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 accurately predict trial failure, but do not provide actionable remedies. To fill this gap, this paper proposes ClinicalReTrial, a multi-agent system 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 a hierarchical memory that captures iteration-level feedback within trials and distills transferable redesign patterns across trials. Empirically, ClinicalReTrial turns 56.7% of failed protocols into predicted successes under the simulation environment, with a mean success probability gain of 7.4% at negligible cost ($0.156 per trial). Extensive retrospective case studies further show that ClinicalReTrial recovers clinically meaningful modification patterns that align with real-world expert-driven protocol changes. The code is anonymously available at: https://anonymous.4open.science/r/ClinicalFailureReasonReTrial-DD17/.

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. (2024a); 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 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 principal modifications or enhancements informed by identified sources of risk (Dagenais et al., 2022; Baumfeld Andre et al., 2020).

To bridge this gap, in this work, as shown in Figure 1, we propose ClinicalReTrial, a self-evolving AI agent that moves beyond static prediction toward actionable intervention through the optimization of the trial protocol, while continuously

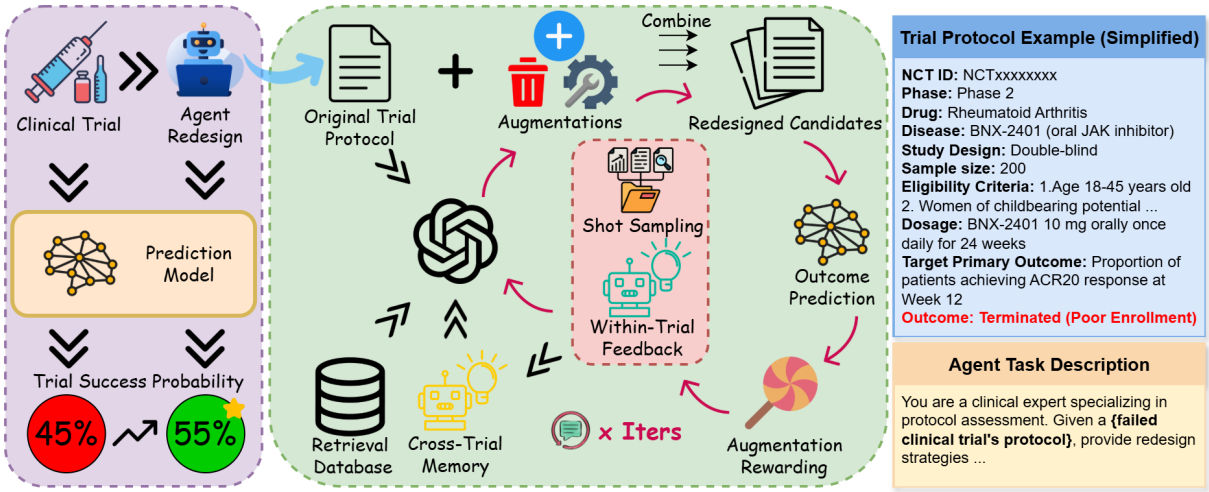

Figure 1: `ClinicalReTrial` Agent architecture. The system operates through iterative refinement: agents analyze failures, generate modifications, and receive rewards from the simulation environment. Historical explorations are extracted into structured knowledge that guides subsequent iterations, enabling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

improving its redesign policies. The language-rich nature of clinical trial protocols makes this an ideal domain for LLM-based optimization. Our framework instantiates 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 trial protocols with a mean success probability gain $\Delta p = 7.4\%$, which turns 56.7% of failed protocols into predicted successes under the simulation environment, at negligible cost ($0.156/trial). Moreover, across 8 real-world retrospective case studies, `ClinicalReTrial` rediscovers clinically meaningful redesign strategies from failed protocols, producing modifications that closely align with independently successful expert-driven trial amendments.

Main contributions (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 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, 2023) proposed HINT, integrating drug molecules, ICD-10 codes, and eligibility criteria; Chen et al. (2024c); Lu et al. (2024) added uncertainty quantification and interpretability; Wang et al. (2024); Yue et al. (2024b) 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, ClinicalAgent (Yue et al., 2024a) decomposes 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. In contrast, our multi-agent architecture leverages chain-of-thought (Wei et al., 2022) and least-to-most prompting (Zhou et al., 2023) to dynamically diagnose failure causes and refine protocols.

3 Methodology

Overview. `ClinicalReTrial` is a self-improving multi-agent system that redesigns failed clinical trials through reward-driven iterative optimization. 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 2 illustrates the process, and Algorithm 1 provided in Appendix A 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. A clinical trial protocol, including eligibility criteria and drug dosage specifications, defines the study's framework in a standardized and controlled manner. 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, 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 (Chen et al., 2024b) that naturally aligns with this structure as a closed loop performing typed editing over the protocol's textual constraints: the Trial Failure Analyzer (§3.2.1) diagnoses root causes and proposes typed edit operations; the Protocol Refinement Generator §3.2.2) realizes them as concrete protocol text; the Clinical Safety Validator (§3.2.3) filters them against external domain knowledge; and the Protocol Candidate Evaluator (§3.2.4) scores complete candidates with 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 Determination, targeting the protocol feature under consideration; then specifies the action strategy {DELETE, MODIFY, ADD} (with KEEP as the no-op default) and a confidence score.

Protocol Taxonomy. Before being processed through the agent redesign pipeline, protocol features are classified using medical terminology into structured categories: eligibility criteria, safety exclusions, selection criteria, and enrichment criteria. We further categorized dosage and outcomes by safety risk, failure contribution, and modification efficacy. These classifications guide action selection based on the observed failure reason and its corresponding analysis.

Action Determination. The agent assigns each candidate criterion a confidence score derived from its taxonomy classification and the observed failure mode. Two memory signals calibrate these scores: a cross-trial prior over edit-category effectiveness shifts confidence toward historically productive edit types (§3.3.3), while within-trial feedback from prior iterations penalizes already-explored or previously-failed criteria and rewards unexplored ones (§3.3.2). A mechanism alignment

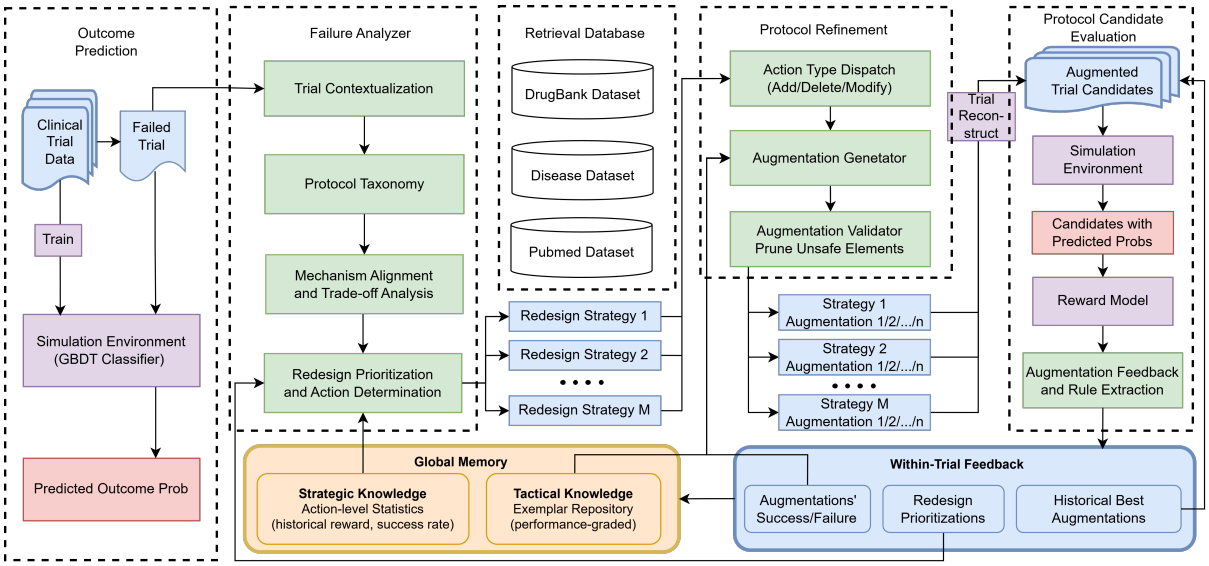

Figure 2: Detailed `ClinicalReTrial` Agent architecture and iterative redesign workflow, comprising diagnosis, augmentation, and validation, with a simulation environment and hierarchical memory to progressively optimize failed clinical trial protocols.

check synthesizes protocols within the intervention's mechanism of action. The prioritized targets, balancing domain knowledge with empirical feedback, guide the Generator Agent (§3.2.2).

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.

Type Conditioned Constrained 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 are first validated for clinical safety (§3.2.3), then proceed to formal evaluation.

3.2.3 Clinical Safety Validator

To ensure the proposed augmentations satisfy clinical safety standards, the system uses *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 in medical assessments. Validated candidates that passed the validation proceed to the Evaluator (§3.2.4) for simulation-based evaluation.

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 a success probability assigned that guides modification 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 a trial candidates' outcome predictor from encoded trial features (Appendix B), serving as a learned scalar verifier 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. While validated augmentations enter this stage, the reward is defined over a clinically admissible candidate space.

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: Local 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 integrates $\mathcal{K}_t^s$ through coverage-based confidence adjustment over the prior iteration's outcomes: criteria whose augmentations yielded positive reward are down-weighted (discouraging redundant re-editing), modification strategies whose entire augmentation set failed validation or produced negative reward are excluded, and criteria left unexplored receive a confidence bonus to encourage coverage. To prevent the search from re-converging on the same aspects, a fraction of selection slots in later iterations is reserved for criteria not selected in any prior iteration. The Generator Agent samples performance stratified exemplars $\mathcal{K}_t^t$ for few-shot prompting (Brown et al., 2020). The Evaluator maintains a redesign pool of 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, its reward distribution $\mathcal{R}_t$ is distilled via LLM synthesis into a global memory pool, so that every trial both benefits from and contributes to an evolving knowledge base (Parisi et al., 2019). For each Edit category (protocol-aspect-action) triple, the pool records the empirical success rate and mean reward over validated augmentations. This cross-trial prior is consumed at two stages:

Confidence Calibration. The Analyzer maps each candidate criterion to an edit category and shifts its confidence toward categories with high historical return, which optionally consider high-performing redesign patterns by upgrading a default KEEP criterion to MODIFY. A bounded number of additional aspect slots is reserved for calibrated criteria that the base ranking would otherwise miss.

Budget Allocation. Under a fixed per-trial generation budget, the Generator distributes samples across aspects according to each category's historical return, starving categories with consistently low validated returns. Because reallocation operates at the category level — not over specific augmentation texts — and every augmentation passes the independent Safety Validator before reward attribution, this calibration cannot bypass clinical constraints to exploit predictor artifacts.

4 Experiment

We evaluate `ClinicalReTrial` across two dimensions: (1) simulation environment performance,

validating that simulation environment achieves sufficient accuracy to serve as reliable feedback oracles, and (2) Agent optimization quality, demonstrating that our multi-agent system successfully redesigns failed trials through iterative learning.

4.1 Experimental Setup

Our evaluation consists of two parts. First, we evaluate the simulation environment at scale to verify that it provides a reliable feedback signal. Second, we evaluate whether `ClinicalReTrial` can effectively improve failed trial protocols through iterative redesign while producing protocol modifications that remain clinically plausible.

Simulation Environment. We use over 12000 annotated Phase I–IV trials from TrialBench (Chen et al., 2025). We encode multi-modal trial features into 6,173-dimensional embeddings and train LightGBM (Ke et al., 2017) classifiers to predict trial outcome. The detailed training process is deferred to Appendix B.2. We follow the official TrialBench train-test split and further split the training set 8:2 for training and validation.

Our simulation environment model is compared against *baseline* approaches, including TrialBench (Chen et al., 2025) and HINT (Fu et al., 2022), prior state-of-the-art systems operating on the original TrialBench feature space. Additionally, a Logistic Regression model is trained on the same encoded features (Appendix B.2) for reference.

Multi-Agent Redesign. We use GPT-4o-mini as our base model and evaluate the full redesign pipeline at the *protocol level*. Specifically, we focus on a stratified set of 60 failed trials from the held-out test set, with 20 trials from each failure mode: enrollment, safety, and efficacy. The sample is selected to cover different failure modes and trial phases, with additional dataset statistics reported in Appendix B.1. We use exhaustive search when the combinatorial space has fewer than 1,000 candidates, and otherwise use beam search with width $k = 8$. We measure agent effectiveness using predicted probability improvement, threshold achievement rate, and convergence efficiency, and compare `ClinicalReTrial`'s performance with GPT-4o LLM and the ReAct agent framework. We further complement the quantitative evaluation with diverse case studies against *real-world expert redesigns*, examining whether `ClinicalReTrial`'s proposed changes are clinically plausible and consistent with real-world trial design decisions.

4.2 Failure-specific Prediction

Implemented in `ClinicalReTrial` Agent, the simulation environment 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 reports 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, providing reliable discriminative feedback. All models achieve Failure Detection Rates of 69.5-74.0% 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 B.1). Also, 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) (Chen et al., 2025). Table 7 shows that the base model used in our simulation environment outperformed the baselines (Appendix B.3).

Table 1: Performance comparison across three binary prediction tasks. We use "*" to denote tasks where `ClinicalReTrial` outperform the best baseline significantly (*i.e.*, using statistical hypothesis testing).

Task	Model	ROC-AUC (↑)	PR-AUC (↑)	Fail Detection (↑)
Enrollment	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`	**0.676** ± 0.009*	**0.754** ± 0.010*	**0.740** ± 0.012*
Safety	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`	**0.656** ± 0.018*	**0.925** ± 0.007*	**0.695** ± 0.028*
Efficacy	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`	**0.746** ± 0.013*	**0.914** ± 0.007*	**0.725** ± 0.021*

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 B.4).

4.3 Protocol Optimization

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

Table 2: Probability shift analysis by failure mode. Threshold reports the share of trials whose optimized p_{final} reaches the task-specific success cutoff. p_0/p_{final}: initial/optimized success probabilities; IQR: 25th–75th percentile of Δp.

Mode	Threshold	p_0 / p_{final}	Δp	IQR
Enrollment	13/20 (65%)	0.506 / 0.609	+0.103	[+0.081, +0.131]
Safety	5/20 (25%)	0.790 / 0.851	+0.061	[+0.030, +0.079]
Efficacy	16/20 (80%)	0.820 / 0.878	+0.058	[+0.039, +0.070]
Overall	**34/60 (56.7%)**	**0.705 / 0.779**	**+0.074**	**[+0.033, +0.093]**

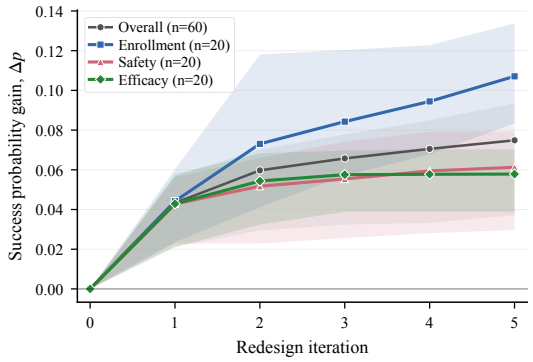

Figure 3: Iterative redesign trajectories by failure mode. Curves show mean cumulative improvement in predicted success probability (Δp).

Convergence Analysis. Table 2 reports convergence and probability-shift statistics across all failure modes. In general, the Agent consistently improves the design of the protocols in all tasks and achieves an average success probability gain of $\Delta p = +0.074$, increasing the probability of predicted success from 0.705 to 0.779 in general. Enrollment shows the largest improvement, with $\Delta p = +0.103$ and 13/20 trials exceeding the predefined improvement threshold. Overall, tested trials achieve a 56.7% threshold-passing rate, suggesting that `ClinicalReTrial` produces reliable improvements across heterogeneous failure modes.

Iterative Improvement Trajectories. We further examine the redesign trajectory over iterations. As shown in Figure 3, the Agent steadily improves predicted success probability across iterations, reaching an overall mean gain of approximately +0.074 by iteration 5. Performance heterogeneity across failure modes reflects their different levels of amenability to protocol-level redesign. Enrollment failures achieve the largest gains, suggesting that recruitment-related issues can often be improved through eligibility and design refinements. Safety and efficacy also improve, but their gains are more constrained by drug-specific safety profiles and underlying therapeutic effectiveness.

Table 3: Ablation study of the `ClinicalReTrial` framework across enrollment, safety, and efficacy failure modes. Each component is individually ablated to isolate its contribution. Reported values are pooled across the three tasks. The primary metric is Iter5 cumulative improvement, and significance is tested against the full system using one-sided paired t-tests with H_1: Full > ablation. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Method	N	Iter5 Δp	p-value	Cohen's d_z
Full System	60	$+0.0741_{\pm 0.048}$	—	—
w/o Memory	60	$+0.0532_{\pm 0.044}$	$(3.58\times10^{-8})^{***}$	0.79
w/o Pool	60	$+0.0487_{\pm 0.041}$	$(3.82\times10^{-9})^{***}$	0.87
w/o Analysis	60	$+0.0522_{\pm 0.039}$	$(3.72\times10^{-8})^{***}$	0.79
GPT-4o	60	$+0.0265_{\pm 0.019}$	$(2.99\times10^{-19})^{***}$	1.68
ReAct	60	$+0.0465_{\pm 0.039}$	$(4.81\times10^{-9})^{***}$	0.86

Moreover, 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.

Computational Cost and Efficiency. The system demonstrates practical feasibility with a mean cost of \$0.156 per trial across 60 trials, a negligible fraction of the typical \$2.6 billion drug development cost, establishing `ClinicalReTrial` as practical for systematic protocol optimization at scale. Full results on the computational cost are deferred to Appendix C.1.

Ablation Study on Self-improvement. To validate the contribution of each architectural component, we conduct paired ablation studies across enrollment, safety, and efficacy failure modes. For each ablation, we remove one component while keeping the rest of the pipeline unchanged, and evaluate Iter5 cumulative improvement as the primary metric. Results pooled across all three tasks are reported in Table 3, with task-specific ablation results provided in Section C.2 (Table 10).

The full system achieves the highest Iter5 improvement ($\Delta p = +0.0741$), while all ablations significantly degrade performance. Removing the redesign pool causes the largest drop ($\Delta p = +0.0487$, $p = 3.824\times10^{-9}$), followed by removing analysis ($\Delta p = +0.0522$, $p = 3.723\times10^{-8}$) and memory ($\Delta p = +0.0532$, $p = 3.582\times10^{-8}$). These results show that pool-based exploitation, structured diagnosis, and cross-trial memory all effectively contribute to the self-improvement ca-

NCT00294723 — FAILED

NCTID:	**NCT00294723**	Date: 2006.2
Phase:	Phase 3	**Drug: liraglutide**
Condition:	Type 2 diabetes, drug-naive or early-stage	
Target Primary Outcome:	Change in HbA1c from baseline at Weeks 52, 104, and 156	
Outcome:	Terminated, insufficient subjects	

NCT01617434 — SUCCESS

NCTID:	**NCT01617434**	Date: 2012.9
Phase:	Phase 3	**Drug: liraglutide**
Condition:	Type 2 diabetes on established basal insulin background	
Target Primary Outcome:	Change in HbA1c from baseline to Week 26	
Outcome:	Completed — Positive Outcome	

METHOD COMPARISON

ASPECT	EXPERT	CLINICAL RETRIAL	W/O ANALYZER	GPT-4o	VERDICT
Therapy requirement	**Require stable basal insulin ≥20 U/day ± metformin** Reward: **+0.017**	Eligible if currently on stable doses of insulin for 3 months Reward: +0.020	Subjects treated with insulin within the last 3 months, including those who can demonstrate stable glucose control Reward: −0.004	Allow participants with varied treatment backgrounds Pruned	Background therapy as a retention safety net.
Observation window	**Shorten 260 weeks → 26 weeks** Reward: **+0.029**	Plan different shorter milestones Reward: +0.031	Not addressed	Not addressed	Compresses the attrition exposure window.
Primary endpoint	**KEEP HbA1c Change**	Preserved	Preserved	Preserved	"HbA1c change" is the diabetes efficacy gold standard.
Protocol scope	**Only protocol-internal changes**	Only protocol-internal changes	Only protocol-internal changes	Proposes wearables, stipends, adaptive features, flexible visits, advisory boards — leaves protocol scope	Only protocol changes are fileable as an amendment.

Aligned with expert redesign Misaligned / not addressed

Figure 4: Case Study on one Poor Enrollment redesign: **(NCT00294723, 2006-02)**, and the real-world redesign **(NCT01617434, 2012-09)**. We demonstrating `ClinicalReTrial`'s strategic alignments with expert, comparing aginst w/o analyzer and zero-shot GPT-4o baseline.

pability of `ClinicalReTrial`.

4.4 Case Studies

Alignment with Real-world Trial Redesigns. We analyze diverse 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 provide insight into clinical applicability. Such trial pair cases are chosen based on the mechanistic interpretability for systematic alignment measures.

In Figure 4, we present a poor enrollment redesign case in which the same sponsor (Novo Nordisk's Victoza®) re-executed a failed design and achieved a positive readout. The study drug, *liraglutide*, is a foundational GLP-1 agonist for type 2 diabetes (T2D). `ClinicalReTrial`is compared against the expert (Ahmann et al., 2015), an ablation without the Failure Analyzer (w/o Analyzer), and zero-shot GPT-4o. We defer the remaining case studies to Appendix D. As shown in Figure 4, `ClinicalReTrial` achieves alignment with the expert: it matches the direction of primary fixes, preserves the criterion the expert deliberately leaves intact, while staying within the regulatory amendment scope. The scalar verifier reliably separates direction-aligned from direction-opposed actions by reward sign, while the upstream validator prunes operations that would compromise scientific safety.

RAG-Grounded Validator Case Studies. We further present three validator case studies to highlight the role of RAG in improving the clinical reliability of `ClinicalReTrial`. Across these examples, disease-specific, literature-based, and drug-level retrieval provides external medical evidence that enables the Validator Agent to identify unsafe or misaligned redesigns that would otherwise appear plausible at the language level. Full case descriptions are deferred to Appendix E.

5 Conclusion

In this work, we propose `ClinicalReTrial`, a 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. Extensive case studies on trial pairs validate that `ClinicalReTrial` exhibits strong strategic alignment with independently derived real-world trial modifications, suggesting generalizable principles for protocol redesign.

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, 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.

References

A. Ahmann, H. W. Rodbard, J. Rosenstock, J. T. Lahtela, L. de Loredo, K. Tornøe, A. Boopalan, M. A. Nauck, and NN2211-3917 Study Group. 2015. Efficacy and safety of liraglutide versus placebo added to basal insulin analogues (with or without metformin) in patients with type 2 diabetes: a randomized, placebo-controlled trial. *Diabetes Obes Metab*, 17(11):1056–64. Epub 2015 Sep 10.

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, and Jimeng Sun. 2023. Automated prediction of clinical trial outcome. US Patent App. 17/749,065.

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).

Yingzhou Lu, Tianyi Chen, Nan Hao, Capucine Van Rechem, Jintai Chen, and Tianfan Fu. 2024. Uncertainty quantification and interpretability for clinical trial approval prediction. *Health Data Science*, 4:0126.

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. 2024a. 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.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024b. Trialenroll: Predicting clinical trial enrollment success with deep & cross network and large language models. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9.

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.

Appendix

A `ClinicalReTrial` Algorithm **12**

B Simulation Environment Details **12**

B.1 Dataset Statistics 12

B.2 Encode Details 13

B.3 Trial Bench Prediction Task 14

B.4 Ablation Study 15

C Deferred Self-Evolving Multi-Agent Experiments **16**

C.1 Computation Cost 16

C.2 Full Ablation Study 16

D Deferred Full Case Study **16**

E Deferred Full RAG-Grounded Case Studies **21**

F Analyzer Agent Details **22**

G Generator Agent Details **25**

G.1 Few-shot Learning Mechanism 25

G.2 Modification Prompts by Failure Mode 26

H Agent Output Template **27**

I Ethics Statement **30**

I.1 Potential Risks 30

I.2 Data Consent 30

I.3 Ethics Review Board Approval 30

J Use of Large Language Model **31**

A `ClinicalReTrial` Algorithm

Algorithm 1 Clinical trial 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_best ← 0, T* ← T_0, M_0^local ← ∅, H_0 ← ∅
for t = 1 to N_max do
    K_t^s, K_t^t ← LoadMemory(M^global[y], t)
    S_t ← AnalyzerAgent(T_{t-1}, y, M_{t-1}^local, K_t^s)
    A_t ← GeneratorAgent(S_t, T_{t-1}, M_{t-1}^local, K_t^t)
    R_t, r_max ← ExploreSearch(A_t, H_{t-1}, T_{t-1})
    if r_max > r_best then r_best ← r_max, T* ← arg max_{T'∈𝕋} f_θ(T')
    K_t ← DistillKnowledge(R_t); H_t ← H_{t-1} ∪ ExtractPool(R_t)
    M_t^local ← M_{t-1}^local ∪ {K_t, R_t, H_t}
end for
M^global ← M^global ∪ TransferMemory(T*, M_{N_max}^local);
return T*, r_best
```

B Simulation Environment Details

B.1 Dataset Statistics

Dataset used for training the prediction models comprises 20,769 clinical trials from TrialBench's failure reason dataset. Table 4 shows the label distribution across four categories.

Table 4: 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 5 presents the phase composition.

Table 5: 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

B.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 B).

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 B).

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 6, features are encoded into total dim=6,173.

Table 6: Feature specification summary. Novel contributions include sentence-level eligibility parsing and fine-tuned molecular-disease encoders.

Category	Component	Dim	Method
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.

B.3 Trial Bench Prediction Task

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) (Chen et al., 2025). Table 7 shows that the base model used in our simulation environment outperformed most baselines, achieving a ROC-AUC improvement of 6% to 19%.

Table 7: 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**

B.4 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.

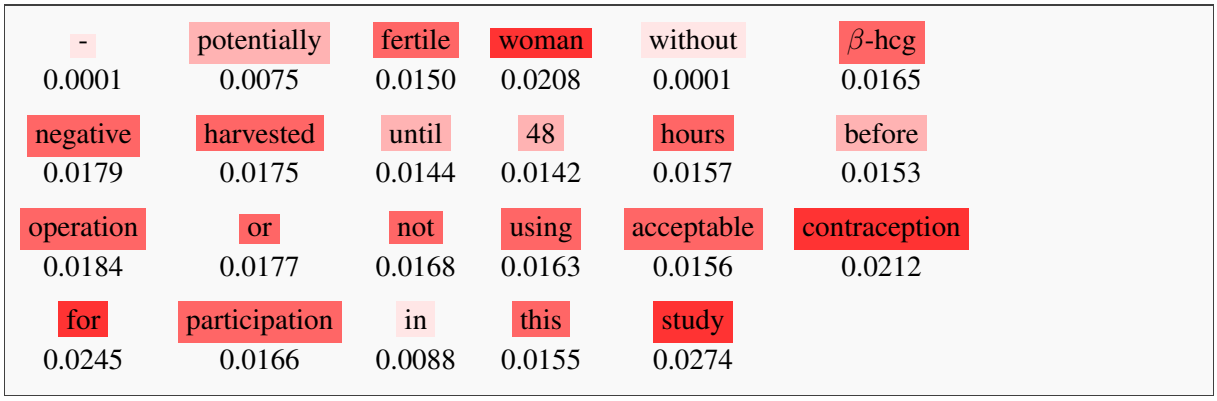

Figure 5: Visualization of text segments in the BioBERT encoder's output, illustrating Shapley values derived from Clinical Trials. Shapley values correspond to attention weights, with darker colors indicating higher weights.

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.

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.

Table 8: 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

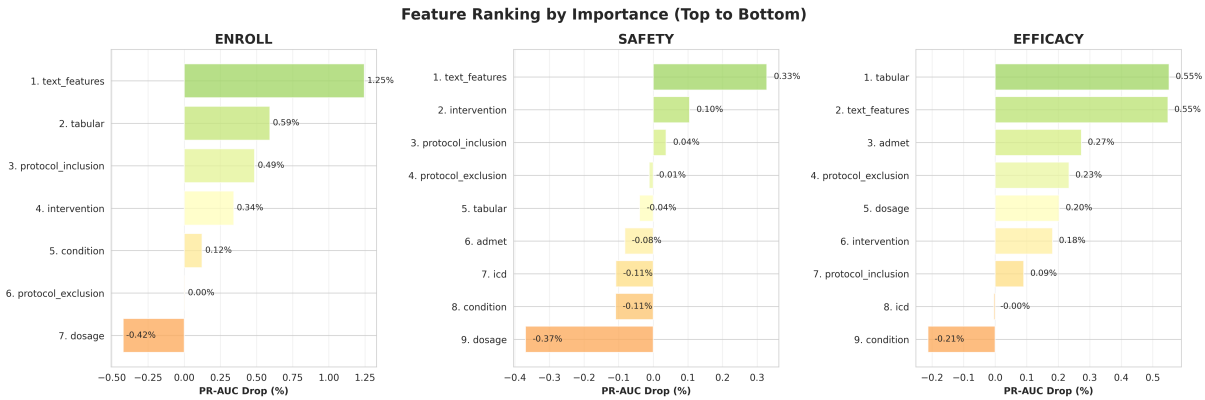

Figure 6: Feature importance of each encoder on 3 classification tasks (enrollment, safety, and efficacy), measured by PR-AUC drop when the encoder is masked out during prediction.

C Deferred Self-Evolving Multi-Agent Experiments

C.1 Computation Cost

Table 9 summarizes the per-trial computational cost of `ClinicalReTrial` across the three failure modes. Overall, the optimization process costs only $0.156 per trial on average, indicating that the proposed self-evolving search procedure remains inexpensive even when applied across diverse protocol redesign tasks. The cost-efficiency ratio is also stable across modes, with an overall cost of $2.11 per unit improvement in predicted success probability.

C.2 Full Ablation Study

This section provides the task-specific ablation results complementing the pooled analysis in Table 3. As shown in Table 10, the full system consistently achieves the highest Iter5 improvement across all three failure modes. Each ablated variant leads to a statistically significant drop, indicating that memory, pool-based exploitation, and structured analysis each contribute to the redesign performance across different trial failure types.

D Deferred Full Case Study

We further validate `ClinicalReTrial` Agent's reasoning against additional real-world protocol redesigns from the training-pair audits. These cases complement the appendix examples by covering two efficacy-inadequacy failures and three poor-enrollment failures. For cases where the agent correctly diagnosed

Table 9: Computational cost and efficiency across failure modes. Cost and Δp report per-trial averages; Cost/Δp measures dollars spent per unit improvement in predicted success probability.

Mode	**N**	**Cost ($)**	Δp	**Cost/Δp**
Enrollment	20	0.247	+0.103	2.40
Safety	20	0.144	+0.061	2.35
Efficacy	20	0.078	+0.058	1.34
Overall	**60**	**0.156**	**+0.074**	**2.11**

Table 10: Ablation study of the `ClinicalReTrial` framework on three failure modes: enrollment, safety, and efficacy. The primary metric is Iter5 cumulative improvement. For each ablated variant, we report the one-sided paired t-test against the full system with alternative hypothesis $H_1 : \Delta p_5^{\text{ablation}} < \Delta p_5^{\text{full}}$. Effect sizes are reported as Cohen's d_z. Significance levels are denoted by $^*p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Task	**Method**	**N**	**Iter5 Δp**	**p-value**	**Cohen's d_z**
Enrollment	Full System	20	$+0.1032_{\pm 0.0438}$	—	—
	w/o Memory	20	$+0.0624_{\pm 0.0454}$	$(2.289\times10^{-6})^{***}$	1.41
	w/o Pool	20	$+0.0503_{\pm 0.0392}$	$(7.035\times10^{-8})^{***}$	1.81
	w/o Analysis	20	$+0.0540_{\pm 0.0338}$	$(1.131\times10^{-8})^{***}$	2.04
Safety	Full System	20	$+0.0613_{\pm 0.0436}$	—	—
	w/o Memory	20	$+0.0521_{\pm 0.0464}$	0.002^{**}	0.72
	w/o Pool	20	$+0.0500_{\pm 0.0443}$	$(7.632\times10^{-5})^{***}$	1.05
	w/o Analysis	20	$+0.0545_{\pm 0.0446}$	0.009^{**}	0.58
Efficacy	Full System	20	$+0.0578_{\pm 0.0457}$	—	—
	w/o Memory	20	$+0.0450_{\pm 0.0394}$	0.011^{*}	0.56
	w/o Pool	20	$+0.0458_{\pm 0.0415}$	0.013^{*}	0.54
	w/o Analysis	20	$+0.0480_{\pm 0.0412}$	0.029^{*}	0.45

an issue but could not express it as a typed augmentation, we separate diagnosis-layer alignment from action-layer alignment.

Efficacy Inadequacy: Brivaracetam in Partial-Onset Seizures. We analyze NCT00699283, a Phase 3 brivaracetam conversion-to-monotherapy study in adults with partial-onset seizures. The failed design tapered background antiepileptic drugs and attempted to show monotherapy efficacy against a fixed historical-control withdrawal benchmark. Successful brivaracetam efficacy studies instead used adjunctive therapy on stable background antiseizure medications, a concurrent placebo control, and seizure-frequency reduction outcomes. Table 11 compares this redesign with `ClinicalReTrial` Agent's proposals.

The agent's strongest contribution in this case is strategic: it correctly moved away from a pure power explanation and recognized that a fixed historical-control benchmark made the original efficacy target structurally difficult. It also moved the dose in the clinically plausible direction. However, the central real-world fix was not simply "more dose" or "a better endpoint"; it was changing how brivaracetam was used relative to existing therapy. Maintaining background antiseizure medications is clinically important because patients with uncontrolled focal seizures are vulnerable to seizure worsening during AED withdrawal, and stable background therapy creates a more interpretable add-on efficacy signal while the agent missed this preservation.

Efficacy Inadequacy: Intranasal Dexmedetomidine in Post-Bunionectomy Pain. We analyze NCT02169336, a Phase 2 trial of intranasal dexmedetomidine (DEX-IN) 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 12 compares the real-world redesign with `ClinicalReTrial` Agent's proposals.

The primary cause of trial failure was insufficient statistical power: only 95 participants were spread

Table 11: Real-world validation (NCT00699283, Efficacy Inadequacy): We compare the real-world brivaracetam redesign pattern with `ClinicalReTrial` Agent's proposed modifications, categorizing alignment as: ✓(match), ∼ (strategic alignment, tactical differences), or × (missed or incorrect).

Change Type	Real-World Redesign	`ClinicalReTrial` Agent Proposal	Match	Impact Level
Major Redesigns (Critical to Efficacy Success)				
Treatment Paradigm	SHIFTED from conversion-to-monotherapy with background AED taper to adjunctive brivaracetam while maintaining stable concomitant antiseizure medications	Missing: agent treated dose and comparator as modifiable, but did not propose keeping background therapy and testing brivaracetam as add-on treatment	×	Major, missed linchpin regimen redesign
Comparator and Endpoint	REPLACED historical withdrawal benchmark with **concurrent placebo comparison**; measured percent reduction in adjusted seizure frequency and responder rate	DIAGNOSED external benchmark as too ambitious and proposed **concurrent control plus event-frequency reduction endpoint**; endpoint augmentations were not retained in final tracker	∼	Strong diagnosis, incomplete action preservation
Dose Range	MOVED toward **higher** effective adjunctive doses (100–200 mg/day range in pivotal trials) rather than low-dose monotherapy	MODIFIED dose **upward**, dropping the lowest arm and adding higher-exposure options	✓	Directionally correct pharmacologic adjustment
Preservations and Tactical Gaps)				
Background Therapy Stability	PRESERVED stable baseline antiseizure therapy to avoid destabilizing refractory epilepsy patients during efficacy assessment	Missing: agent did not recognize that preserving background therapy was itself assay-sensitivity protection	×	Important clinical preservation missed

Table 12: Real-world validation (NCT02169336, Efficacy Inadequacy): We compare the real-world dexmedetomidine redesign with `ClinicalReTrial` Agent's proposed modifications, categorizing alignment as: ✓(match), ∼ (strategic alignment, tactical differences), or × (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 → 168 participants (**+77% increase**)	DIAGNOSED underpowering and **increased sample size** to ∼100 participants	∼	Correct direction, underestimated magnitude
Preservations (Critical NOT to Change)				
Dosing Regimen	KEPT identical (35mcg&50mcg q6h)	KEEP existing 35/50mcg dosing	✓	Preserved endpoint
Eligibility Criteria	KEPT **original bunionectomy eligibility unchanged**; no biomarker enrichment	Step-3 recommended **KEEP for inclusion/exclusion criteria**, but lower-ranked augmentations still proposed NRS, pain-catastrophizing, or biomarker enrichment	∼	Correct preservation signal with enrichment noise

across two active doses and placebo, leaving small per-arm sample sizes for a continuous pain endpoint. The real-world redesign was deliberately simple: scale enrollment, focus on the 50 mcg dose, remove the PRN extension that could dilute the fixed 48-hour efficacy signal, and preserve SPID48 and eligibility. `ClinicalReTrial` effectively identified POWER as the root cause and recognized SPID48 as an appropriate endpoint and proposed to increase the sample size to ~ 100, which shows strong strategic alignment but with minor magnitude shift.

Safety/Adverse Events: EXPAREL in Post-Surgical Pain. We analyze NCT01919190, a Phase 4 trial of EXPAREL (liposomal bupivacaine) via transversus abdominis plane (TAP) infiltration for post-surgical pain in lower abdominal procedures that failed due to severe adverse events, including postoperative abdominal hemorrhage. Sponsored by Pacira Pharmaceuticals, the drug was subsequently redesigned and successfully executed as NCT02199574 in a different surgical context. Table 13 compares the real-world redesign with `ClinicalReTrial` Agent's proposals.

The primary safety issue in the failed trial was postoperative abdominal hemorrhage, plausibly linked to high-volume TAP infiltration in hemorrhage-prone lower-abdominal surgical sites. Both the real-world

Table 13: Real-world validation (NCT01919190, Safety/Adverse Events): We compare the real-world EXPAREL redesign with `ClinicalReTrial` Agent's proposed modifications, categorizing alignment as: ✓(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** characterization: original failed trial attempted to prove opioid-sparing efficacy and improved OBAS scores in a heterogeneous surgical population	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%**: 266 mg/20 mL (60 mL total volume) → 133 mg/10 mL single dose	**REDUCED by ∼50%**: proposed 133 mg in 20 mL saline per validated option (40 mL total volume)	✓	Correct magnitude and direction
Surgical Model	CHANGED procedure entirely: lower abdominal surgeries with TAP infiltration → tonsillectomy with local surgical-site infiltration	Missing: agent did not propose leaving the hemorrhage-prone lower-abdominal/TAP context	×	Missed anatomical risk elimination
Minor Refinements (Safety Improvements, Non-Critical to Success)				
Eligibility Criteria	SIMPLIFIED: removed TAP-specific anatomical exclusions, complex surgical requirements, chronic opioid exclusions, pain-medication washout, metastatic disease, and substance-abuse exclusions; retained core safety exclusions	ADDED bleeding-specific exclusions and liver dysfunction criteria, while keeping the original complex exclusion structure	×	Over-engineered restrictions vs. radical simplification
Safety Focus	PRESERVED **key safety exclusions** such as local-anesthetic hypersensitivity, pregnancy/nursing, investigational-drug washout, and significant medical conditions	Correctly emphasized bleeding and systemic-exposure risk, but did so by **tightening exclusions** within the wrong surgical model	∼	Clinically relevant concern, wrong implementation level

redesign and `ClinicalReTrial` Agent correctly identified two central safety moves: pivoting from an efficacy trial to PK/safety characterization and reducing the EXPAREL dose by approximately half. These are strong clinically plausible alignments because they directly address systemic bupivacaine exposure and safety monitoring. On the other hand, the expert redesign went further by changing the surgical model entirely, thereby removing the anatomical context in which hemorrhage had emerged. The agent instead tried to manage risk within the same broad surgical context by adding bleeding-specific restrictions. That proposal is clinically understandable, but it misses the structural safety move: eliminate the risky use context rather than merely screening around it.

Poor Enrollment: Mircera in CKD Anemia. We analyze NCT00576602, a Mircera (C.E.R.A.) trial for anemia in chronic kidney disease after kidney transplantation. The failed trial targeted a narrow post-transplant population and randomized anemic patients to Mircera versus no ESA therapy. The successful redesign (NCT01342640) shifted to a broader and more accessible pre-dialysis CKD anemia population with eGFR-defined eligibility and a simpler open-label structure. Table 14 compares the redesign with `ClinicalReTrial` Agent's proposals.

The agent captured several local signals from the failed protocol: the post-transplant population was enrollment-limiting, the randomized no-ESA control could deter anemic CKD patients, and Hb $<$10 g/dL was a clinically meaningful severity threshold for observing a C.E.R.A. response. These points align with parts of the successful redesign. The main gap is that the expert redesign solved access by shifting to a pre-dialysis CKD population defined by eGFR 20–60 mL/min, rather than by further relaxing the original post-transplant criteria. Preserving the ESA washout also remained important because recent ESA exposure could confound hemoglobin response attribution; therefore, the agent's ESA-allowance proposals were clinically plausible as enrollment aids but misaligned with the final successful protocol.

Poor Enrollment: Stannsoporfin in Neonatal Jaundice. We analyze NCT00850993, a Stannsoporfin trial in neonatal hyperbilirubinemia. The failed trial recruited a broad all-comer neonatal jaundice

Table 14: Real-world validation (NCT00576602, Poor Enrollment): We compare the real-world Mircera redesign with `ClinicalReTrial` Agent's proposed modifications, categorizing alignment as: ✓(match), ∼ (strategic alignment, tactical differences), or × (missed or incorrect).

Change Type	Real-World Redesign	`ClinicalReTrial` Agent Proposal	Match	Impact Level
Major Redesigns (Critical to Enrollment Success)				
Population Scope	SHIFTED to a **larger, easier-to-access** CKD anemia pool: from patients 6 months–5 years post-kidney transplant to pre-dialysis CKD patients	Recognized that the original post-transplant population restricted enrollment and proposed **broader access** but implemented broadening only through transplant-timing and anemia-threshold changes	∼	Matched broadening rationale, missed target population
Renal Function Definition	ADDED core eGFR eligibility (not on dialysis; eGFR 20–60 mL/min)	Missing: no eGFR criterion proposed because the failed trial did not contain an eGFR row to modify	×	Missing
RCT / No-Treatment Barrier	REMOVED the randomized Mircera versus **no-ESA control** by moving to a single-arm open-label design, reducing reluctance to risk no active anemia treatment	Identified the randomized **no-ESA control** as an ethical and recruitment barrier for anemic CKD patients, but could not translate this diagnosis into the expert's single-arm open-label redesign	∼	Matched ethical-barrier diagnosis, missed design action
Dose Schedule	MODIFIED Mircera exposure from 0.6 mcg/kg q2w to 1.2 mcg/kg q4w	No dosing schedule change are proposed	×	Missing
Preservations and Misalignments)				
ESA Washout	PRESERVED no ESA therapy during the 3 months before study start to measure Mircera response cleanly	Proposed allowing stable ESA-treated patients within 3 months, often with Hb stability conditions	×	Misaligned; would confound hemoglobin response
Hemoglobin Severity Threshold	REQUIRED anemia severity with **Hb<10 g/dL** in the successful pre-dialysis CKD population	Proposed Hb severity thresholds overlapping the successful protocol, including **Hb<10 g/dL** or narrow Hb ranges near 8.0–10.5 g/dL	✓	Captures severity signal

population, while the successful redesign (NCT01887327) enriched for neonates with hemolysis indicators and added a mandatory phototherapy backbone. Table 15 summarizes alignment.

The successful redesign narrowed enrollment to neonates with hemolysis-driven hyperbilirubinemia while preserving key neonatal safety guardrails. `ClinicalReTrial` Agent identified the correct enrichment target, repeatedly detected the need for hemolysis-based enrichment, and generated a close Rh/ABO plus TSB-threshold criterion. It also recognized the recruitment barrier created by the lack of guaranteed phototherapy and correctly preserved the major abnormality/infection exclusion. However, these aligned signals were not promoted into the final redesign: the agent did not add mandatory phototherapy, missed the 48-hour bilirubin endpoint refinement, and favored broadening changes.

Poor Enrollment: Daprodustat Renal Pharmacokinetics. We analyze NCT01406340, a daprodustat (GSK1278863) renal pharmacokinetic study. The failed protocol required parallel recruitment across renal-function cohorts, including a difficult-to-fill end-stage renal disease hemodialysis cohort. The successful redesign (NCT02243306) simplified the study into a standalone peritoneal-dialysis ESRD cohort. Table 16 compares the real-world redesign with `ClinicalReTrial` Agent's proposals.

The real-world fix was structural simplification: the failed protocol required simultaneous recruitment across multiple renal-function cohorts, so a stalled Stage 5 hemodialysis cohort could block the entire study. The successful redesign removed this cross-cohort dependency by converting the study into a single focused ESRD cohort and further refocusing recruitment on peritoneal dialysis patients. `ClinicalReTrial` Agent aligned at the diagnostic level by identifying the failure as a structural/logistical design problem and by generating renal-function threshold modifications, indicating that it recognized renal-function definition as the relevant clinical axis. However, its proposed changes remained criterion-local: reducing visit burden, relaxing CrCl or renal-impairment thresholds, and selecting peripheral boilerplate criteria rather than decoupling cohorts or switching dialysis modality.

Table 15: Real-world validation (NCT00850993, Poor Enrollment): We compare the real-world Stannsoporfin redesign with `ClinicalReTrial` Agent's proposed modifications, categorizing alignment as: ✓(match), ∼ (strategic alignment, tactical differences), or × (missed or incorrect).

Change Type	Real-World Redesign	`ClinicalReTrial` Agent Proposal	Match	Impact Level
Major Redesigns (Critical to Enrollment and Assay Sensitivity)				
Operational Hemolysis Criterion	REQUIRED specific hemolysis indicators (**ABO/Rh incompatibility** with Coombs positivity or G6PD deficiency) with bilirubin-centered assessment	Specified **Rh/ABO incompatibility** plus baseline TSB ≥9 mg/dL but was not promoted (rank too low)	∼	Strong partial action-layer match, ranked too low
Background Therapy	ADDED **mandatory phototherapy backbone** for all arms to remove ethical concern about withholding standard care	Failure analysis noted that families may **prefer phototherapy**, but no augmentation required phototherapy across arms	∼	Diagnosis hit, action miss
Endpoint Refinement	REFINED endpoint to short-term 48-hour total serum bilirubin percent change	No endpoint refinement proposed	×	Missed efficacy-readout clarification
Preservations (Critical Safety Guardrails)				
Infection / Major Abnormality Safety	PRESERVED exclusions for **major abnormalities**, clinically significant infection, or serious neonatal conditions	Correctly KEEP-locked the **abnormalities/infections** exclusion; validator also invalidated ADD proposals that would include neonates with manageable infections	✓	Important neonatal safety preservation
Gestational Age and Weight	PRESERVED near-term/term eligibility and birth weight ≥2500 g	Proposed lowering gestational-age and weight thresholds to broaden access	×	Misaligned; weakens neonatal safety guardrails
QT/Cardiac Safety	PRESERVED and strengthened exclusions for QT-prolonging medications and cardiac risk	Relaxed QT-medication exclusion	×	Clinically concerning in vulnerable neonates

E Deferred Full RAG-Grounded Case Studies

Case 1: Disease-RAG prevents misaligned safety filtering in a Type 2 diabetes trialcolback

Clinical trial case. *Trial:* NCT00294723. *Condition:* Type 2 diabetes (liraglutide safety evaluation). *Proposed redesign:* exclude patients with eGFR < 60 mL/min at screening.

Retrieved evidence (Disease-RAG). *Query:* `Diabetes`. Retrieved guidelines emphasized management of **cardiovascular disease risk factors** as a core part of diabetes care.

Validation outcome. *Decision:* **Invalid redesign.**

Reason. Combined with the adverse event summary, which included myocardial infarctions, the retrieved disease context showed that the proposed renal exclusion was overly broad and not well aligned with the trial's main safety risks. The Validator Agent therefore pruned the modification.

Case 2: PubMed-RAG prevents unsafe diagnostic relaxation in a hypertension-in-pregnancy trialcolback

Clinical trial case. *Trial:* NCT03872336. *Condition:* Hypertension in pregnancy (labetalol treatment). *Proposed redesign:* accept a single severe blood pressure reading ($\geq 160/110$ mmHg) for diagnosis.

Retrieved evidence (PubMed-RAG). *Query:* `Hypertension diagnosis measurement criteria`. PMID 29574470 and PMID 24652215 both support diagnosis based on repeated blood pressure measurements over time.

Validation outcome. *Decision:* **Invalid redesign.**

Reason. The Validator Agent determined that accepting a single severe reading is inconsistent with standard diagnostic practice and may cause false-positive diagnoses, unnecessary intervention, or inappropriate enrollment.

Table 16: Real-world validation (NCT01406340, Poor Enrollment): We compare the real-world daprodustat renal-PK redesign with `ClinicalReTrial` Agent's proposed modifications, categorizing alignment as: ✓(match), ∼ (strategic alignment, tactical differences), or × (missed or incorrect).

Change Type	Real-World Redesign	`ClinicalReTrial` Agent Proposal	Match	Impact Level
Major Redesigns (Critical to Enrollment Success)				
Cohort Structure	DECOUPLED the **multi-cohort renal-impairment design** (normal, Stage 3, Stage 4, Stage 5 HD) into a single focused ESRD cohort, eliminating the cross-cohort recruitment bottleneck	Correctly diagnosed enrollment failure as a **structural/logistical design** problem, but translated this into visit-burden reduction and threshold relaxation rather than cohort decoupling	∼	Matched structural failure category, missed exact cohort-level fix
Population Refocusing	SWITCHED from the stalled hemodialysis cohort to peritoneal dialysis ESRD patients	Missing: no PD-specific population refocusing was proposed	×	Missing
Renal-Function Axis	FOCUSED eligibility on a clinically specific **renal-function** stratum: ESRD patients receiving peritoneal dialysis	Generated CrCl and renal-impairment threshold modifications, showing recognition that **renal-function** definition was relevant, but broadened thresholds rather than focusing on PD ESRD	∼	Correct axis recognized, wrong modality and direction
Subgroup Interdependency	REMOVED requirement to complete parallel renal-function strata, eliminating last-cohort bottleneck	Missing: No action proposed to split or decouple cohorts.	×	Missing
Preservations and Peripheral Misalignments				
Clinical Aspect Selection	Needed cohort-level design aspects rather than header text or boilerplate criteria	Agent often selected section headers/preamble and peripheral rows as modifiable aspects	×	Data/selection failure limits clinical utility

Case 3: Drug-RAG preserves a critical safety exclusion in a labetalol trialcolback

Clinical trial case. *Trial:* NCT03872336. *Condition:* Hypertension in pregnancy (labetalol treatment). *Proposed redesign context:* remove or weaken the exclusion criterion on persistent mild–moderate asthma.

Retrieved evidence (Drug-RAG). *Query:* `labetalol toxicity`. Retrieved safety evidence noted risk of hypotension, bradycardia, and bronchospasm.

Validation outcome. *Decision:* **Preserve existing criterion.**

Reason. The Validator Agent identified a drug–disease interaction between labetalol and asthma risk. Because bronchospasm is a foreseeable adverse effect, removing the asthma exclusion would be clinically unsafe.

F 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 17 summarizes failure-mode-specific adaptations.

Table 17: 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>
```

G Generator Agent Details

Table 18: 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

G.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.

G.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>
```

H 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]"
          ]
        }
      }
    ]
  }
```


I Ethics Statement

I.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.

I.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.

I.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 (§4.4) 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.

J Use of Large Language Model

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