madsen, Jen Webster, and Lauren Becnel. 2022. Use of real-world evidence to drive drug development strategy and inform clinical trial design. *Clinical Pharmacology & Therapeutics*, 111(1):77–89.

Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. 2016. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47:20–33.

David B Fogel. 2018. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemporary clinical trials communications*, 11:156–164.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics.

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.

Kaitlyn M. Gayvert, Neel S. Madhukar, and Olivier Elemento. 2016. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chemical Biology*, 23(10):1294–1301.

Kenneth A. Getz and Rafael A. Campo. 2017. Trends in clinical trial design complexity. *Nature Reviews Drug Discovery*, 16(5):307.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L. Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E. Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30 of *NeurIPS*, pages 3146–3154.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33 of *NeurIPS*, pages 9459–9474.

Fengze Liu, Haoyu Wang, Joonhyuk Cho, Dan Roth, and Andrew W. Lo. 2025. AutoCT: Automating interpretable clinical trial prediction with LLM agents. *arXiv preprint arXiv:2506.04293*.

Andrew W. Lo, Kien Wei Siah, and Chi Heem Wong. 2019. Machine learning with statistical imputation for predicting drug approvals. *Harvard Data Science Review*, 1(1).

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30 of *NeurIPS*, pages 4765–4774.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71.

Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, and 1 others. 2025. A systematic survey of automatic prompt optimization techniques. *arXiv preprint arXiv:2502.16923*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Yue Wang, Yingzhou Lu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Honghao Gao, and Jian Wu. 2024. TWIN-GPT: Digital twins for clinical trials via large language model. *arXiv preprint arXiv:2404.01273*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35 of *NeurIPS*, pages 24824–24837.

David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanveer Sajed, Daniel Johnson, Camille Li, Zinat Sayeeda, Nasrin Assempour, Imad Iynkkaran, Yifeng Liu, Adam Maciejewski, Natali Gale, Anson Wilson, Lee Chin, Robert Cummings, Daniel Le, and 3 others. 2018. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082.

Satoshi Yamaguchi, Mika Kaneko, and Mamoru Narukawa. 2021. Approval success rates of drug candidates based on target, action, modality, application, and their combinations. *Clinical and Translational Science*, 14(3):1113–1122.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, ICLR.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024. ClinicalAgent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB, pages 1–10.

Bin Zhang, Lu Zhang, Qiuying Chen, Zhe Jin, Shuyi Liu, and Shuixing Zhang. 2023. Harnessing artificial intelligence to improve clinical trial design. *Communications Medicine*, 3(1):191.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, ICLR.

A Simulation Environment Details

A.1 Dataset Statistics

Dataset used for training the prediction models comprises 20,769 clinical trials from TrialBench's failure reason dataset. Table 9 shows the label distribution across four categories.

Table 9: Distribution of failure reason labels in the dataset.

Failure Reason	Count	Percentage
Success	9,939	47.8%
Poor Enrollment	7,229	34.8%
Inefficacy	2,217	10.7%
Adverse Effect	1,384	6.7%
Total	**20,769**	**100.0%**

The class imbalance reflects real-world trial outcomes: enrollment challenges are the most common failure mode (34.8%), followed by efficacy gaps (10.7%), while safety failures are relatively rare (6.7%) due to rigorous preclinical screening. Success cases (47.8%) include trials that completed without major protocol violations or early termination.

Due to computational cost constraints, we randomly select a stratified sample of 60 trials from the test set (20 enrollment, 20 safety, 20 efficacy), ensuring representation across failure modes and trial phases. Table 10 presents the phase composition.

Table 10: Test distribution by trial phase.

Phase	Count	%
Phase 1	8	13.3
Phase 2	27	45.0
Phase 3	15	25.0
Phase 4	10	16.7
Total	**60**	**100**

A.2 Encode Details

This appendix provides comprehensive implementation details for the Simulation Environment described in §3.2.4, including encoder pretraining procedures, model training hyperparameters, and detailed validation results.

Text Features. Textual contents are encoded using BioBERT (Lee et al., 2020), a domain-adapted language model pre-trained on PubMed abstracts and PMC full-text articles. Critically, we diverge from prior work by decomposing eligibility criteria at the sentence level rather than treating them as monolithic text blocks. For each text field $\mathcal{T}$, we decompose it into sentences $\mathcal{T} = \{s_1, s_2, \ldots, s_n\}$. Each sentence is encoded via BioBERT and the final text embedding is obtained via max pooling:

$$\mathbf{e}_{s_i} = \text{BioBERT}(s_i), \quad \mathbf{h}_{\mathcal{T}} = \max_{i=1}^{n} \mathbf{e}_{s_i} \tag{2}$$

This sentence-level representation preserves granularity essential for aspect-specific modification: `ClinicalReTrial` Agent can target individual criteria rather than generic protocol summaries.

Graph Features. We incorporate pre-trained molecular and disease encodings to capture pharmacological properties and disease characteristics.

Drug Molecular Graphs. Each drug molecule m is represented as a graph $\mathcal{G}_m = (\mathcal{V}, \mathcal{E})$ where nodes $v \in \mathcal{V}$ are atoms and edges $(u, v) \in \mathcal{E}$ are bonds. We employ Message Passing Neural Networks (MPNNs) to aggregate neighborhood information over L iterations:

$$\mathbf{m}_{uv}^{(l)} = \text{ReLU}\left(W_i \cdot [\mathbf{f}_u \oplus \mathbf{f}_{uv}] + W_h \cdot \sum_{w \in \mathcal{N}(u) \setminus v} \mathbf{m}_{wu}^{(l-1)} \right), \tag{3}$$

where $\mathbf{m}_{uv}^{(l)} \in \mathbb{R}^{d_{\text{mpnn}}}$ is the message from atom u to atom v at layer l, $\mathcal{N}(u)$ denotes neighbors of u, $\oplus$ denotes concatenation, and W_i, W_h are learnable transformation matrices. After L message passing iterations, node embeddings are computed as:

$$\mathbf{h}_u = \text{ReLU}\left(W_o \cdot \left[\mathbf{f}_u \oplus \sum_{v \in \mathcal{N}(u)} \mathbf{m}_{vu}^{(L)} \right] \right). \tag{4}$$

The graph-level drug embedding is obtained via global average pooling:

$$\mathbf{h}_{\text{drug}} = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \mathbf{h}_u \in \mathbb{R}^{d_{\text{mpnn}}} \tag{5}$$

For trials with multiple drugs, we average their embeddings. The MPNN encoder is pretrained on pharmacokinetic (ADMET) tasks, then fine-tuned on trial outcome labels (details in Appendix A).

Disease Hierarchical Encoding. Each disease is represented by an ICD-10 code d_i following a hierarchical taxonomy with ancestors $\mathcal{A}(d_i) = \{a_1, a_2, \ldots, a_p\}$. We use Graph-based Attention Model (GRAM) to encode hierarchical disease information. Each code c has a learnable base embedding $\mathbf{e}_c \in \mathbb{R}^{d_{\text{gram}}}$. The hierarchical embedding for disease d_i is computed as an attention-weighted sum over itself and its ancestors:

$$\mathbf{h}_{d_i} = \sum_{a_j \in \mathcal{A}(d_i) \cup \{d_i\}} \alpha_{ji} \cdot \mathbf{e}_{a_j} \tag{6}$$

where the attention weight α_{ji} measures the relevance of ancestor a_j to the current disease d_i:

$$\alpha_{ji} = \frac{\exp(\phi([\mathbf{e}_{a_j} \oplus \mathbf{e}_{d_i}]))}{\sum_{a_k \in \mathcal{A}(d_i) \cup \{d_i\}} \exp(\phi([\mathbf{e}_{a_k} \oplus \mathbf{e}_{d_i}]))} \tag{7}$$

where $\phi(\cdot) : \mathbb{R}^{2d_{\text{gram}}} \to \mathbb{R}$ is a learnable single-layer network. For trials targeting multiple diseases, we average their embeddings. The GRAM encoder is initialized with the ICD-10 hierarchical ontology, then fine-tuned on historical trial success rates (details in Appendix A).

Tabular Features. We encode structured trial metadata through a modular pipeline that processes categorical attributes, demographic constraints, administrative properties, and enrollment characteristics. The pipeline extracts 29 numerical features.

Problem Formulation and Dataset. We formulate clinical trial outcome prediction as a binary classification problem over three distinct failure modes: poor enrollment, safety/drug adverse effect, and drug inefficacy. We train models separately for each failure mode, enabling `ClinicalReTrial` Agent to target specific causes during protocol optimization. Our experiments utilize the TrialBench dataset (Chen et al., 2025), which contains over 12,000 annotated clinical trials spanning Phase I through Phase IV, with each trial labeled according to outcome. The dataset provides multi-modal features including drug molecular structures, disease ICD-10 codes, eligibility criteria text, trial metadata, and intervention details. Following standard practice to avoid temporal leakage, we partition data chronologically by trial completion year. According to Table 11, features are encoded into total dim=6,173.

Table 11: Feature specification summary. Novel contributions include sentence-level eligibility parsing and fine-tuned molecular-disease encoders.

Category	Component	Dim	Method	Novel
Text	Study Design	768	BioBERT	
	Dosage	768	BioBERT	
	Intervention	768	BioBERT + pooling	✓
	Condition	768	BioBERT + pooling	✓
	Eligibility Inclusion	768	BioBERT + pooling	✓
	Eligibility Exclusion	768	BioBERT + pooling	✓
Graph	Drug (ADMET)	768	MPNN (fine-tuned)	✓
	Disease (ICD)	768	GRAM (fine-tuned)	✓
Tabular	Categorical Features	18	One-Hot	✓
	Age constraints	2	Unit normalization	✓
	Multi-hot indicators	9	Binary encoding	✓
Total		**6,173**		

Feature Concatenation and Prediction. All feature modalities are concatenated into a single input vector:

$$\mathbf{x}_{\text{trial}} = [\mathbf{h}_{\text{design}}; \mathbf{h}_{\text{dose}}; \mathbf{h}_{\text{interv}}; \mathbf{h}_{\text{cond}}; \mathbf{h}_{\text{incl}}; \mathbf{h}_{\text{excl}}; \mathbf{h}_{\text{drug}}; \mathbf{h}_{\text{disease}}; \mathbf{f}_{\text{tabular}}] \in \mathbb{R}^{6173} \tag{8}$$

where semicolons denote concatenation. For each failure mode $\tau \in \{\text{enrollment, safety, efficacy}\}$, we train a separate LightGBM classifier $\mathcal{M}_\tau$ that predicts trial success probability. The predicted probability $\hat{y} = \mathcal{M}_\tau(\mathbf{x}_{\text{trial}}) \in [0, 1]$ serves as the reward signal for evaluating protocol modifications in the agent system.

Model Training and Validation. We employ LightGBM (Ke et al., 2017) for its computational efficiency with high-dimensional sparse features. Three independent models are trained for enrollment, safety, and efficacy failure prediction using cross-validation with early stopping. The trained GBDT models achieve strong predictive performance across all failure modes (PR-AUC > 0.75) with well-calibrated probability estimates, validating the simulation environment as a reliable proxy for real trial outcomes.

A.3 Ablation Study

Word-Level Attention Analysis. Figure 5 demonstrates the word-level attention weights captured by BioBERT embeddings in the TrialDura model, visualized through Shapley values. The heatmap reveals that clinical keywords such as "woman," "contraception," receive the highest attention weights (0.0208–0.0274), while functional words like prepositions and conjunctions are assigned lower weights. This attention distribution indicates that the model effectively focuses on medically relevant terminology when processing eligibility criteria, suggesting that domain specific language models can automatically identify critical phrases without explicit feature engineering.

Figure 5: Visualization of text segments in the BioBERT encoder's output, illustrating Shapley values derived from Clinical Trials. Shapley values correspond to attention weights, with darker colors indicating higher weights.

Sentence-Level Eligibility Weights. Table 6 illustrates a example of sentence-level importance scores within the inclusion criteria for trial NCT01102504, normalized across all eligibility statements, with weighted importance calculated on predict probability shift if masking out each eligibility protocols. The model assigns highest weights (0.20–0.25) to sentences describing acute cerebrovascular events such as "Transient ischemic attack (TIA)" and "Stroke (ipsilaterally to the stenotic artery)," while demographic criteria like age receive minimal attention (0.07). Notably, the quantitative stenosis threshold "$>30\%$ stenosis on initial B-mode ultrasonography imaging" receives substantial weight (0.18), indicating that the model prioritizes disease severity markers and clinical events over basic demographic qualifications when predicting trial outcomes.

Table 12: Inclusion Criteria with Sentence Importance (Color-coded)

NCT01102504 Eligibility Criteria Protocols	**Weight**
Inclusion Criteria:	0.03
- Age 40–90 years old,	0.07
- Clinically documented carotid symptomatic atherosclerotic disease (symptomatic disease will be considered if one of the following has occurred within 2 months prior to symptoms:)	0.12
1. Amaurosis fugax	0.10
2. Transient ischemic attack (TIA)	0.20
3. Stroke (ipsilaterally to the stenotic artery)	0.25
- $>30\%$ stenosis on initial B-mode ultrasonography imaging,	0.18
- Written, informed consent.	0.05

Encodes Contributions Revealed Through Ablation Analysis. Figure 6 presents the relative importance of different encoders across three prediction tasks through systematic masking experiments. By individually masking each encoder and measuring the resulting PR-AUC drop, we quantify each component's contribution to enrollment, safety, and efficacy outcome predictions. The analysis reveals task-specific dependency patterns: certain encoders prove critical for particular outcomes, with their removal causing substantial performance degradation, while showing minimal impact on other tasks. This heterogeneous importance distribution demonstrates that different aspects of trial design and patient characteristics drive distinct clinical endpoints. The varying magnitudes of PR-AUC drops across tasks validate the multi-task learning framework's ability to capture task-specific representations while identifying which shared features are most crucial for each prediction objective.

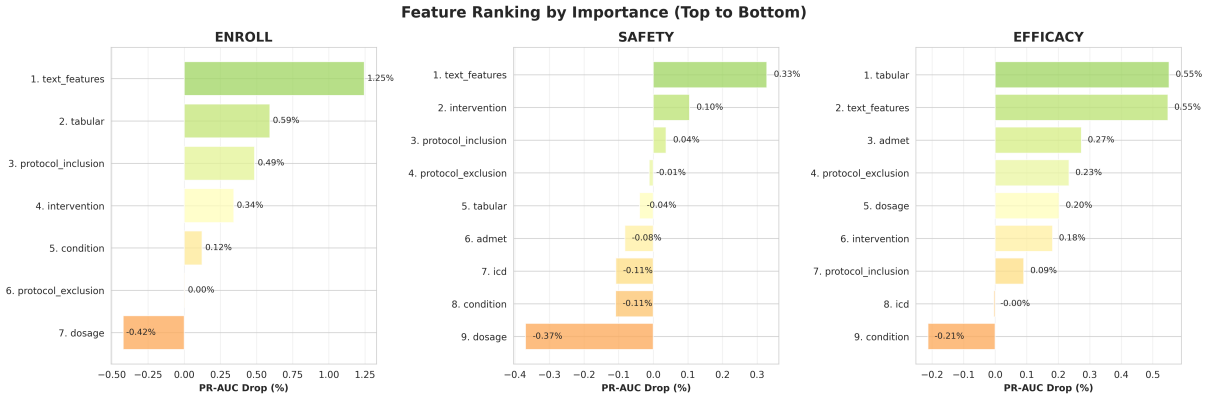

Figure 6: Feature importance of each encoder on 3 classification tasks (enrollment, safety, and efficacy), measured by PR-AUC drop when the encoder is masked out during prediction.

B Analyzer Agent Details

The Analyzor Agent implements a domain-aware ReAct reasoning pipeline adapted by failure mode (enrollment, safety, efficacy). Novel components include adverse event profiling, statistical power assessment, and design-level pivots. Table 13 summarizes failure-mode-specific adaptations.

Table 13: Analyzer Agent variants by failure mode.

Component	Enrollment	Safety	Efficacy
Profiling	None	Adverse Event Profiling (severity, organ systems, root cause)	Efficacy Gap Profiling (observed vs expected)
Classification	4 categories	5 categories (adds `safety_inadequate`)	4 categories (weights enrichment higher)
Assessments	None	Dosage + AE profile	Dosage + Outcome + Power analysis
Prioritization	Confidence-based	Safety-first (toxicity reduction priority)	Simplicity-first tiered (PRIMARY/SECONDARY/TERTIARY)

Protocol Classification

Role: Clinical researcher classifying eligibility criteria
Task: Classify criteria into 4 categories with confidence scores [0-1]
Context:

```
Phase: Phase 2
Mechanism: X inhibits Y pathway
Endpoint: Measuring Z at 12 weeks
```

Criteria to Classify:

```
<criterion aspect_name="eligibility/inclusion_criteria" index="1">
Must wait for fellow eye surgery until study completion
</criterion>
<criterion aspect_name="eligibility/exclusion_criteria" index="2">
Any prior participation in drug trials within 12 months
</criterion>
```

Categories:

1. PARTICIPATION_BARRIER: Timing/waiting requirements, administrative hurdles
   -
2. SAFETY_EXCLUSION: Medical risks (allergies, drug interactions, severe conditions)
3. SELECTION_CRITERION: Defines WHO is eligible (disease type, procedure type, demographics)
4. ENRICHMENT_CRITERION: Selects likely responders (biomarkers, mechanism-aligned traits)

For each criterion, assign scores [0-1] to ALL categories, pick PRIMARY (highest), give 1-sentence reason.
Output Format:

```
<classification aspect_name="eligibility/inclusion_criteria" index="1">
<participation_barrier_score>0.92</participation_barrier_score>
<safety_exclusion_score>0.05</safety_exclusion_score>
<selection_criterion_score>0.20</selection_criterion_score>
<enrichment_criterion_score>0.10</enrichment_criterion_score>
<primary_category>PARTICIPATION_BARRIER</primary_category>
<reasoning>Waiting requirement for fellow eye surgery is a strong
participation barrier with no medical justification.</reasoning>
</classification>
```

Mechanism Alignment Check

Role: Clinical researcher evaluating mechanism alignment
Task: Check if criteria select mechanism-appropriate patients and detect missing enrichment
Questions:

1. Do we select patients who HAVE the target condition this mechanism treats?
2. Do we select patients with baseline values allowing measurement of endpoint Y?
3. Are safety exclusions too broad, blocking potential responders?

If missing enrichment (no criteria selecting treatment-responsive patients):

- Propose ONE objective criterion with: measurement method, threshold, timing
- Must be measurable (grades/scores/labs), not subjective ("anticipated"/"likely")

Output Format:

```
<mechanism_analysis>
Current criteria define cataract surgery candidates but lack enrichment
for inflammation severity. Waiting requirement blocks eligible patients
without medical benefit.
</mechanism_analysis>

<missing_enrichment_criterion>
Add inclusion: Baseline anterior chamber cell grade $\geq$2+ (SUN criteria)
measured within 7 days of enrollment. Selects patients with measurable
inflammation for mechanism-aligned response assessment.
</missing_enrichment_criterion>
```

Adverse Event Profiling

Role: Clinical researcher analyzing safety failures
Task: Parse and categorize adverse events for safety redesign
Input:

```
Adverse events: Hepatotoxicity (Grade 3, 25%), elevated AST/ALT (Grade 2, 40%)
Intervention: Drug X (oral, 100mg daily for 28 days)
Mechanism: Inhibits enzyme Y in Z pathway
```

Instructions:

1. SEVERITY CLASSIFICATION: Extract Grade 3-5 events (dose-limiting), Grade 2 (tolerability)
2. ORGAN SYSTEM MAPPING: Map toxicity to organ (Liver, Kidney, Bone marrow, Heart, GI)
3. MECHANISM CONSISTENCY: Does toxicity match expected mechanism?
4. DOSE-RESPONSE INFERENCE: Dose-dependent? Acute or cumulative?
5. PRIORITY RANKING: CRITICAL (Grade 3+ >10%), HIGH (Grade 2+ >30% OR any Grade 4+)
6. ROOT CAUSE HYPOTHESIS: Excessive dose, inadequate exclusions, off-target effects?

Output Format:

```
<adverse_event_profile>
<primary_toxicity>
  <event>Hepatotoxicity</event>
  <grade>3</grade>
  <incidence>25%</incidence>
  <organ_system>Liver</organ_system>
  <priority>CRITICAL</priority>
```

```
  <dose_dependent>likely</dose_dependent>
</primary_toxicity>

<mechanism_consistency>
UNEXPECTED - mechanism does not predict liver toxicity
</mechanism_consistency>

<root_cause_hypothesis>
Likely excessive dose (100mg exceeds typical range) or missing hepatic
impairment exclusion. Drug metabolism may saturate at high doses.
</root_cause_hypothesis>

<critical_gaps>
<gap>Exclude patients with baseline AST/ALT >2x ULN</gap>
<gap>Exclude patients with Child-Pugh Class B or C cirrhosis</gap>
</critical_gaps>
</adverse_event_profile>
```

Design-Level Pivots

Role: Clinical trial designer proposing trial-level redesign
Task: Propose high-level trial redesign (not just criteria tweaks)
Context:

```
Phase: Phase 2
Mechanism: Inhibits enzyme Y
Failure: Grade 3 hepatotoxicity 25%
Redesign archetype: PK_SAFETY_FOLLOWUP
Primary outcome: Safety assessment at 28 days
Dosage assessment: EXCESSIVE (100mg daily exceeds safe exposure)
```

Design Pivot Rules:

- If archetype is PK_SAFETY_FOLLOWUP or main failure is safety-driven:
  - Prefer PK_SAFETY or DOSE_FINDING trial type
  - Prefer PK-focused primary endpoints
  - Prefer simpler care model with lower background risk
  - Prefer simpler dosing (single-dose or short-duration)
- When systemic toxicity suspected:
  - Consider more local/regional route to reduce systemic exposure
  - Consider smaller, denser design (PK_SINGLE_ARM with intensive sampling)

Output Format:

```
<design_pivots>
<trial_type>PK_SAFETY</trial_type>
<endpoint_family>PK_SAFETY</endpoint_family>
<dose_regimen_direction>SIMPLER</dose_regimen_direction>
<route_change>CONSIDER_ALTERNATIVE_ROUTE</route_change>
<proposed_route>Consider single 25mg dose with intensive PK sampling
over 7 days, or switch to subcutaneous administration to reduce
first-pass hepatic metabolism</proposed_route>
<sample_size_direction>SMALLER</sample_size_direction>
<design_structure>PK_DOSE_FINDING</design_structure>
<proposed_primary_outcome>Area under curve (AUC) and peak liver enzyme
elevation (AST/ALT) at 24h, 48h, 72h post-dose</proposed_primary_outcome>
<summary>Pivot from Phase 2 efficacy trial to Phase 1b/2a PK safety
study. Reduce dose to 25mg single administration with intensive PK and
liver function monitoring. Alternative route (subcutaneous) may bypass
hepatic first-pass effect. Smaller sample (N=20-30) adequate for PK
characterization. Expected to reduce Grade 3+ hepatotoxicity from 25%
to <5%.</summary>
</design_pivots>
```

Trade-off Analysis

Role: Clinical pharmacologist analyzing dosage for EFFICACY failure
Task: Analyze DOSAGE trade-offs (not safety)
Context:

```
Current dosage: 50mg oral daily for 21 days
Dosage assessment: SUBOPTIMAL
  Classification reasoning: Phase 1 MTD was 100mg daily. Current 50mg
  dose is at 50% of MTD with acceptable safety. PK data shows linear
  dose-response up to 80mg.
  Suggested: Escalate to 75mg daily
Mechanism: Inhibits receptor X
Efficacy Gap: ORR 15% vs 30% (gap: 15%)
Power Assessment: LIKELY underpowered, Root Cause: BOTH
```

Instructions:

1. RECOMMENDATION: MODIFY (escalate) or KEEP (defer)
2. IMPACTS: efficacy_signal [++], enrollment [0], safety [-], mechanism [ALIGNED]
3. CONFIDENCE: High (0.80-0.90) if clear PK/PD data
4. REASONING: Include feasibility (Time: X-Ymo; Burden: LOW|MED|HIGH; Cost: Zx)

Output Format:

```
<dosage_tradeoff>
<recommendation>MODIFY</recommendation>
<efficacy_signal>++</efficacy_signal>
<enrollment>0</enrollment>
<safety>-</safety>
<mechanism_alignment>ALIGNED</mechanism_alignment>
<confidence>0.85</confidence>
<reasoning>Escalating to 75mg (75% of MTD) expected to improve ORR by
10-15 percentage points based on linear PK and Phase 1 exposure-response.
Safety risk manageable (Grade 2 toxicity may increase from 20% to 30%).
FEASIBILITY: Time: 1-3mo; Burden: LOW; Cost: 1.2x (simple dose
adjustment, no formulation change).</reasoning>
</dosage_tradeoff>
```

C Generator Agent Details

Table 14: Generator Agent novel features by failure mode. Process differ a little in dosage modification strategy across failure modes.

Feature	Enrollment	Safety	Efficacy
Dosage Strategy	N/A	**Reduce** (↓25-50%, fractionation, pulse)	**Escalate** (↑25-50%, loading, dose-dense)
Outcome Strategy	N/A	Add safety qualifications	Switch to feasible endpoint
Domain Focus	Enrich participation	Tighten safety exclusions	Add biomarker enrichment

C.1 Few-Shot Learning Mechanism

Few-Shot Example Injection (Shared Structure)

Matching Logic:

```
# LIST aspects (eligibility criteria)
prev_rules["seen_indices"][aspect_name][str(aspect_index)]

# STRING aspects (dosage, target_primary_outcome)
prev_rules["seen_indices"][aspect_name]["None"]
```

Injected Section in MODIFY Prompts (Iteration 2+):

```
<few_shot_examples>
```

```
Previous iteration examples for THIS EXACT criterion:

EXCELLENT:
  - [Example that led to excellent validation score]
  - [Another excellent example]

GOOD:
  - [Example that led to good validation score]

MODERATE:
  - [Example with moderate validation score]

BAD:
  - [Example that validation agent rejected]

BANNED:
  - [Example that was explicitly banned (safety violation)]

Generate variations that learn from EXCELLENT/GOOD patterns,
avoid BAD patterns, and NEVER replicate BANNED augmentations.
</few_shot_examples>
```

Effect: LLM learns from previous iteration's successes/failures. Only available iteration 2+ after prev_rules established.

C.2 Modification Prompts by Failure Mode

Eligibility Example

Role: Clinical researcher generating criterion variations
Task: Generate `num_augment` variations with few-shot guidance
Input:

```
Original criterion: "Must wait for fellow eye surgery until completion"
Strategy: "Delete waiting requirement to increase enrollment"
Failure mode: Enrollment
Adaptive num_augment: 3 (medium variance)
```

Few-Shot Examples (if iteration 2+):

```
EXCELLENT: "No waiting period required between surgeries"
GOOD: "Fellow eye surgery allowed concurrent with study"
BAD: "Reduced wait from 6 months to 3 months" (still a barrier)
BANNED: "Must complete fellow eye surgery before enrollment" (contradicts)
```

Universal Requirements:

- Each variation MUST directly implement the Strategy
- Preserve clinical intent, make more operational/measurable/specific
- Objective and quantifiable (use thresholds, time windows, methods)
- Avoid vague language: "anticipated", "expected", "likely", "may", "severe"
- Maintain consistency with safety and mechanism of action
- All variations distinct from each other

Output:

```
<augmentations>
<augmentation>No waiting period required between fellow eye surgeries</augmentation>
<augmentation>Fellow eye surgery allowed at any time during study</augmentation>
<augmentation>Bilateral surgery candidates eligible without delay</augmentation>
</augmentations>
```

Dosage Example

Role: Clinical pharmacologist reducing dosage to minimize toxicity
Task: Generate `num_augment` dosage reductions
Input:

```
Original dosage: 100mg oral daily for 28 days
Adverse events: Hepatotoxicity (Grade 3, 25%), AST/ALT elevation (Grade 2, 40%)
Strategy: Reduce dose to decrease Grade 3+ hepatotoxicity to <10%
Adaptive num_augment: 5 (high variance)
```

Few-Shot Examples (iteration 3):

```
EXCELLENT: "50mg oral daily (50% reduction, expected toxicity <8%)"
GOOD: "50mg BID (fractionated, reduces Cmax and hepatic load)"
MODERATE: "75mg oral daily (25% reduction, may be insufficient)"
BAD: "90mg oral daily (only 10% reduction)"
BANNED: "100mg every other day (same cumulative exposure)"
```

Dosage Reduction Strategies:

1. DOSE REDUCTION: Reduce total daily dose by 25-50%
2. FRACTIONATED DOSING: Split dose to reduce C_{max} (peak $\rightarrow$ peak toxicity)
3. TITRATION SCHEDULE: Start low, escalate if tolerated
4. INTERMITTENT/PULSE DOSING: Reduce cumulative exposure for cumulative toxicities
5. PATIENT-FACTOR ADJUSTED: Reduce dose for vulnerable populations
6. LOADING DOSE ELIMINATION: Remove if causing acute toxicity

Requirements:

- Reduce estimated Grade 3+ toxicity by $\geq$30%
- Maintain dose intensity $\geq$60% of original (preserve efficacy)
- Specify exact mg, frequency (QD/BID/TID), duration
- If conditional, specify threshold/trigger (*e.g.*, "if AST <2$\times$ULN")

Output:

```
<augmentations>
<augmentation>
<dosage_modification>50mg oral daily for 28 days</dosage_modification>
<rationale>50% dose reduction expected to reduce hepatotoxicity
from 25% to <8% based on linear dose-toxicity relationship</rationale>
</augmentation>
<augmentation>
<dosage_modification>40mg BID (total 80mg daily, fractionated)</dosage_modification>
<rationale>Fractionated dosing reduces Cmax by ~40%, lowering peak
hepatic exposure while maintaining 80% dose intensity</rationale>
</augmentation>
<augmentation>
<dosage_modification>50mg on days 1-5, off days 6-7 each week</dosage_modification>
<rationale>Pulse dosing (71% intensity) allows hepatic recovery,
expected to reduce Grade 3+ events to <10%</rationale>
</augmentation>
</augmentations>
```

D Agent Output Template

This section presents the structured output format produced by the agent pipeline. The complete output is stored as JSON and includes trial data, ReAct reasoning traces, and generated protocol modifications.

Agent Pipeline Output Structure (Generic Template)

```
{
  "trial_data": {
    "nct_id": "NCT########",
    "phase": "Phase X",
    "condition": "[Disease/Condition]",
    "intervention/intervention_name": "[Intervention Name]",
    "failure_reason": "[enrollment|safety|efficacy]",
    "adverse_events": "[Adverse event summary or 'Not specified']",
    "eligibility/inclusion_criteria": [
      "[Inclusion criterion 1]",
      "[Inclusion criterion 2]",
      "..."
    ],
    "eligibility/exclusion_criteria": [
      "[Exclusion criterion 1]",
      "[Exclusion criterion 2]",
      "..."
    ],
```

```
    "dosage": "[Dosage regimen]",
    "target_primary_outcome": "[Primary outcome description]"
  },

  "trial_context": {
    "phase": "Phase X",
    "mechanism_of_action": "[Mechanism description]",
    "primary_endpoint_type": "[Endpoint type description]",
    "redesign_archetype": "[PK_SAFETY_FOLLOWUP | DOSE_FINDING_REDESIGN |
                           ENRICHED_EFFICACY_RETRY | OTHER]",
    "index_surgical_model": "[Care/procedural model description]"
  },

  "react_reasoning": {
    "step0_contextualize": {
      "phase": "Phase X",
      "mechanism_of_action": "[Mechanism extracted by LLM]",
      "adverse_event_profile": {
        "primary_toxicity": {
          "event": "[Primary adverse event]",
          "grade": "[0-5]",
          "incidence": "[X%]",
          "priority": "[CRITICAL|HIGH|MEDIUM|LOW]"
        },
        "root_cause_hypothesis": "[Root cause analysis by LLM]"
      },
      "dosage_assessment": {
        "classification": "[EXCESSIVE|BORDERLINE|APPROPRIATE|SUBOPTIMAL]",
        "reasoning": "[Dosage assessment reasoning]"
      }
    },

    "step1_classification": [
      {
        "aspect_name": "eligibility/[inclusion|exclusion]_criteria",
        "aspect_index": N,
        "criterion_text": "[Original criterion text]",
        "participation_barrier_score": 0.X,
        "safety_exclusion_score": 0.X,
        "selection_criterion_score": 0.X,
        "enrichment_criterion_score": 0.X,
        "primary_category": "[PARTICIPATION_BARRIER | SAFETY_EXCLUSION |
                             SELECTION_CRITERION | ENRICHMENT_CRITERION]",
        "reasoning": "[Classification reasoning]"
      },
      {
        "aspect_name": "eligibility/[inclusion|exclusion]_criteria",
        "aspect_index": M,
        "criterion_text": "[Original criterion text]",
        "primary_category": "[Category]",
        "reasoning": "[Classification reasoning]"
      }
    ],

    "step2_mechanism_alignment": "[3-4 sentences on whether existing
                                  criteria + dosage maximize success
                                  probability for this failure mode]",

    "step3_tradeoff_analysis": [
      {
        "aspect_name": "eligibility/[inclusion|exclusion]_criteria",
        "aspect_index": N,
        "enrollment_impact": "[--|-|0|+|++]",
        "efficacy_signal_impact": "[--|-|0|+|++]",
        "safety_risk_impact": "[--|-|0|+|++]",
        "mechanism_alignment": "[ESSENTIAL|ALIGNED|NEUTRAL|MISALIGNED]",
        "net_recommendation": "[KEEP|MODIFY|DELETE|ADD]",
        "confidence": 0.XX,
```

```
      "reasoning": "[Trade-off reasoning with feasibility encoding]"
    },
    {
      "aspect_name": "[dosage|target_primary_outcome|surgical_model|...]",
      "aspect_index": null,
      "enrollment_impact": "[Impact symbol]",
      "safety_risk_impact": "[Impact symbol]",
      "net_recommendation": "[MODIFY|ADD]",
      "confidence": 0.XX,
      "reasoning": "[Trade-off reasoning]"
    }
  ],

  "step4_prioritization": "[6-8 sentences with tiered recommendations
                           (PRIMARY/SECONDARY/TERTIARY), timeline, and
                           confidence level]",

  "step5_synthesis": "[4-6 sentences synthesizing failure analysis with
                      quantification, expected benefits, trade-offs, and
                      overall confidence]"
},

"aspect_li": [
  {
    "aspect_name": "eligibility/[inclusion|exclusion]_criteria",
    "aspect_index": N,
    "original_value": "[Original criterion text]",
    "aspect_type": "list",
    "analysis": {
      "timestamp": "YYYY-MM-DDTHH:MM:SS",
      "failure_analysis": "[Analysis from step3 trade-off reasoning]",
      "impact_level": "[MAJOR|MINOR|NOT_RELATED]",
      "action_type": "[MODIFY|DELETE]",
      "strategy": "[Strategy from Analysis Agent]",
      "confidence": 0.XX
    },
    "augment": {
      "timestamp": "YYYY-MM-DDTHH:MM:SS",
      "augment_val_li": [
        "[Augmentation 1]",
        "[Augmentation 2]",
        "[Augmentation 3]"
      ]
    }
  },
  {
    "aspect_name": "eligibility/[inclusion|exclusion]_criteria",
    "aspect_index": null,
    "original_value": "N/A",
    "aspect_type": "list",
    "analysis": {
      "timestamp": "YYYY-MM-DDTHH:MM:SS",
      "failure_analysis": "[Analysis for ADD action]",
      "impact_level": "MAJOR",
      "action_type": "ADD",
      "strategy": "[Strategy from Analysis Agent]",
      "confidence": 0.XX
    },
    "augment": {
      "timestamp": "YYYY-MM-DDTHH:MM:SS",
      "augment_val_li": [
        "[New criterion 1]",
        "[New criterion 2]",
        "[New criterion 3]"
      ]
    }
  },
  {
```

```
      "aspect_name": "[dosage|target_primary_outcome]",
      "aspect_index": null,
      "original_value": "[Original value for string aspect]",
      "aspect_type": "string",
      "analysis": {
        "timestamp": "YYYY-MM-DDTHH:MM:SS",
        "failure_analysis": "[Analysis for string aspect]",
        "impact_level": "MAJOR",
        "action_type": "MODIFY",
        "strategy": "[Strategy from Analysis Agent]",
        "confidence": 0.XX
      },
      "augment": {
        "timestamp": "YYYY-MM-DDTHH:MM:SS",
        "augment_val_li": [
          "[Modified value 1]",
          "[Modified value 2]",
          "[Modified value 3]"
        ]
      }
    }
  ]
}
```

E Case Study Details

We validate ClinicalReTrial Agent's reasoning against real-world protocol modifications provides critical insight into clinical applicability. We analyze three trial pairs where investigators redesigned and successfully re-executed failed protocols, enabling direct comparison between expert redesign decisions and ClinicalReTrial Agent's proposals. Each case represents a distinct failure mode: NCT01298752 (poor enrollment), NCT01919190 (safety/adverse effects), and NCT02169336 (efficacy inadequacy).

Poor Enrollment. To validate agent redesign quality against real-world outcomes, we analyze NCT01298752, a Phase 3 trial of Mapracorat (anti-inflammatory ophthalmic suspension) for post-cataract surgery inflammation that failed due to poor enrollment. Sponsored by Bausch & Lomb, the trial was subsequently redesigned and successfully executed as NCT01591161. Table 15 compares the real-world redesign with ClinicalReTrial Agent's proposals.

Table 15: Agent-proposed modifications alignment check with real-world protocol redesign for poor enrollment, ClinicalReTrial Agent's proposed modifications, categorizing alignment as: ✓(perfect match), $\sim$ (strategic alignment, tactical differences), or $\times$ (missed or incorrect).

Modification Type	Real-World Redesign	Agent Proposal	Match	Impact Level
Enrollment Barrier	DELETE: "subjects must be willing to wait to undergo cataract surgery..."	DELETE: "subjects must be willing to wait to undergo cataract surgery..."	✓	Major, removed primary barrier
Quality Enrichment	ADDED: AC cells $\geq$ Grade 2 (6-15 cells)	ADD: Require baseline AC cells $\geq$2 within 7 days	✓	Major, critical enrichment criteria
Safety Standardization	Exclude inflammation/pain $>$ Grade 1 at screening. Exclude active external ocular disease, POD1 + VA$\geq$20/200	Include pain$>$2 at screening (negative reward); Exclude serious ocular conditions (negative reward)	$\times$	Major, Maintained safety, reduced over-restriction

The primary enrollment barrier in the failed trial was a timing restriction requiring subjects to "wait to undergo cataract surgery on the fellow eye until after the study has been completed"—a constraint that excluded bilateral cataract patients unwilling or unable to delay their second surgery. Both the real-world redesign and ClinicalReTrial Agent correctly identified this as the critical obstacle and proposed its removal. Additionally, both approaches recognized the need for enrichment criteria: the real-world redesign added specific postoperative inflammation thresholds (AC cells $\geq$ Grade 2) to ensure enrolled patients exhibited measurable inflammation suitable for treatment evaluation, while ClinicalReTrial Agent proposed conceptually similar criteria targeting "mild to moderate inflammation". However, the agent failed to capture domain-specific refinements present in the real-world redesign, including

baseline safety standardization (requiring Grade 0 inflammation at screening) and operational clarity improvements (specifying exclusion of active external ocular disease). These tactical gaps highlight the agent's limitations in translating strategic insights into clinically precise protocol language.

Safety/Adverse Events. To validate agent redesign quality against real-world outcomes, we analyze NCT01919190, a Phase 4 trial of EXPAREL (liposomal bupivacaine) via TAP infiltration for post-surgical pain in lower abdominal procedures that failed due to severe adverse events (postoperative abdominal hemorrhage, 33.3% incidence). Sponsored by Pacira Pharmaceuticals, the drug was subsequently redesigned and successfully executed as NCT02199574 in a different surgical context. Table 16 compares the real-world redesign with `ClinicalReTrial` Agent's proposals.

Table 16: Real-world validation (NCT01919190, Safety/Adverse Events): We compare the real-world changes with `ClinicalReTrial` Agent's proposed modifications, categorizing alignment as: ✓(perfect match), ∼ (strategic alignment, tactical differences), or × (missed or incorrect).

Change Type	Real-World Redesign	`ClinicalReTrial` Agent Proposal	Match	Impact Level
Major Redesigns (Critical to Safety Success)				
Trial Type & Primary Outcome	PIVOTED to PK_SAFETY: original failed trial tried to prove opioid-sparing efficacy and improved OBAS scores in a heterogeneous surgical population; while modified trial completely pivoted to PK endpoints (half-life, AUC, Cmax, Tmax, λz)	MODIFIED to PK_SAFETY: "Evaluate plasma levels of bupivacaine and safety metrics following a single administration of EXPAREL"	✓	Fundamental redesign addressing root cause
Dosage Reduction	REDUCED by 50%: 266mg/20mL (60mL total volume) → 133mg/10mL (single dose, no dilution specified)	REDUCED by ∼50%: Proposed 133mg in 20mL saline per validated option (total 40mL)	✓	Correct magnitude and direction
Surgical Model	CHANGED procedure entirely: Lower abdominal surgeries (laparoscopic hysterectomy/myomectomy/colectomy with TAP infiltration) → Tonsillectomy (intraoperative infiltration to surgical site)	Missing	×	Missing
Minor Refinements (Safety Improvements, Non-Critical to Success)				
Eligibility Criteria	SIMPLIFIED: Removed all TAP-specific anatomical exclusions, complex surgical requirements, chronic opioid exclusions, pain medication washout requirements, metastatic disease exclusions, substance abuse history exclusions; retained only: hypersensitivity to local anesthetics, investigational drug washout, pregnancy/nursing exclusions, and general "significant medical conditions" clause	ADDED bleeding-specific exclusions: "Patients with history of bleeding disorders or on anticoagulant therapy" + liver dysfunction (Child-Pugh B/C) criteria; KEPT all 10 original complex exclusions including chronic opioid use, metastatic disease, substance abuse history, pain medication restrictions	×	Over-engineered restrictions vs. radical simplification

The primary safety issue in the failed trial was postoperative abdominal hemorrhage (33.3% incidence), attributed to excessive systemic exposure from high-volume TAP infiltration in hemorrhage-prone surgical sites. Both the real-world redesign and `ClinicalReTrial` Agent correctly identified the fundamental need to pivot from an efficacy trial to a PK/safety study and to reduce dosage by 50%, demonstrating strong diagnostic capability and appropriate dose-finding reasoning. However, the real-world approach implemented several structural changes largely absent from or contradicted by the agent's proposal: radical surgical model change (lower abdominal surgeries → tonsillectomy), eliminating hemorrhage-prone anatomical sites entirely rather than attempting to "broaden" or "standardize" the same problematic surgical context; drastic scope reduction to a 12 patient PK characterization study rather than maintaining Phase 4 scale; and dramatic eligibility simplification, removing 6 of 10 complex exclusion criteria (chronic opioid use, metastatic disease, substance abuse, pain medication washout, TAP-specific anatomical

concerns) to focus enrollment on the core safety profile.

Efficacy Inadequacy. To validate agent redesign quality against real-world outcomes, we analyze NCT02169336, a Phase 2 trial of intranasal Dexmedetomidine for acute post-operative pain following bunionectomy that failed due to lack of observed efficacy. Sponsored by Baudax Bio/Lotus Clinical, the trial was subsequently redesigned and successfully executed as NCT02284243. Table 17 compares the real-world redesign with `ClinicalReTrial` Agent's proposals.

Table 17: Real-world validation (NCT02169336, Efficacy Inadequacy): We compare the real-world changes with `ClinicalReTrial` Agent's proposed modifications, categorizing alignment as: ✓(perfect match), $\sim$ (strategic alignment, tactical differences), or $\times$ (missed or incorrect).

Change Type	Real-World Redesign	`ClinicalReTrial` Agent Proposal	Match	Impact Level
Major Redesigns (Critical to Efficacy Success)				
Statistical Power	INCREASED sample size: 95 $\rightarrow$ 168 participants (+77%)	INCREASE to $\sim$100 participants (power_multiplier=1.0x)	$\sim$	Correct direction, underestimated magnitude
Minor Refinements (Non-Critical to Success)				
Primary Outcome	KEPT SPID48 unchanged	KEEP SPID48 as primary outcome	✓	Preserved endpoint
Dosing Regimen	KEPT identical (35mcg & 50mcg q6h)	KEEP existing 35/50mcg dosing	✓	No modifications
Enrichment Criteria	KEPT (no biomarker screening)	ADD Central Sensitization Inventory (CSI $\geq$50) on top of existing criteria	$\times$	Unnecessary restrictiveness (would exclude 80-85%)
Enrichment Criteria	KEPT (no biomarker screening)	ADD BDNF levels ($\geq$15 ng/ml) on top of existing criteria	$\times$	Over-engineered (would exclude 80%)
Enrichment Criteria	KEPT (no genetic screening)	ADD COMT Val158Met polymorphism screening on top of existing criteria	$\times$	Invalid (flagged by validation, would exclude 70%)

The primary cause of trial failure was insufficient statistical power to detect the treatment effect, with only 95 participants enrolled. Both the real-world redesign and `ClinicalReTrial` Agent correctly identified underpowering as the root cause and proposed sample size increase as the primary solution, demonstrating strong diagnostic capability. However, the real-world approach implemented a single, decisive change—increasing enrollment to 168 participants (+77%)—while maintaining 100% protocol fidelity across eligibility criteria, primary outcomes, and dosing. In contrast, `ClinicalReTrial` Agent underestimated the required sample size (proposing $\sim$100 vs. actual 168, representing only a 5% increase) and additionally proposed layering biomarker enrichment criteria atop the existing protocol. The agent simultaneously proposed adding three unecessary new enrichment requirements. Notably, the agent's own validation system flagged the COMT polymorphism proposal as invalid due to insufficient evidence. This case illustrates a critical limitation: while `ClinicalReTrial` Agent exhibits strong strategic reasoning (correct root cause identification, appropriate prioritization of power), it defaults to mechanistic over-optimization when pragmatic simplicity proves more effective. The real-world success through power-only expansion—requiring zero design complexity—validates Occam's Razor in trial redesign: sometimes "more participants" decisively outperforms "smarter selection."

Implications. The case study reveals that `ClinicalReTrial` Agent excels at *strategic-level redesign* (identifying root causes, removing barriers, preserving safety constraints) but lacks *tactical-level domain expertise* (selecting specific biomarkers, anticipating data quality needs, distinguishing between validity-preserving and validity-threatening modifications). This suggests that future work should integrate specialized biomarker databases and safety constraint ontologies to bridge the gap between strategic reasoning and actionable clinical knowledge.

F Ethics Statement

F.1 Potential Risks

This work focuses on developing AI systems to optimize clinical trial protocols through simulation-based evaluation. While the system demonstrates potential to improve trial design efficiency, we acknowledge several important limitations and risks:

Decision Support, Not Replacement: `ClinicalReTrial` is designed as a decision support tool for clinical trial designers and should not replace human expert judgment. All system-generated protocol modifications require review by qualified medical professionals and regulatory compliance verification before real-world implementation.

Simulation Environment Limitations: Our prediction models achieve PR-AUC > 0.75, but prediction errors could lead to suboptimal redesign recommendations. The system's suggestions should be validated through standard clinical trial design processes and regulatory review.

Retrospective Validation: Our case studies demonstrate alignment with real-world redesigns but are retrospective analyses. Prospective validation in collaboration with clinical trial sponsors is necessary before deployment.

Generalization Constraints: The system is trained on historical clinical trial data and may not generalize to novel therapeutic mechanisms, rare diseases, or emerging trial paradigms not well-represented in the training data.

F.2 Data Consent

This study exclusively utilizes publicly available datasets that do not require additional consent:

- **TrialBench Dataset** (Chen et al., 2025): Publicly released benchmark containing anonymized clinical trial protocols from ClinicalTrials.gov
- **ClinicalTrials.gov**: Public registry of clinical trials maintained by the U.S. National Library of Medicine
- **PubMed**: Public database of biomedical literature abstracts
- **DrugBank** (Wishart et al., 2018): Publicly available bioinformatics and cheminformatics database
- **Disease Database** (Chen et al., 2024a): Publicly available disease ontology database

All data sources are publicly accessible and designed for research purposes. No patient-level identifiable information is used in this study. Clinical trial protocols contain only de-identified, aggregate information as required by ClinicalTrials.gov data sharing policies.

F.3 Ethics Review Board Approval

This computational study analyzes publicly available, de-identified clinical trial metadata and does not involve human subjects research, prospective clinical interventions, or collection of new patient data. The retrospective case studies (§4.3) analyze publicly registered clinical trials with outcomes already recorded in ClinicalTrials.gov, constituting secondary analysis of publicly available data exempt from human subjects research requirements.

G Use of Large Language Model

Within our data construction workflow, we utilize large language models for agent in-context learning, reasoning, and augmentations generating. Additionally, we employ LLMs such as ChatGPT to help improve the clarity and fluency of our written content.