DEEPIMAGINE: CLINICAL TRIAL OUTCOME PREDICTION VIA STEPWISE LOCAL COUNTERFACTUAL IMAGINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the outcomes of prospective clinical trials remains a major challenge. Clinical trial outcomes result from complex interactions among experimental factors such as drug interventions, participant demographics, and protocols. Here, we introduce DeepImagine, a framework that predicts target trial outcomes through stepwise counterfactual imagination anchored on historical trials with observed results. Starting from a relevant historical trial, DeepImagine sequentially modifies one differing experimental factor at a time. With each step a large language model (LLM) is posed a local counterfactual: how would the current imagined outcome change with this single perturbation? The updated result is carried forward as the input to the next step, until the historical configuration matches the target more closely, yielding the final prediction. Empirically, DeepImagine consistently outperforms traditional clinical trial outcome prediction methods and direct one-step prediction across several off-the-shelf LLMs, with further gains when multiple imagination pathways, initiated from different historical anchors, are aggregated. We also construct natural counterfactuals augmented with synthetic reasoning traces and train a family of specialized language models, each dedicated to learning one factor's local counterfactual transition. Integrating these learned local operators into DeepImagine yields substantial improvements over general-purpose LLM baselines. Our findings position stepwise counterfactual imagination, distinct from both correlational prediction and explicit structural causal modeling, as a promising direction for clinical trial outcome prediction. All code, training scripts, and evaluation scripts are available at https://github.com/deepimagine-counterfactual/DeepImagine.

1 INTRODUCTION

Despite rapid advances in artificial intelligence (AI) and biomedical science, predicting the outcome of a prospective clinical trial remains a notoriously difficult problem. Our analysis of 14,451 recent Phase II and Phase III drug trials with posted results on clinicaltrials.gov reveals that 50.45% failed to achieve their stated primary objectives, directly affecting 3,639,346 enrolled participants. Recent work (Wang et al., 2026) further shows that AI models perform poorly on this task. To fill this research gap, we introduce a novel framework for both inference and training of language models that substantially improves prediction accuracy.

Even a closely related historical trial may differ in its regimen and endpoint. An overall-survival result cannot simply be copied into a target outcome asking about the progression-free survival under another treatment. Retrieval methods provide a relevant historical anchor. The remaining problem is how to adapt that historical observation to the target trial of interest. In practice, trial outcomes depend on intricate interactions among many factors that no closed-form model can fully capture. This likely explains the absence of successful applications of classical causal inference in this setting: the standard pipeline requires specifying a complete structural causal model or learning a causal DAG from data (Pearl, 2009; Heckerman et al., 1999), both of which presuppose a tractable causal structure that does not exist here. Direct language-model prediction fares no better, and recent retrieval-augmented (RAG) approaches (Wang et al., 2026) that supplement an LLM with similar historical trials yield only minimal and unstable improvements.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

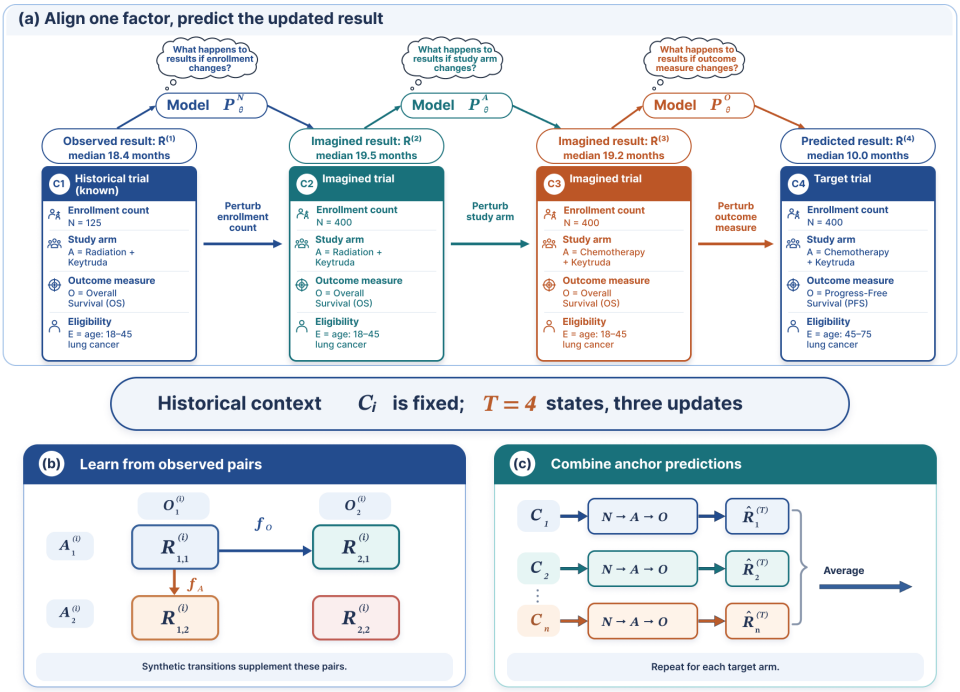

Figure 1: DeepImagine uses one local transition for prediction and supervision. **(a)** A historical result is updated as enrollment, study arm, and endpoint are perturbed towards the target trial. Only the starting result is observed. **(b)** Within-trial result pairs provide local transitions when other attributes match. **(c)** Different anchors produce separate predictions for each target arm. Results are averaged.

Inspired by Wang et al. (2026), we propose to learn how the observed result of a historical trial would change if its experimental factors were modified to match the target trial's configuration more closely. This bypasses the need for a structural causal model: the model only needs to learn the hidden mechanism governing this counterfactual transition. However, supervised fine-tuning on this task is empirically ineffective because any historical trial differs from the target along too many factors at once. We therefore propose **DeepImagine**, which decomposes the prediction into a Markov chain of single-factor perturbations: starting from a historical trial with its observed result, DeepImagine modifies one experimental factor at a time to match the target, and at each step the language model predicts how the current intermediate result would change if only that single factor were perturbed, *ceteris paribus*, i.e., holding all others fixed (Figure 1). We refer to each step as a local counterfactual imagination, using "counterfactual" in a procedural sense rather than in the formal sense of Pearl (2009). Empirically, we only change three experimental factors: enrollment count N, study arm A, outcome measure O, as we can learn local counterfactual transitions for these factors because each controlled clinical trial already contains naturally occurring counterfactual experiments on them (Section 4.3). We leave learning to change other factors to future work.

Our experiments yield three main findings. First, DeepImagine generalizes effectively to off-the-shelf LLMs such as GPT-5, producing consistent improvements over one-shot RAG-style prediction and traditional clinical trial outcome prediction methods, indicating that even general-purpose LLMs predict more accurately when prompted to reason via local counterfactual imagination. Second, aggregating predictions across multiple pathways anchored on different historical trials yields significant additional gains; this is not an artifact of majority voting, as multiple pathways anchored on the same trial do not produce the same improvement. Third, language models trained within the DeepImagine framework substantially outperform baselines fine-tuned to predict outcomes directly, suggesting that learning a single-factor transition is far easier than learning the joint transition over all factors at once.

2 Related Work

Clinical trial prediction and evaluation. Clinical trial prediction spans statistical imputation and structured features (Lo et al., 2019; Siah et al., 2021) to models of drug–disease–protocol interactions (HINT; Fu et al. 2022), sequential trial structure (SPOT; Wang et al. 2023c), and multimodal representations (Zhang et al., 2025; Zheng et al., 2025). PyTrial provides a broader evaluation toolkit (Wang et al., 2023b), and CTP-LLM predicts phase transitions from protocol text (Reinisch et al., 2024). For data, we use the Summer 2025 cohort of CT Open (Wang et al., 2026), in which no trial had results released before 2025-08-31; this temporal separation reduces, but does not formally rule out, pretraining contamination.

Grounded, decomposed, and counterfactual reasoning. Prior work grounds LLM predictions in retrieved or self-generated examples (Lewis et al., 2020; Wang & Sun, 2022; Wang et al., 2026; Yasunaga et al., 2024), decomposes problems into simpler or specialized subtasks (Zhou et al., 2023; Khot et al., 2023; Yue et al., 2024), aggregates sampled reasoning paths (Wang et al., 2023a), and supervises reasoning with generated reasoning or step-level correctness labels (Zelikman et al., 2022; Lightman et al., 2024). DeepImagine is related to these ideas but gives each step concrete state. Starting from an observed historical result, each operator receives the current result and a single configuration change and predicts the next result. We vary the observed starting trial and average arm-level predictions before the final comparison. For supervision, natural counterfactual pairs link two reported trial results through a generated explanation, and synthetic pathways add generated intermediate results without step-level correctness labels. Classical causal inference answers interventional and counterfactual questions under explicit assumptions about a data-generating process (Pearl, 2009; Heckerman et al., 1999), and LLM studies examine causal and counterfactual reasoning in various forms (Kiciman et al., 2024; Jin et al., 2023; Miliani et al., 2025; Chen et al., 2026; Vashishtha et al., 2025). In contrast, DeepImagine is given no causal model, and its single-factor updates are model-based approximations rather than identified causal effects.

3 Preliminary

A trial C_i provides study arms, enrollment counts, and outcome measures. We define these quantities and their associated results. Index i identifies a trial, k a study arm, and j an outcome measure.

Definition 1 (**Study Arm**). A study arm is a group of participants assigned to a specific intervention regimen within a clinical trial. Its description can include the interventional drug, dose, and schedule information. A treatment arm receives the target intervention, whereas a comparator arm receives a placebo, standard care, or another intervention. We denote the k-th arm of trial C_i by $A_k^{(i)}$.
Example: A trial may include one arm receiving 200 mg of pembrolizumab every three weeks, and a comparator arm receiving standard chemotherapy. Pembrolizumab's biological mechanism of action is PD-1 blockade which restores T-cell anti-tumor activity via immune cell checkpoint inhibition.

Definition 2 (**Enrollment Count**). $N_k^{(i)}$ is the number of participants allocated to study arm $A_k^{(i)}$.

Definition 3 (**Outcome Measure**). An outcome measure (also called an endpoint) is a pre-specified quantity used to evaluate the effect of a drug intervention. It defines both what will be measured (e.g., 5-year overall survival rate, progression-free survival) and the criterion under which success is declared (e.g., a statistically significant improvement over the comparator at $p < 0.05$). We denote the j-th outcome measure of trial C_i by $O_j^{(i)}$.
Example: "5-year overall survival rate, with success defined as statistically significant improvement over the comparator at $p < 0.05$" is an outcome measure.

Definition 4 (**Trial Result**). A trial result is the observed value of an outcome measure for a particular study arm after the trial has been conducted. For trial C_i, outcome $O_j^{(i)}$, and study arm $A_k^{(i)}$ with enrollment $N_k^{(i)}$, we denote the corresponding result by

$$R_{j,k}^{(i)} = \text{Result of}\left(N_k^{(i)}, A_k^{(i)}, O_j^{(i)}, C_i\right).$$

This representation separates the three explicitly perturbed factors from auxiliary trial context. Trial results are mostly numerical.
Example: "85% of patients in arm $A_k^{(i)}$ survived for at least 5 years" is a trial result.

We refer to a set of these factors $X^{(i,j,k)} = \left(N_k^{(i)}, A_k^{(i)}, O_j^{(i)}\right)$ as a **configuration**. Enrollment is integer-valued whereas arms and outcome measures are textual descriptions. Along a pathway, (t) indexes a state rather than a trial. A historical anchor provides a configuration and its observed result, denoted $(X^{(1)}, R^{(1)})$. Its auxiliary trial context C_i, including eligibility criteria, is provided separately and held fixed within the pathway. It may differ across anchors.

A CT Open question (C, O, A_{k_1}, A_{k_2}) asks whether the first study arm will show a statistically significant improvement over its comparator arm at outcome measure O (Wang et al., 2026). The target description and historical observations are available at inference time and the target results are withheld. DeepImagine predicts arm-level results before forecasting this binary comparison. Macro-F1 evaluates the final answers, not intermediate results.

4 The DeepImagine Framework

DeepImagine asks a local question before making a global prediction: given a historical result, what would change if only one experimental factor were different? Figure 1 illustrates why the intermediate result must be carried forward. After enrollment count is updated, the study arm operator reasons about the new enrollment rather than starting from the original trial. Then, the outcome measure operator reasons about the updated regimen. We first formalize these updates, then explain the single-path approximation and multi-anchor aggregation, and finally show how the same updates become supervised training tasks.

Definition 5 (Local Counterfactual Imagination). Let $\Phi_{t-1} \in \{N, A, O\}$ be the only factor that differs between two clinical trial configurations, $X^{(t-1)}$ and $X^{(t)}$, taking values ϕ_{t-1} and ϕ_t, respectively, within this transition. Given the observed or imagined result $R^{(t-1)}$, a local counterfactual imagination is the mapping

$$f_{\Phi_{t-1}} : \left(R^{(t-1)}, X^{(t-1)}, X^{(t)}\right) \mapsto R^{(t)},$$

which produces an imagined result $R^{(t)}$ by setting the factor Φ_{t-1} to be ϕ_t instead of ϕ_{t-1}.

Definition 6 (Probabilistic Parametrization). For factor Φ_{t-1}, let $P_\theta^{\Phi_{t-1}}$ be a conditional language model with parameters θ. We use the argmax of a probabilistic language model $P_\theta^{\Phi_{t-1}}$ to calculate the local counterfactual imagination operator $f_{\Phi_{t-1}}$:

$$f_{\Phi_{t-1}}\left(R^{(t-1)}, X^{(t-1)}, X^{(t)}\right) = \operatorname*{argmax}_{R^{(t)}} P_\theta^{\Phi_{t-1}}\left(R^{(t)} \,\middle|\, R^{(t-1)}, X^{(t-1)}, X^{(t)}\right),$$

where θ denotes learnable parameters. The three-factor framework uses one specialized operator for each factor in $\{N, A, O\}$.

This result-level mode is an idealization: token-wise decoding and result extraction need not recover it. The analysis below requires its stated concentration condition at the retained result Appendix A.1.

Definition 7 (Stepwise Local Counterfactual Imaginations). A pathway of stepwise local counterfactual imaginations from source configuration $X^{(1)}$ with observed result $R^{(1)}$ to a target configuration $X^{(T)}$ is a chain

$$\left(X^{(1)}, R^{(1)}\right) \xrightarrow{f_{\Phi_1}} \left(X^{(2)}, R^{(2)}\right) \xrightarrow{f_{\Phi_2}} \cdots \xrightarrow{f_{\Phi_{T-1}}} \left(X^{(T)}, R^{(T)}\right).$$

where, for each $t = 2, \ldots, T$, the configurations $X^{(t-1)}$ and $X^{(t)}$ differ only in factor Φ_{t-1}, and $R^{(t)} = f_{\Phi_{t-1}}\left(R^{(t-1)}, X^{(t-1)}, X^{(t)}\right)$ is the result produced by the corresponding local counterfactual imagination.

We use $T = 4$ and $(\Phi_1, \Phi_2, \Phi_3) = (N, A, O)$, so three updates connect four states. The enrollment update changes only enrollment, keeping the historical study arm and outcome. The arm update then takes the enrollment-updated result as input, and the outcome update takes the arm-updated result. Because each update conditions on the preceding predicted result, later predictions can depend on earlier changes. The intermediate configurations are imagined experiments, rather than additional observed trials. We denote predicted results by $\hat{R}^{(t)}$, with $\hat{R}^{(1)} = R^{(1)}$. The final prediction is $\hat{R}^{(4)}$.

Note that only one factor changes at each step, but its predicted effect depends on the current configuration and result. The composition therefore does not assume additive factor effects nor invariance to the order in which factors are changed. For example, the effect predicted for an arm change can depend on the enrollment already updated and on the outcome specified at that step. Section 5.4 examines order dependence through an ablation study.

4.1 Approximating the Marginal with a Single Greedy Pathway

Our key empirical observation underlying DeepImagine is that directly predicting the target result $R^{(T)}$ from $(R^{(1)}, X^{(1)}, X^{(T)})$ in one step is hard, whereas single-factor counterfactual transitions are substantially easier to learn. However, by law of total probability, the marginal for $R^{(T)}$ via stepwise local counterfactual imaginations is intractable as it requires integrating over all possible values for intermediate results $R^{(2)}$ to $R^{(T-1)}$ Equation 1 in Appendix A.2 DeepImagine avoids this integration by single-factor perturbation at each step.

Suppose that, along this single-factor perturbation pathway, each local conditional places mass at least $1 - \delta_t$ on $\hat{R}^{(t)}$ Definition 8. The following approximation result characterizes when this single trajectory approximates the composed model.

Proposition 1 (Greedy Approximation of the Marginal). *Suppose each $P_\theta^{\Phi_{t-1}}$ is δ_t-dominant for $t = 2, \ldots, T$. The absolute difference between the greedy pathway probability (see Equation 2) and the model marginal at $\hat{R}^{(T)}$ is at most $\sum_{t=2}^{T} \delta_t$. For $T = 4$, the bound is $\delta_2 + \delta_3 + \delta_4$.*

The full inequality and proof of Proposition 1 are in Appendix A.3. The argument uses the chain rule and the elementary inequality $\prod_t (1 - \delta_t) \geq 1 - \sum_t \delta_t$. Empirically, this approximation works well and since we keep T relatively small, the chance of error propagation is reduced and the error bound is also tightened.

4.2 Convergence of Aggregating Multiple Counterfactual Pathways

Another key empirical finding is that aggregating pathways from different historical anchors improves on single-anchor prediction. We formalize the role of averaging under independent sampling and calibration assumptions.

Different anchors provide different starting observations for the same target $X^{(T)}$. Write $\bar{P}_n$ for the average of n pathway marginals and M_θ for its expectation under the data-generating law p_*. The analysis assumes i.i.d. complete historical records, including their fixed contexts. Calibration requires $M_\theta(\cdot|X^{(T)}) = P_{p_*}(\cdot|X^{(T)})$, the true target conditional (Definition 9). Appendix A.4 gives the complete setup.

Proposition 2 (Convergence of Aggregating Counterfactual Pathways). *Under i.i.d. anchor sampling and calibration, for every fixed target configuration $X^{(T)}$ and fixed measurable event $B \in \mathcal{B}$,*

$$\frac{1}{n} \sum_{i=1}^{n} P\Big(B \,\Big|\, R_i^{(1)}, X_i^{(1)}, X^{(T)}\Big) \;\longrightarrow\; P_{p_\star}\Big(B \,\Big|\, X^{(T)}\Big) \qquad \textit{almost surely as } n \to \infty.$$

The strong law makes the average approach M_θ. Calibration identifies that limit with the target law. With ϵ-approximate calibration and the common-density conditions in Appendix A.5, the aggregate's asymptotic total-variation distance from the target law is at most ϵ. Averaging reduces sampling variance under these assumptions, but does not remove calibration error.

The calibration assumption is also plausible by construction. Our training procedure for the local operators P_θ^Φ (Section 4.3) explicitly targets it: the natural counterfactuals used during training are observed conditional realizations of the true law $p_\star$, and the training objective minimizes a divergence between P_θ^Φ and the empirical conditional. Calibration is thus not an external condition imposed on the analysis but the operational target of training itself.

4.3 Training with Natural and Synthetic Counterfactuals

The local conditional is both an inference operation and a supervised prediction task. We train one operator P_θ^Φ for each factor $\Phi \in \{N, A, O\}$ on tuples $(R^{(t-1)}, X^{(t-1)}, R^{(t)}, X^{(t)})$. The preceding result and both configurations are inputs; the updated result is the target. We collect such tuples from two complementary sources: **natural counterfactuals**, whose paired results are both observed, and **synthetic counterfactuals**, which include synthetically generated intermediate results.

Natural Counterfactuals from a Controlled Clinical Trial: A controlled clinical trial typically reports several outcome measures across multiple study arms. For one arm k in trial i, $R^{(i)}_{j_1,k}$ and $R^{(i)}_{j_2,k}$ are observed under two outcome measures $O^{(i)}_{j_1}$ and $O^{(i)}_{j_2}$. The corresponding configurations differ in the outcome field. This gives a natural pair for learning how the result changes when the measured quantity changes, while keeping the intervention unchanged. We prompt GPT-5 to explain the transition between the two observed results and train P_θ^O to generate this explanation followed by $R^{(i)}_{j_2,k}$. GPT-5 supplies the explanation, not the result label. Appendix B.1 retains the full conditional training expression.

Analogously, for outcome measure O_j and study arms k_1, k_2, the observed results $R^{(i)}_{j,k_1}, R^{(i)}_{j,k_2}$ provide input-target pairs for P_θ^A, where the same trial can provide several local tasks rather than one trial-success label. These are aggregate comparisons, not both potential outcomes for individual patients. Exact single-factor matching requires the remaining attributes, including enrollment and analysis population, to agree, and randomization alone does not ensure this (Appendix C.4).

Synthetic Counterfactuals from Similar Trial Pairs. Natural local counterfactuals are unavailable for the enrollment factor N, because constructing one would require running an alternate version of the same trial *ceteris paribus*, varying only the factor of interest. To address this gap, we develop a synthetic counterfactual pipeline. We first embed each historical trial with NV-Embed-v2 (Lee et al., 2025), retrieve candidate pairs based on cosine similarity, and use GPT-5 to verify the semantic similarity of each retrieved pair. In practice, however, even the most similar trials still differ along all three endogenous factors $\{N, A, O\}$, so they cannot be used directly as single-factor local counterfactuals.

Concretely, for historical trials C_1, C_2 with configurations $(N^{(1)}, A^{(1)}, O^{(1)})$ and $(N^{(2)}, A^{(2)}, O^{(2)})$ with observed results $R^{(1)}$ and $R^{(2)}$, we prompt GPT-5 to imagine a pathway that changes enrollment, study arm, and outcome measure in that order. Both observed results are provided to guide the generated transitions. The source context is held fixed; the second trial's result is reference evidence, not an observed label for an imagined configuration with potentially different context. Each step and its accompanying explanation yield a training tuple. These tuples provide the enrollment operator's supervision and augment the natural sets for P_θ^A and P_θ^O. Appendix B.1 gives the complete three-step bridge and distinguishes observed from generated results.

5 Results and Analyses

5.1 Experimental Setup

Training data. Our data source is 14,000 randomized controlled drug trials with posted results on clinicaltrials.gov, restricted to trials with enrollment count $N \geq 50$ to filter out underpowered studies. Following Section 4.3, we extracted natural counterfactuals: 57k for training P_θ^A and 82k for training P_θ^O. We also extracted synthetic counterfactuals: 45k per factor $\Phi \in \{N, A, O\}$. For P_θ^A and P_θ^O, these augment the natural sets; for P_θ^N, no natural counterfactual is available, so only synthetic data is used. GPT-5 generates the accompanying explanations. We hold out 10% of each operator's training set for hyperparameter tuning. Appendix B.2 reports the full settings.

To assess the quality of our training data, a biomedical expert reviewed 90 generated explanations: 87 were clinically sound, two were potentially problematic, and one was uncertain. Most counterfactual results are therefore derived through sound biomedical reasoning, although a lower percentage of flawed explanations add some noise to training.

Table 1: Prediction strategies and trained predictors on the 701-question CT Open cohort (macro-F1, 0–100). Values after ± are reported 95% bootstrap CI half-widths. Bold marks the largest GPT-row or baseline-family estimate (Appendix C.2). Relative to RAG within each backbone, ↑ and ↓ indicate nominal increases and decreases, respectively, while ↑↑ denotes statistically significant improvements.

A. LLMs	One-step prediction		DeepImagine		
Backbone	RAG	Direct	1 anchor 1 pathway	1 anchor 7 pathways	7 anchors 1 pathway each
GPT-5	72.4±3.32	69.4±3.41 ↓	77.1±3.14 ↑	75.8±2.99 ↑	**79.0**±3.09 ↑↑
GPT-5-mini	71.7±3.38	72.8±3.34 ↑	73.3±3.31 ↑	76.0±3.16 ↑	**78.5**±3.06 ↑↑
o3-mini	71.0±3.31	73.1±3.38 ↑	72.9±3.27 ↑	72.6±3.43 ↑	75.6±3.34 ↑

B. Conventional classifiers	macro-F1
PLR (ℓ_2)	**70.8**
Feed-forward NN	66.4
XGBoost	66.2
Random forest	65.9
kNN imputation + RF	60.3
AdaBoost	59.8

C. Clinical predictors	macro-F1
MEXA-CTP	**69.8**
HINT	59.5
SPOT	59.3
LIFTED	58.6
Trained local operators	
DeepImagine (Gemma)	**78.7**

Evaluation benchmarks. We evaluate the time-stamped CT Open benchmark (Wang et al., 2026): *Summer 2025* (cutoff 2025-08-31) with 701 questions including 442 positives and 259 negatives. Each question is a binary decision over a tuple (C, O, A_{k_1}, A_{k_2}) asking whether R_{k_1} shows a statistically significant improvement over R_{k_2}. DeepImagine produces separate estimates $\hat{R}_{k_1}$ and $\hat{R}_{k_2}$ via stepwise local counterfactual imagination, and for the off-the-shelf LLM backbones, the same LLM converts these predictions into a binary answer. We report macro-F1, the unweighted mean of the two class-specific F1 scores, on a 0–100 scale.

Models. We evaluate three off-the-shelf LLMs as DeepImagine backbones: GPT-5-2025-08-07, GPT-5-mini-2025-08-07, o3-mini-2025-01-31. Their stated knowledge cutoffs precede the benchmark cutoff. We instantiate each P_θ^Φ, for $\Phi \in \{N, A, O\}$, as a separate Gemma-3-12B-it model (Gemma Team & Google DeepMind, 2025), fine-tuned on its corresponding training set. We also evaluate six conventional classifiers (Lo et al., 2019; Tranchevent et al., 2019; Fan et al., 2020; Siah et al., 2021) and trial prediction methods including HINT (Fu et al., 2022), SPOT (Wang et al., 2023c), LIFTED (Zheng et al., 2025), and MEXA-CTP (Zhang et al., 2025). The conventional models combine numerical, categorical, and NV-Embed-v2 text features. Appendix C.2 gives full details about data processing, method adaptation, and hyperparameters.

For off-the-shelf LLMs, we compare four strategies:

Direct: The LLM sees only the target tuple (C, O, A_{k_1}, A_{k_2}) before giving a binary answer.
RAG. For each arm A_k, we retrieve up to 7 historical trials (with observed results) that are most relevant to that arm. The LLM sees these retrieved results and the target configuration before giving an answer. This is the standard RAG baseline of Wang et al. (2026).
DeepImagine (n-anchor). For each arm, $1 \leq n \leq 7$ distinct retrieved trials initialize separate three-update pathways. RAG and DeepImagine use the same retrieved evidence. The n predictions are averaged, following Proposition 2.
DeepImagine (same-anchor, n pathways). An ablation in which all n pathways start from the same historical trial. This violates the i.i.d. anchor assumption of Proposition 2 and controls for whether multi-anchor improvement can be attributed to majority voting alone.

5.2 Ablation Study Setup

To isolate the contributions that anchoring on historical trials and stepwise factorization used in DeepImagine for training language models, we train two ablation variants of Gemma, both with and without GPT-5-generated reasoning traces.

(A1) Direct Configuration-to-Result Prediction.
Gemma is trained to map a configuration $X = (N, A, O)$ and its available trial context C directly to its result R. This setting tests whether a language model can directly capture the latent structural mechanism governing clinical trial outcomes. We use 127k configuration-result pairs to ensure data scale is not the bottleneck.

(A2) One-Step Prediction from a Historical Trial Result
Gemma is trained to predict $R^{(T)}$ from a historical configuration and result $\left(X^{(1)}, R^{(1)}\right)$ together with the target configuration. For each of the 127k targets, NV-Embed-v2 retrieval followed by GPT-5 verification identifies a similar predecessor, yielding 127k source-target pairs. A2 learns a direct predictor of the target result. It does not compute the marginal of the trained local operators in Equation 1.

5.3 Analyses and Discussions

Seven-anchor DeepImagine improves all GPT backbones. Table 1 shows that even with a single anchor ($n = 1$), DeepImagine improves over RAG by 4.7 points for GPT-5, 1.6 for GPT-5-mini, and 1.9 for o3-mini. The single-anchor result does not, however, exceed Direct for o3-mini (72.9 vs. 73.1). At seven distinct anchors, DeepImagine exceeds both baselines for all three models tested. RAG itself is below Direct for GPT-5-mini and o3-mini, indicating that retireved historical trials are not always helpful when provided as raw context. DeepImagine, by contrast, consumes the same retrieved trials but processes them through stepwise local counterfactual imagination rather than direct correlation, and the resulting gains are stable across LLMs of different scales and providers.

DeepImagine improves on a broader set of prediction baselines. Table 1 B-C reports all ten conventional and task-adapted clinical trial prediction methods. PLR with ℓ_2 regularization achieves the strongest result at 70.8 macro-F1, followed by MEXA-CTP at 69.8. The trained Gemma system reaches 78.7, significantly outperforming all traditional baseline methods. These comparisons extend beyond a single feed-forward reference model. They evaluate adaptations to our arm–endpoint task, not scores copied from the methods' original trial-level benchmarks; input representations and training procedures differ across methods (Appendix C.2).

Aggregating across multiple distinct anchors yields substantial additional gains. Moving from n=1 to n=7 different anchors improves DeepImagine by 1.9 for GPT-5 (to 79.0) and 5.2 for GPT-5-mini (to 78.5). Figure 2 shows the reported anchor-subset comparison. Though this comparison increases both the amount of historical evidence and the number of pathways, the same-anchor control below compares distinct evidence with repeated generation.

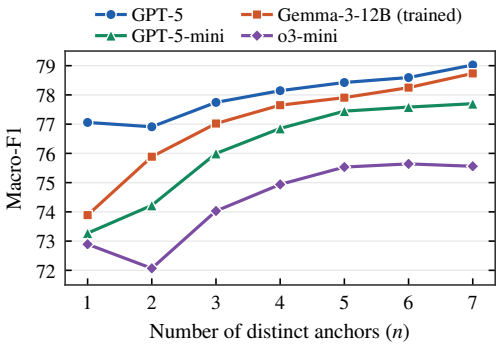

Figure 2: Macro-F1 on the CT Open Summer 2025 as the number of distinct anchors n increases. For $n \geq 2$, each point is the mean over all $\binom{7}{n}$ anchor subsets.

Local-operator training outperforms the one-step Gemma variants. Table 2 compares the two one-step training schemes with local-operator training. A1 reaches 54.8/56.9 macro-F1 with/without explanations; adding a historical result in A2 raises these scores to 60.2/63.2. DeepImagine reaches 78.7: 15.5 points above the best A2 setting and 21.8 above the best A1 setting. Explanations alone do not account for the improvement, since both one-step variants score lower with them. The result favors the combined local-supervision and multi-anchor strategy, while the unequal data and inference budgets prevent attributing the entire gap to factorization.

Table 2: Gemma training comparisons (macro-F1). "With/without" denotes generated explanations. A dash is an unreported setting.

Task	Anchors	With	Without
A1: direct	0	54.8	56.9
A2: one-step	1	60.2	63.2
DeepImagine	7	**78.7**	—

Distinct anchors offer more than repeated prediction from one anchor. At the same count of seven pathways, distinct anchors outperform repeated prediction from one trial by 3.2, 2.5, and 3.0 points for GPT-5, GPT-5-mini, and o3-mini, respectively. Repeating one anchor changes the single-pathway scores by -1.3, 2.7, and -0.3 points. Thus, the reported multi-anchor gains is not demonstrated by additional generation from unchanged evidence. The contrast supports the value of different observed starting results rather than repeated use of the same evidence. Its relationship to Proposition 2 is discussed in Section 4.2.

Trained Gemma operators match GPT-5 within DeepImagine. The Gemma family of local operators within DeepImagine (78.7) matches GPT-5 within the same framework (79.0) at $n = 7$, despite a much smaller parameter count. This indicates that explicitly teaching a language model to perform local counterfactual imagination from natural and synthetic counterfactuals is comparable in effectiveness to relying on the innate reasoning of a much larger generalist LLM, while being substantially more efficient at deployment.

5.4 Perturbation Order, Anchor Relevance, and Explanation Quality

The preceding results motivate two further questions: does the order of local changes matter, and how much does prediction depend on the starting observation? The following ablation studies vary these choices separately to investigate the effects.

The order of local updates matters. We use the ordering $N \to A \to O$ because it reduces error propagation and reflects dependencies among the factors. We update N first because enrollment changes primarily affect statistical power while leaving percentage-based results largely unchanged, making this counterfactual the most reliable. We then update A before O, since the interpretation of an outcome measure can depend on its associated study arm. This ordering was chosen in consultation with our biomedical expert and used throughout data generation and experiments. We further evaluate all six permutations on the 701 CT Open questions using GPT-5 with DeepImagine; because GPT-5 is not trained on our generated data, this analysis is independent of the generation order. As shown in Table 3, the canonical $N \to A \to O$ order achieves the highest macro-F1 of 77.06.

Table 3: GPT-5 order sensitivity (macro-F1). Anchors, instructions, and judge are fixed. Δ is relative to $N \to A \to O$.

Update order	Macro-F1	Δ
$N \to A \to O$	**77.06**	—
$A \to O \to N$	73.71	-3.35
$A \to N \to O$	73.24	-3.82
$O \to A \to N$	73.03	-4.03
$N \to O \to A$	72.99	-4.07
$O \to N \to A$	72.45	-4.61

Relevant anchors help, with different sensitivity across models. In another ablation study, we investigate the effect of anchor relevancy to the DeepImagine framework. A relevant anchor is selected by GPT-5 from the top 10 NV-Embed-v2 candidates. Alternatively, we select a random training-corpus anchor. This changes the relevance of the starting configuration to the target configuration. As shown in Table 4, Random selection lowers GPT-5 from 75.90 to 71.20 and Gemma from 73.75 to 73.09. The smaller drop in Gemma model suggests its lower sensitivity to anchor selection, which could be the result of varied counterfactual settings it has been trained on.

Table 4: Anchor relevance, a separate study. Macro-F1; Δ is random minus relevant.

Model	Relevant	Random	Δ
GPT-5	75.90	71.20	-4.70
Gemma	73.75	73.09	-0.66

6 Conclusion and Limitation

We introduce DeepImagine, a step-wise local counterfactual imagination framework for predicting clinical trial outcomes. On the Summer 2025 split of CT Open, it outperforms traditional machine learning baselines. Applying DeepImagine enables off-the-shelf LLMs to predict trial outcomes more accurately over baselines. Training with DeepImagine also enables the much smaller Gemma-3-12B-it nearly match GPT-5. Limitations include limited coverage of trial factors and reliance on GPT-5 to synthesize intermediate steps when natural counterfactuals are unavailable. Improving the fidelity of these synthetic intermediates, e.g., via verification against future trial readouts, is a natural next step.

AI USE STATEMENT

GPT-5 is used for generated explanations, synthetic intermediate results, retrieval relevance evaluation, and prediction as described in Sections 4.3 and 5. Generative AI also assisted manuscript editing and figure/table layout.

ETHICS STATEMENT

The study uses reported aggregate clinical trial information published in clinicaltrials.gov. Its predictions and imagined intermediate results are research outputs – they by no means serve as clinical recommendations or identified patient-level effects.

REPRODUCIBILITY STATEMENT

Section 5.1 describes the evaluation setup and predictive strategies. Appendices A, B.1, and B.2 provide proofs, supervision details, and training settings. All code are available in the provided GitHub link, which is anonymous for review purpose.

REFERENCES

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruixiang Tang. Counterbench: Evaluating and improving counterfactual reasoning in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30350–30358, 2026. doi: 10.1609/aaai.v40i36.40287. URL https://doi.org/10.1609/aaai.v40i36.40287.

Zhao Fan, Fanyu Xu, Cai Li, and Lili Yao. Application of kpca and adaboost algorithm in classification of functional magnetic resonance imaging of alzheimer's disease. *Neural Computing and Applications*, 32:5329–5338, 2020. doi: 10.1007/s00521-020-04707-y.

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Hint: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4):100445, 2022. doi: 10.1016/j.patter.2022.100445.

Gemma Team and Google DeepMind. Gemma 3 technical report. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf.

David Heckerman, Christopher Meek, and Gregory F. Cooper. A bayesian approach to causal discovery. In Gregory F. Cooper and Clark Glymour (eds.), *Computation, Causation, and Discovery*, pp. 141–165. AAAI Press / MIT Press, Cambridge, MA, 1999.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, 2023.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks, 2023. URL https://arxiv.org/abs/2210.02406.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=mqoxLkX210. Featured Certification.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2025. URL https://arxiv.org/abs/2405.17428.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Andrew W. Lo, Kien Wei Siah, and Chi Heem Wong. Machine learning with statistical imputation for predicting drug approvals. *Harvard Data Science Review*, 1(1), 2019. doi: 10.1162/99608f92.5c5f0525. URL https://hdsr.mitpress.mit.edu/pub/ct67j043.

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia C. Passaro, Irene Sucameli, and Alessandro Lenci. Explica: Evaluating explicit causal reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

OpenAI. Gpt-5, 2025. URL https://openai.com/index/introducing-gpt-5/.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Michael Reinisch, Jianfeng He, Chenxi Liao, Sauleh Ahmad Siddiqui, and Bei Xiao. Ctp-llm: Clinical trial phase transition prediction using large language models, 2024. URL https://arxiv.org/abs/2408.10995.

Kien Wei Siah, Nicholas W. Kelley, Steffen Ballerstedt, Björn Holzhauer, Tianmeng Lyu, David Mettler, Sophie Sun, Simon Wandel, Yang Zhong, Bin Zhou, Shifeng Pan, Yingyao Zhou, and Andrew W. Lo. Predicting drug approvals: The novartis data science and artificial intelligence challenge. *Patterns*, 2(8):100312, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100312. URL https://www.sciencedirect.com/science/article/pii/S2666389921001550.

Léon-Charles Tranchevent, Francisco Azuaje, and Jagath C. Rajapakse. A deep neural network approach to predicting clinical outcomes of neuroblastoma patients. *BMC Medical Genomics*, 12 (Suppl 8):178, 2019. doi: 10.1186/s12920-019-0628-y. URL https://doi.org/10.1186/s12920-019-0628-y.

Aniket Vashishtha, Qirun Dai, Hongyuan Mei, Amit Sharma, Chenhao Tan, and Hao Peng. Executable counterfactuals: Improving llms' causal reasoning through code, 2025. URL https://arxiv.org/abs/2510.01539.

Jianyou Wang, Youze Zheng, Longtian Bao, Hanyuan Zhang, Qirui Zheng, Yuhan Chen, Yang Zhang, Matthew Feng, Maxim Khan, Aditya K. Sehgal, Christopher D. Rosin, Ramamohan Paturi, Umber Dube, and Leon Bergen. Ct open: An open-access, uncontaminated, live platform for the open challenge of clinical trial outcome prediction. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2604.16742.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023a. URL https://arxiv.org/abs/2203.11171.

Zifeng Wang and Jimeng Sun. Trial2Vec: Zero-shot clinical trial document similarity search using self-supervision. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6377–6390, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.476. URL https://aclanthology.org/2022.findings-emnlp.476/.

Zifeng Wang, Brandon Theodorou, Tianfan Fu, Cao Xiao, and Jimeng Sun. Pytrial: Machine learning software and benchmark for clinical trial applications. *arXiv preprint arXiv:2306.04018*, 2023b.

Zifeng Wang, Cao Xiao, and Jimeng Sun. Spot: sequential predictive modeling of clinical trial outcome with meta-learning. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–11, 2023c.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. Large language models as analogical reasoners, 2024. URL https://arxiv.org/abs/2310.01714.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning, 2024. URL https://arxiv.org/abs/2404.14777.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. URL https://arxiv.org/abs/2203.14465.

Yiqing Zhang, Xiaozhong Liu, and Fabricio Murai. Mexa-ctp: Mode experts cross-attention for clinical trial outcome prediction. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*, pp. 467–475. SIAM, 2025. doi: 10.1137/1.9781611978520.51.

Wenhao Zheng, Liaoyaqi Wang, Dongshen Peng, Hongxia Xu, Yun Li, Hongtu Zhu, Tianfan Fu, and Huaxiu Yao. Lifted: Multimodal clinical trial outcome prediction via large language models and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7503–7517. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.396. URL https://aclanthology.org/2025.findings-emnlp.396/.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. URL https://arxiv.org/abs/2205.10625.

Appendix: Table of Contents

A. Formal Definitions and Complete Theoretical Analysis 15
1. Notation, Conditioning, and Result Distributions 15
2. Complete Marginal and Greedy-Pathway Formulation 15
3. Proof of Proposition: Greedy Approximation of the Marginal 16
4. Proof of Proposition: Convergence of Aggregating Counterfactual Pathways 18
5. Relaxed Convergence under Approximate Calibration 19
B. Supervision and Training Details 20
1. Observed Labels and Synthetic Intermediate Results 20
2. Training Details 20
C. Evaluation Protocols and Additional Analyses 21
1. Retrieval, Prediction, and Aggregation 21
2. Conventional and Task-Adapted Clinical Predictors 22
3. Sensitivity Studies and Statistical Reporting 23
4. Clinical Interpretation and Explanation Assessment 23

A Formal Definitions and Complete Theoretical Analysis

This appendix develops the full mathematical formulation underlying Section 4. Section A.2 gives the marginal and greedy pathway. Appendices A.3-A.5 give proofs and approximate-calibration extension. We use T states and $T-1$ updates throughout.

A.1 Notation, Conditioning, and Result Distributions

Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{R}, \mathcal{B})$ be the measurable configuration and result spaces. Superscripts index trials in $R^{(i)}_{j,k}$ and pathway states in $R^{(t)}$; the surrounding definition specifies which is intended. Singletons, contexts, maps, and probability kernels are assumed measurable. For an anchor C_i, the full local conditional is $P_\theta^{\Phi_{t-1}}(\cdot \mid R^{(t-1)}, X^{(t-1)}, X^{(t)}; C_i)$. The historical context is held fixed along the pathway and suppressed in the equations to preserve the original notation. Across anchors, expectations are over complete records $(X^{(1)}, R^{(1)}, C)$, not merely the three configuration fields and result. The target and factor order are fixed throughout each application of the propositions.

In the training expression for $\mathrm{concat}(\text{reasoning}, R)$, P_θ^Φ denotes a complete-response law. In the result-level analysis, it denotes the induced law after a measurable result parser is applied. Different responses can yield the same result. This does not equate the probability of a result with a confidence score stated in an explanation. A failure symbol may make the parsing map total; that mathematical convention does not specify an empirical invalid-output policy.

The modal definition in Definition 6 assumes that a result-level mode and a fixed measurable tie-breaking rule exist. Token-wise greedy decoding need not recover that mode. The greedy proof requires the stated mass bound at the retained result; it does not require exact modal decoding if that bound holds. Probabilities evaluated at a result denote singleton masses, not density heights. Finite result strings or their parsed summaries give a countable result space. More general spaces are allowed, but purely non-atomic conditionals cannot satisfy near-unit singleton dominance.

A.2 Complete Marginal and Greedy-Pathway Formulation

The law of total probability gives the following marginal over a fixed pathway. We retain the original integral notation. For a countable result representation, each $dR^{(t)}$ denotes counting measure and each P_θ^Φ in the integrand is a mass function. Under a density interpretation, integration uses the corresponding reference measure; the proofs below use probability kernels explicitly.

$$P\big(R^{(T)} \,\big|\, R^{(1)}, X^{(1)}, \ldots, X^{(T)}\big) = \int_{R^{(2)}} \cdots \int_{R^{(T-1)}} \prod_{t=2}^{T-1} P_\theta^{\Phi_{t-1}}\big(R^{(t)} \,\big|\, R^{(t-1)}, X^{(t-1)}, X^{(t)}\big) \times P_\theta^{\Phi_{T-1}}\big(R^{(T)} \,\big|\, R^{(T-1)}, X^{(T-1)}, X^{(T)}\big)\, dR^{(2)} \cdots dR^{(T-1)} \tag{1}$$

Our training procedure that uses natural counterfactuals from controlled experiments and synthetic counterfactuals from similar trials (Section 4.3) encourages each local operator $P_\theta^{\Phi_{t-1}}$ to place most of its probability mass on a single value at each step. We formalize this property below as the dominant local conditional assumption, under which the intractable model marginal can be tightly approximated by the probability product along a single greedy pathway which is precisely the quantity DeepImagine computes at inference.

Concretely, the probability of the greedy pathway starting from observed result $R^{(1)}$ is

$$\begin{aligned} &\prod_{t=2}^{T} P_\theta^{\Phi_{t-1}}\left(\hat{R}^{(t)} \mid \hat{R}^{(t-1)}, X^{(t-1)}, X^{(t)}\right) \\ &= P_\theta^{\Phi_1}\left(\hat{R}^{(2)} \mid R^{(1)}, X^{(1)}, X^{(2)}\right) \\ &\times \prod_{t=3}^{T} P_\theta^{\Phi_{t-1}}\left(\hat{R}^{(t)} \mid f_{\Phi_{t-2}}\left(\hat{R}^{(t-2)}, X^{(t-2)}, X^{(t-1)}\right), X^{(t-1)}, X^{(t)}\right). \end{aligned} \tag{2}$$

where $\hat{R}^{(t-1)}$ takes the greedy value using the operator $f_{\Phi_{t-2}}(\cdot)$ for $t \geq 3$ and $P_\theta^{\Phi_{t-1}}$ is a language model that captures a local counterfactual imagination.

Definition 8 (Dominant Local Conditional). A local operator $P_\theta^{\Phi_{t-1}}$ is δ_t-dominant along the greedy trajectory if its mode $\hat{R}^{(t)} := f_{\Phi_{t-1}}(\hat{R}^{(t-1)}, X^{(t-1)}, X^{(t)})$ carries near-unit mass:

$$P_\theta^{\Phi_{t-1}}\big(\hat{R}^{(t)} \,\big|\, \hat{R}^{(t-1)},\, X^{(t-1)},\, X^{(t)}\big) \;\geq\; 1 - \delta_t,$$

for some $0 < \delta_t \ll 1$.

Let $\big\{\big(X_i^{(1)}, R_i^{(1)}\big)\big\}_{i\geq 1}$ be an i.i.d. sequence of historical trial anchors drawn from the unknown true data-generating law $p_\star$, where $\big(X_i^{(1)}, R_i^{(1)}\big)$ is the source configuration and its observed result for the i-th clinical trial anchor. For any target configuration $X^{(T)}$, running DeepImagine from $\big(X_i^{(1)}, R_i^{(1)}\big)$ to $X^{(T)}$ yields the marginal distribution $P\Big(R^{(T)} \,\Big|\, R_i^{(1)}, X_i^{(1)}, X^{(T)}\Big)$ over the target result $R^{(T)}$, see Equation 1. Let $P_{p_\star}\big(R^{(T)} \,\big|\, X^{(T)}\big)$ denote the true conditional distribution of $R^{(T)}$ given $X^{(T)}$ under the law $p_\star$.

A.3 Proof of Proposition 1 (Greedy Approximation of the Marginal)

For convenience we restate the proposition. We write $P_\theta^{\Phi_{t-1}}(\hat{R}^{(t)} \mid \cdots) := P_\theta^{\Phi_{t-1}}(\{\hat{R}^{(t)}\} \mid \cdots)$ for the atom mass that the local operator places at the greedy mode $\hat{R}^{(t)}$. We write P_θ for the model marginal (denoted P in the main text). Marginalization over intermediate trajectories is performed via integration with respect to the local probability measures.

Proposition (restated) (Greedy Approximation of the Marginal). Under the δ_t-dominance assumption (Definition 8) for $t = 2, \ldots, T$,

$$\Big|\prod_{t=2}^{T} P_\theta^{\Phi_{t-1}}\big(\hat{R}^{(t)} \,\big|\, \hat{R}^{(t-1)},\, X^{(t-1)},\, X^{(t)}\big) \;-\; P_\theta\big(\hat{R}^{(T)} \,\big|\, R^{(1)},\, X^{(1)},\, \ldots,\, X^{(T)}\big)\Big| \;\leq\; \sum_{t=2}^{T} \delta_t.$$

The same assumptions also imply the total-variation consequence proved below:

$$\Big\| P_\theta(\cdot \mid R^{(1)}, X^{(1)}, \ldots, X^{(T)}) - \boldsymbol{\delta}_{\hat{R}^{(T)}} \Big\|_{\mathrm{TV}} \leq \sum_{t=2}^{T} \delta_t.$$

Proof. Throughout, all conditioning is on the fixed inputs $(R^{(1)}, X^{(1)}, \ldots, X^{(T)})$, which we abbreviate as $(\cdots)$.

By the probabilistic parametrization (Definition 6) and the chain-rule decomposition (Definition 7), the joint probability measure over the trajectory $(R^{(2)}, \ldots, R^{(T)}) \in \mathcal{R}^{T-1}$ factorizes as

$$P_\theta\big(dR^{(2)}, \ldots, dR^{(T)} \,\big|\, \cdots\big) \;=\; \prod_{t=2}^{T} P_\theta^{\Phi_{t-1}}\big(dR^{(t)} \,\big|\, R^{(t-1)}, X^{(t-1)}, X^{(t)}\big).$$

Restricting both sides to the singleton $\{(\hat{R}^{(2)}, \ldots, \hat{R}^{(T)})\} \subseteq \mathcal{R}^{T-1}$ gives the joint atom mass:

$$P_\theta\big(\hat{R}^{(2)}, \ldots, \hat{R}^{(T)} \,\big|\, \cdots\big) \;=\; \prod_{t=2}^{T} P_\theta^{\Phi_{t-1}}\big(\hat{R}^{(t)} \,\big|\, \hat{R}^{(t-1)}, X^{(t-1)}, X^{(t)}\big). \tag{3}$$

By δ_t-dominance, $P_\theta^{\Phi_{t-1}}(\hat{R}^{(t)} \mid \hat{R}^{(t-1)}, X^{(t-1)}, X^{(t)}) \geq 1 - \delta_t$ for every $t \in \{2, \ldots, T\}$. Multiplying these inequalities,

$$\prod_{t=2}^{T} P_\theta^{\Phi_{t-1}}\big(\hat{R}^{(t)} \,\big|\, \hat{R}^{(t-1)}, \cdots\big) \;\geq\; \prod_{t=2}^{T}(1 - \delta_t) \;\geq\; 1 - \sum_{t=2}^{T} \delta_t, \tag{4}$$

where the final inequality is the Weierstrass product inequality, proved by induction on T. The base case $T = 2$ gives $1 - \delta_2 \geq 1 - \delta_2$, which is trivial. For the inductive step, assuming $\prod_{t=2}^{T}(1-\delta_t) \geq 1 - \sum_{t=2}^{T}\delta_t$,

$$\prod_{t=2}^{T+1}(1-\delta_t) \;=\; (1-\delta_{T+1})\prod_{t=2}^{T}(1-\delta_t) \;\geq\; (1-\delta_{T+1})\Big(1-\sum_{t=2}^{T}\delta_t\Big).$$

Expanding the right-hand side,

$$(1-\delta_{T+1})\Big(1-\sum_{t=2}^{T}\delta_t\Big) \;=\; 1-\sum_{t=2}^{T+1}\delta_t \;+\; \delta_{T+1}\sum_{t=2}^{T}\delta_t \;\geq\; 1-\sum_{t=2}^{T+1}\delta_t,$$

since $\delta_{T+1}\sum_{t=2}^{T}\delta_t \geq 0$ for $\delta_t \in [0,1]$.

The marginal atom mass of $R^{(T)}$ at the greedy value $\hat{R}^{(T)}$ is obtained by integrating over the intermediate result space:

$$P_\theta\big(\hat{R}^{(T)} \,\big|\, \cdots\big) = \int_{\mathcal{R}^{T-2}} P_\theta^{\Phi_{T-1}}\big(\hat{R}^{(T)} \,\big|\, R^{(T-1)}, X^{(T-1)}, X^{(T)}\big) \times \prod_{t=2}^{T-1} P_\theta^{\Phi_{t-1}}\big(dR^{(t)} \,\big|\, R^{(t-1)}, X^{(t-1)}, X^{(t)}\big).$$

The greedy trajectory $(\hat{R}^{(2)}, \ldots, \hat{R}^{(T-1)}, \hat{R}^{(T)})$ contributes the joint atom mass $P_\theta(\hat{R}^{(2)}, \ldots, \hat{R}^{(T)} \mid \cdots)$ to the integral above. Since each local operator has an atom of mass $\geq 1 - \delta_t$ at $\hat{R}^{(t)}$, this contribution is non-negative, and the integral is at least this single contribution:

$$P_\theta\big(\hat{R}^{(T)} \,\big|\, \cdots\big) \;\geq\; P_\theta\big(\hat{R}^{(2)}, \ldots, \hat{R}^{(T)} \,\big|\, \cdots\big).$$

Combining with equation 3 and equation 4,

$$\begin{aligned} 1 \;\geq\; P_\theta\big(\hat{R}^{(T)} \,\big|\, \cdots\big) \;&\geq\; P_\theta\big(\hat{R}^{(2)}, \ldots, \hat{R}^{(T)} \,\big|\, \cdots\big) \\ &= \prod_{t=2}^{T} P_\theta^{\Phi_{t-1}}\big(\hat{R}^{(t)} \,\big|\, \hat{R}^{(t-1)}, \cdots\big) \;\geq\; 1-\sum_{t=2}^{T}\delta_t. \end{aligned} \tag{5}$$

Both the stepwise product and the marginal atom mass at $\hat{R}^{(T)}$ thus lie in the interval $\big[1-\sum_t \delta_t,\, 1\big]$, with the product no larger than the marginal. This yields the first claim:

$$0 \;\leq\; P_\theta\big(\hat{R}^{(T)} \,\big|\, \cdots\big) \;-\; \prod_{t=2}^{T} P_\theta^{\Phi_{t-1}}\big(\hat{R}^{(t)} \,\big|\, \hat{R}^{(t-1)}, \cdots\big) \;\leq\; 1-\Big(1-\sum_{t=2}^{T}\delta_t\Big) \;=\; \sum_{t=2}^{T}\delta_t.$$

For the total variation corollary, recall that for any probability measure P on $(\mathcal{R}, \mathcal{B})$ and any point $r^\star \in \mathcal{R}$,

$$\big\|P - \boldsymbol{\delta}_{r^\star}\big\|_{\mathrm{TV}} \;=\; \sup_{B\in\mathcal{B}} \big|P(B) - \boldsymbol{\delta}_{r^\star}(B)\big| \;=\; 1 - P(\{r^\star\}),$$

where the supremum is attained at $B = \{r^\star\}^c$ (or equivalently $B = \{r^\star\}$). Applying this with $r^\star = \hat{R}^{(T)}$ and equation 5,

$$\Big\|P_\theta\big(\cdot \,\big|\, R^{(1)}, X^{(1)}, \ldots, X^{(T)}\big) \;-\; \boldsymbol{\delta}_{\hat{R}^{(T)}}\Big\|_{\mathrm{TV}} \;=\; 1 - P_\theta\big(\hat{R}^{(T)} \,\big|\, \cdots\big) \;\leq\; \sum_{t=2}^{T}\delta_t.$$

This establishes the corollary and completes the proof. □

Remark (Linear-in-T scaling). The bound grows linearly in the pathway length T, not exponentially. The sum accumulates per-step concentration gaps; it is informative when those gaps are small, rather than a bound on clinical prediction error. For the three-factor setting $T = 4$, the bound reduces to $\delta_2 + \delta_3 + \delta_4$, recovering the bound stated in the main text statement of Proposition 1.

A.4 Proof of Proposition 2 (Convergence of Aggregating Counterfactual Pathways)

For convenience we restate the proposition. Throughout this appendix, the result space $(\mathcal{R}, \mathcal{B})$ follows Appendix A.1; it need not be non-atomic, and the configuration space $\mathcal{X}$ is equipped with a σ-algebra $\mathcal{A}$. All single-pathway predictions $P_\theta(\cdot \mid R^{(1)}, X^{(1)}, X^{(T)})$ are interpreted as probability measures on $(\mathcal{R}, \mathcal{B})$.

Let $H_i = (X_i^{(1)}, R_i^{(1)}, C_i)$ be i.i.d. complete historical records under $p_\star$. In the original pair notation below, $(X_i^{(1)}, R_i^{(1)})$ abbreviates the anchor together with its associated fixed context: the expectations integrate over H_i, and the model conditionals retain C_i. This convention does not assume that a stripped configuration–result pair determines its context. For each fixed target configuration $X^{(T)}$, define the *prior predictive distribution*, the anchor-marginal of DeepImagine's single-pathway prediction:

$$M_\theta\big(B \,\big|\, X^{(T)}\big) \;:=\; \mathbb{E}_{(X^{(1)},R^{(1)})\sim p_\star}\Big[P_\theta\big(B \,\big|\, R^{(1)}, X^{(1)}, X^{(T)}\big)\Big], \qquad B \in \mathcal{B}.$$

Definition 9 (Calibration). The model P_θ is *calibrated* at the target configuration $X^{(T)}$ if its prior predictive coincides with the true target conditional as probability measures: for every fixed measurable event $B \in \mathcal{B}$,

$$M_\theta\big(B \,\big|\, X^{(T)}\big) \;=\; P_{p_\star}\big(B \,\big|\, X^{(T)}\big).$$

Proposition (restated) (Convergence of Aggregating Counterfactual Pathways)**.** Suppose P_θ is calibrated at the target configuration $X^{(T)}$ in the sense of Definition 9. Then, for every measurable event $B \in \mathcal{B}$,

$$\frac{1}{n}\sum_{i=1}^{n} P_\theta\big(B \,\big|\, R_i^{(1)}, X_i^{(1)}, X^{(T)}\big) \;\longrightarrow\; P_{p_\star}\big(B \,\big|\, X^{(T)}\big) \qquad \text{a.s. as } n \to \infty.$$

Proof. Fix a measurable event $B \in \mathcal{B}$ and a target configuration $X^{(T)}$. Define

$$Y_i \;:=\; P_\theta\big(B \,\big|\, R_i^{(1)}, X_i^{(1)}, X^{(T)}\big), \qquad i = 1, 2, \ldots.$$

Each Y_i is a probability, so $0 \le Y_i \le 1$. Because Y_i is a fixed measurable function of the complete record H_i and those records are i.i.d., the random variables $Y_1, Y_2, \ldots$ are also i.i.d. and uniformly bounded, hence integrable.

By Kolmogorov's strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \;\longrightarrow\; \mathbb{E}[Y_1] \qquad \text{a.s. as } n \to \infty.$$

By definition of the prior predictive,

$$\mathbb{E}[Y_1] \;=\; \mathbb{E}_{(X_1^{(1)},R_1^{(1)})\sim p_\star}\Big[P_\theta\big(B \,\big|\, R_1^{(1)}, X_1^{(1)}, X^{(T)}\big)\Big] \;=\; M_\theta(B \mid X^{(T)}).$$

By the calibration assumption (Definition 9),

$$M_\theta(B \mid X^{(T)}) \;=\; P_{p_\star}(B \mid X^{(T)}).$$

$$\frac{1}{n}\sum_{i=1}^{n} P_\theta\big(B \,\big|\, R_i^{(1)}, X_i^{(1)}, X^{(T)}\big) \;\longrightarrow\; P_{p_\star}\big(B \,\big|\, X^{(T)}\big) \qquad \text{a.s.}$$

Since $B \in \mathcal{B}$ was arbitrary, this completes the proof. □

This is almost-sure convergence for each fixed event. The probability-one set may depend on that event; the statement alone is not uniform convergence over all measurable events. The next subsection supplies conditions for total-variation convergence.

A.5 Relaxed Convergence under Approximate Calibration

We state and prove a relaxed version of Proposition 2 that allows the calibration assumption to hold only approximately. This relaxation is important in practice: averaging independent predictions can reduce sampling variance without eliminating their systematic calibration discrepancy. We adopt the same setup and notation as Appendix A.4, including the prior predictive

$$M_\theta\big(\cdot \,\big|\, X^{(T)}\big) \;:=\; \mathbb{E}_{(X^{(1)},R^{(1)})\sim p_\star}\Big[P_\theta\big(\cdot \,\big|\, R^{(1)}, X^{(1)}, X^{(T)}\big)\Big],$$

and the aggregated prediction $\bar{P}_n(\cdot \mid X^{(T)}) := \frac{1}{n}\sum_{i=1}^n P_\theta(\cdot \mid R_i^{(1)}, X_i^{(1)}, X^{(T)})$.

We assume that, for $p_\star$-almost every anchor pair $(X^{(1)}, R^{(1)})$, the predictive measure $P_\theta(\cdot \mid R^{(1)}, X^{(1)}, X^{(T)})$ admits a jointly measurable density $p_\theta(r \mid R^{(1)}, X^{(1)}, X^{(T)})$ with respect to a common σ-finite reference measure ν on $(\mathcal{R}, \mathcal{B})$ (e.g., counting measure for countable result strings, or Lebesgue measure when applicable). Under this assumption, the prior predictive admits density $m_\theta(r \mid X^{(T)}) := \mathbb{E}[p_\theta(r \mid R^{(1)}, X^{(1)}, X^{(T)})]$, and the aggregated predictive admits density

$$\bar{p}_n(r \mid X^{(T)}) \;=\; \frac{1}{n}\sum_{i=1}^n p_\theta\big(r \,\big|\, R_i^{(1)}, X_i^{(1)}, X^{(T)}\big).$$

Definition 10 (**ϵ-Approximate Calibration**). The model P_θ is *ϵ-approximately calibrated* at the target configuration $X^{(T)}$ if

$$\big\|M_\theta\big(\cdot \,\big|\, X^{(T)}\big) \;-\; P_{p_\star}\big(\cdot \,\big|\, X^{(T)}\big)\big\|_{\mathrm{TV}} \;\le\; \epsilon,$$

for some $\epsilon \ge 0$. Strict calibration (Definition 9) corresponds to $\epsilon = 0$.

Proposition 3 (**Relaxed Convergence of Aggregating Counterfactual Pathways**). *Suppose P_θ is ϵ-approximately calibrated at $X^{(T)}$ (Definition 10), and the predictive measures admit densities w.r.t. a common reference measure ν. Then*

$$\limsup_{n\to\infty} \big\|\bar{P}_n\big(\cdot \,\big|\, X^{(T)}\big) \;-\; P_{p_\star}\big(\cdot \,\big|\, X^{(T)}\big)\big\|_{\mathrm{TV}} \;\le\; \epsilon \qquad \textit{a.s.}$$

With $\epsilon = 0$, the additional density assumptions strengthen the eventwise conclusion of Proposition 2 to TV convergence.

Proof. For each fixed $r \in \mathcal{R}$ at which $m_\theta(r \mid X^{(T)}) < \infty$, the random variables $\big\{p_\theta(r \mid R_i^{(1)}, X_i^{(1)}, X^{(T)})\big\}_{i\ge 1}$ are i.i.d., non-negative, with mean

$$\mathbb{E}\big[p_\theta(r \mid R_1^{(1)}, X_1^{(1)}, X^{(T)})\big] \;=\; m_\theta(r \mid X^{(T)}) \;<\; \infty.$$

By Kolmogorov's strong law of large numbers,

$$\bar{p}_n(r \mid X^{(T)}) \;\longrightarrow\; m_\theta(r \mid X^{(T)}) \qquad \text{a.s.}$$

Since $m_\theta(\cdot \mid X^{(T)})$ is a probability density, it is finite ν-a.e., so this convergence holds a.s. for ν-a.e. r.

Define the joint complement set

$$E^c \;:=\; \big\{(\omega, r) \in \Omega \times \mathcal{R} \,:\, \bar{p}_n(r \mid X^{(T)})(\omega) \not\to m_\theta(r \mid X^{(T)}) \text{ as } n \to \infty\big\},$$

which is jointly measurable. For ν-a.e. fixed r, $\mathbb{P}\big(\{\omega : (\omega, r) \in E^c\}\big) = 0$. By Fubini's theorem,

$$0 \;=\; \int_{\mathcal{R}} \mathbb{P}\big(\{\omega : (\omega, r) \in E^c\}\big)\, \nu(dr) \;=\; \int_\Omega \nu\big(\{r : (\omega, r) \in E^c\}\big)\, d\mathbb{P}(\omega).$$

Since the integrand on the right is non-negative, $\nu\big(\{r : (\omega, r) \in E^c\}\big) = 0$ for $\mathbb{P}$-a.e. ω. That is, there exists a probability-one event Ω_∞ on which the convergence $\bar{p}_n(r \mid X^{(T)}) \to m_\theta(r \mid X^{(T)})$ holds simultaneously for ν-a.e. r.

On the event Ω_∞, we have: (i) $\bar{p}_n(\cdot \mid X^{(T)})$ and $m_\theta(\cdot \mid X^{(T)})$ are non-negative probability densities w.r.t. ν, both integrating to one; and (ii) $\bar{p}_n(\cdot \mid X^{(T)}) \to m_\theta(\cdot \mid X^{(T)})$ pointwise ν-a.e. Scheffé's lemma asserts that pointwise a.e. convergence of probability densities implies L^1 convergence:

$$\int_{\mathcal{R}} \left|\bar{p}_n(r \mid X^{(T)}) - m_\theta(r \mid X^{(T)})\right| \nu(dr) \;\longrightarrow\; 0 \qquad \text{on } \Omega_\infty.$$

Since the L^1 distance between densities equals twice the total variation distance between the corresponding probability measures,

$$\left\|\bar{P}_n\big(\cdot \,\big|\, X^{(T)}\big) \;-\; M_\theta\big(\cdot \,\big|\, X^{(T)}\big)\right\|_{\mathrm{TV}} \;\longrightarrow\; 0 \qquad \text{a.s.}$$

By the triangle inequality for total variation,

$$\left\|\bar{P}_n - P_{p_\star}\right\|_{\mathrm{TV}} \;\le\; \underbrace{\left\|\bar{P}_n - M_\theta\right\|_{\mathrm{TV}}}_{\to 0 \text{ a.s.}} + \underbrace{\left\|M_\theta - P_{p_\star}\right\|_{\mathrm{TV}}}_{\le \epsilon \text{ by Def. 10}},$$

where all measures are conditioned on $X^{(T)}$. Taking $\limsup$ as $n \to \infty$,

$$\limsup_{n\to\infty} \left\|\bar{P}_n(\cdot \mid X^{(T)}) - P_{p_\star}(\cdot \mid X^{(T)})\right\|_{\mathrm{TV}} \;\le\; 0 + \epsilon \;=\; \epsilon \qquad \text{a.s.}$$

This completes the proof. □

B Supervision and Training Details

B.1 Observed Labels and Synthetic Intermediate Results

For a natural outcome pair from arm k in trial i, the first observed result is an input and the second is the target. GPT-5 supplies the connecting explanation, and the operator is trained to generate the explanation followed by the observed target result:

$$P_\theta^O\Big(\text{concat}\,(\text{reasoning}, R^{(i)}_{j_2,k}) \;\Big|\; R^{(i)}_{j_1,k},\, X^{(i,j_1,k)},\, X^{(i,j_2,k)}\Big). \tag{6}$$

The corresponding arm-update objective exchanges the arm index while retaining the outcome measure. These response targets are fitted with standard cross-entropy supervised fine-tuning. Observed aggregate comparisons do not supply both potential outcomes for any individual patient.

For two retrieved historical trials C_1, C_2, superscripts $(1), (2)$ identify their observed configurations and results; $(a), (b), (c)$ identify teacher-generated results. The three-factor synthetic bridge is

$$\begin{aligned}&(N^{(1)}, A^{(1)}, O^{(1)}, R^{(1)}) \to (N^{(2)}, A^{(1)}, O^{(1)}, R^{(a)})\\ &\to (N^{(2)}, A^{(2)}, O^{(1)}, R^{(b)}) \to (N^{(2)}, A^{(2)}, O^{(2)}, R^{(c)}),\end{aligned} \tag{7}$$

The historical context C_1 is fixed and suppressed in these tuples. The teacher receives both observed endpoint results $R^{(1)}, R^{(2)}$, but $R^{(a)}, R^{(b)}, R^{(c)}$ remain generated. In particular, $R^{(2)}$ is reference evidence, not automatically the observed label for the last imagined configuration under C_1. Matching N, A, O does not align every contextual attribute. The first synthetic tuple has an observed input result; later tuples have generated input results. This is why the term synthetic does not imply that every endpoint of every tuple is generated.

B.2 Training Details

We use GPT-5 (OpenAI, 2025), with medium reasoning effort and low verbosity, to generate synthetic counterfactual reasoning trajectories for the three endogenous factors: enrollment count N, study arm A, and outcome measure O, yielding 45k synthetic training tuples per factor (Section 4.3). For study-arm and outcome-measure perturbations, these synthetic tuples augment 57k and 82k natural counterfactual pairs, respectively; for enrollment-count perturbations, no natural counterfactuals are available, so training relies solely on synthetic data.

We fine-tuned a separate Gemma-3-12B-it (Gemma Team & Google DeepMind, 2025) model for each of the three local operators P_θ^Φ, $\Phi \in \{N, A, O\}$, using the identical hyperparameter configuration

Table 5: Three-factor supervision counts before the 10% tuning holdout. Counts are training tuples, not independent trials. All accompanying explanations are generated by GPT-5.

Operator	Natural	Synthetic	Total
Enrollment N	—	45k	45k
Study arm A	57k	45k	102k
Outcome measure O	82k	45k	127k
Total	139k	135k	274k

listed in Table 6.The same configuration is also used for the two ablation training schemes in Section 5.2 (A1: direct configuration-to-result prediction; A2: one-step prediction from a historical trial result), in both their with-reasoning and without-reasoning variants. LoRA adapters were applied to all linear layers with rank $r = 32$, scaling factor $\alpha = 64$, and dropout 0.05. We trained for one epoch with a per-device batch size of 1 and gradient accumulation of 2. Optimization used AdamW with a peak learning rate of 5×10^{-5}, cosine decay to a minimum of 10% of the peak learning rate, 5% warm-up steps, zero weight decay, and gradient clipping at 1.0. The maximum sequence length was 8,192 tokens. Training was performed on a single node with 7×80GB H100 GPUs using FlashAttention-4.

Table 6: Hyperparameters used uniformly across all Gemma-based experiments, including the three local operators P_θ^Φ for $\Phi \in \{N, A, O\}$ and both ablation training schemes (A1, A2).

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Model and fine-tuning		*Hardware and precision*	
Base model	Gemma-3-12B-it	Compute	7×GPU, single-node DDP
Fine-tuning method	LoRA (all linear layers)	Attention	FlashAttention-4
LoRA rank / α / dropout	32 / 64 / 0.05		
Optimization		*Batching and schedule*	
Optimizer	AdamW	Epochs	1
Maximum learning rate	5×10^{-5}	Per-GPU batch size	1
Minimum learning rate	5×10^{-6}	Gradient accumulation	2
LR scheduler	Cosine decay	Effective batch size	14
Warmup ratio	5%	Max sequence length	8,192
Weight decay	0.0		
Gradient clipping	1.0		

C Evaluation Protocols and Additional Analyses

C.1 Retrieval, Prediction, and Aggregation

The evaluation cohort used here comprises 701 Summer 2025 CT Open questions: 442 positives and 259 negatives, with a result-availability cutoff of 2025-08-31 (Wang et al., 2026). This is the cohort specified in our experiments, not a claim that every public CT Open release contains exactly 701 questions. The API snapshots are GPT-5-2025-08-07, GPT-5-mini-2025-08-07, and o3-mini-2025-01-31. Each method predicts one binary answer per question; macro-F1 averages the two class-specific F1 scores.

The retrieval representation uses the available registry description, excluding results. NV-Embed-v2 (Lee et al., 2025) supplies text embeddings and cosine similarity; the retrieval protocol uses language-model relevance verification. Selected historical results are then available to the prediction stage, whereas target results are withheld. Direct receives only the target question and description. RAG adds up to seven historical trials and observed results per target arm. The corresponding DeepImagine comparison uses the same historical evidence to initialize local pathways.

For each target arm, the three-factor pathway replaces enrollment, then study arm, then outcome measure. Each operator conditions on the current and next configurations, the preceding result, and

fixed historical context. Its response contains an explanation and updated result; the parsed result is passed to the next operator. Terminal predictions are averaged on the target endpoint's numerical scale before the final binary comparison. This is not a vote over intermediate answers. For API-model comparisons, the generating backbone performs the final comparison. Forecasting that a trial will meet a significance criterion is not itself a significance test on target observations.

For two arms, n anchors per arm, and $T-1$ updates, the local stage uses $2n(T-1)$ operator calls. At $T=4$, this is $6n$ calls, excluding retrieval verification and final comparison. This is a structural call count, not a measurement of tokens or latency. The same-anchor control matches pathway counts, not token lengths or cost. The Gemma strategy comparison likewise uses the same fine-tuning hyperparameters but not the same number of tuples: A1 and A2 each use 127k pairs, whereas the three operators use 274k tuples in total before the tuning holdout.

C.2 Conventional and Task-Adapted Clinical Predictors

Shared features and preprocessing. The six conventional baselines use available numerical, categorical, and textual trial features, including trial descriptions, study arms, and outcome measures. Numerical features are used directly, with column-median imputation for missing values; the kNN+RF baseline instead uses nearest-neighbor imputation (Lo et al., 2019). Categorical features are one-hot encoded, with missing entries replaced by the most frequent category. Text fields are embedded with NV-Embed-v2 (Lee et al., 2025); missing fields are represented by zero vectors. PLR uses the full embeddings, whereas the other five models reduce text representations to 32 dimensions using PCA. PLR therefore differs in both classifier and representation dimension; its advantage should not be attributed to the classifier alone.

Feed-forward network. The FFNN maps the concatenated trial features directly to the prediction label. It has two hidden layers of sizes 512 and 128 with ReLU activations, regularization coefficient 10^{-4}, initial learning rate 10^{-3}, and at most 200 iterations. The checkpoint with the best validation performance is used for evaluation. Unlike local-operator training, this model does not receive supervision on intermediate result transitions. Table 7 preserves the full supplied conventional-model settings.

Clinical prediction models. HINT (Fu et al., 2022) combines representations of drug molecules, diseases, and eligibility criteria. In the task adaptation, drug SMILES, disease ICD codes, and eligibility descriptions retain their original roles, while outcome descriptions and the compared arms supply the information needed for an arm–endpoint question. SPOT (Wang et al., 2023c), LIFTED (Zheng et al., 2025), and MEXA-CTP (Zhang et al., 2025) are likewise adapted to the present prediction target while retaining their core architectures. Hyperparameters follow the original implementations, and checkpoints are selected by validation performance. These scores evaluate the supplied task adaptations; they are not the methods' published performance on their original trial-level tasks.

Table 7: Conventional-model hyperparameters supplied with the updated experiments. Triplets denote Endpoint / Superiority / Comparative Effect settings; a single value applies to all three. The extra task settings are retained for reproducibility of the source protocol, not presented as additional evaluations in this paper.

Model	Hyperparameter	Setting
PLR (Lo et al., 2019)	Penalty	ℓ_2
	Inverse regularization C	1.0
	Solver	`lbfgs`
	Maximum iterations	3000
Random forest (Lo et al., 2019)	Estimators	300 / 200 / 200
	Maximum depth	None
	Minimum samples per leaf	6 / 4 / 4
kNN + random forest (Lo et al., 2019)	Imputation neighbors	5
	Neighbor weights	Uniform
	Estimators	200 / 100 / 100
	Maximum depth	None
	Minimum samples per leaf	4
FFNN (Tranchevent et al., 2019)	Hidden layers	(512, 128)
	Activation	ReLU
	Regularization α	10^{-4}
	Initial learning rate	10^{-3}
	Maximum iterations	200
AdaBoost (Fan et al., 2020)	Estimators	300 / 300 / 200
	Learning rate	1.0
	Base-tree maximum depth	2
	Base-tree minimum samples per leaf	2
XGBoost (Siah et al., 2021)	Estimators	300
	Learning rate	0.1 / 1.0 / 1.0
	Maximum depth	12

C.3 Sensitivity Studies and Statistical Reporting

The order experiment evaluates every permutation of N, A, O with GPT-5 on 701 questions, holding anchors, instructions, and the judge fixed within that study. The anchor-relevance experiment uses the canonical $N \rightarrow A \rightarrow O$ order and compares a relevant historical anchor with a random training-corpus anchor. Relevant selection uses a top-10 NV-Embed-v2 shortlist followed by GPT-5 selection. The studies retain their distinct controls; neither is substituted for a seven-anchor measurement.

The reported bootstrap intervals quantify uncertainty in individual macro-F1 estimates, rather than standard deviations or uncertainty in paired differences between methods. Therefore, interval overlap or separation alone does not establish statistical significance. Uncertainty estimates for the Gemma ablations, conventional baselines, and sensitivity studies remain unreported. Training-seed variability, paired uncertainty, and measured token or latency costs also remain unreported. In particular, the 3.35–4.61-point differences in the perturbation-order study are descriptive and do not establish that all pairwise differences are statistically significant.

C.4 Clinical Interpretation and Explanation Assessment

DeepImagine uses counterfactual imagination procedurally, not as a claim that each update is an identified causal intervention (Pearl, 2009). The three factors have different roles. Enrollment concerns sampling design: with population and intervention fixed, sample size need not change the population effect, but can affect estimate precision and the chance of meeting a significance criterion. The arm update concerns an intervention contrast. The outcome update changes the measured quantity, not the assigned treatment.

Natural arm pairs are observed aggregate comparisons from a controlled trial, not both potential outcomes for individual participants. Exact single-factor matching additionally requires agreement of enrollment and the other encoded attributes; randomization alone does not guarantee equal arm sizes. Natural outcome pairs may differ in assessment time, missingness, and analysis population. The reported extraction counts do not themselves establish that every pair satisfies strict matching. Retrieval identifies related examples but does not establish transportability across patient populations or remove contextual differences.

A biomedical expert assessed 90 generated explanations from the source supervision corpus: 87 were clinically sound, two potentially problematic, and one uncertain. The assessment was not broken down by factor or by natural versus synthetic supervision. It checks explanation plausibility, not unobserved intermediate numerical labels. Observed natural-pair labels remain distinct from the teacher-generated explanations that accompany them.