THE UNIVERSITY OF CHICAGO

Predicting Clinical Trial Completion and Success Using Machine Learning and Natural Language Processing

By

Jiazheng Li

May 2025

A paper submitted in partial fulfillment of the requirements for
the Master of Arts degree in the Master of Arts in
Computational Social Science

Faculty Advisor: Professor Yuan Ji

Preceptor: Fabricio Vasselai

Abstract

This study introduces a dual-task machine learning framework for predicting both the operational completion and scientific success of clinical trials using data from ClinicalTrials.gov. Leveraging structured trial metadata and unstructured textual descriptions, we develop predictive models that assess whether trials are likely to complete and whether they meet their primary endpoints. For the first task, ensemble models like XGBoost significantly outperform traditional baselines, particularly when enriched with contextual embeddings derived from BioLinkBERT. For the second task, we propose a novel large language model (LLM)-driven annotation pipeline using GPT-4o-mini to label trial success based on publication content. Human evaluation confirms its high accuracy. Across both tasks, our framework demonstrates the value of combining structured features, natural language processing, and scalable LLM-based labeling to improve the understanding and forecasting of clinical trial performance. This approach not only enhances predictive accuracy but also contributes to better utilization of large-scale biomedical data.

Keywords: Clinical Trials; Machine Learning; Natural Language Processing; Clinical Trial Completion; Outcome Prediction; GPT-4; BioLinkBERT; ClinicalTrials.gov; Publication Analysis

1 Introduction

Clinical trials play a pivotal role in advancing medical knowledge by evaluating the safety and efficacy of new treatments, interventions, and diagnostic methods. As these trials are essential for validating medical innovations, predicting their outcomes can offer significant insights for researchers, practitioners, and policymakers. Despite the importance of clinical trials, many face challenges such as prolonged time frames, recruitment difficulties, and rising costs, often leading to premature termination or inconclusive results.

This study seeks to predict two key outcomes: (1) the completion or termination status of clinical trials and (2) the effectiveness of trials, measured by whether the associated publication reports positive or negative results—as determined by a large language model (LLM) analyzing the publication abstract for evidence of primary endpoint achievement. By employing large language models (LLMs) to tokenize and analyze the associated published research, this study aims to uncover patterns and predictors that influence both trial completion and effectiveness.

The primary data source for this research is the ClinicalTrials.gov database, a comprehensive repository that provides detailed information on clinical studies conducted around the world. Established under the Food and Drug Administration Amendments Act (FDAAA) of 2007, this database contains records of trials that meet specific criteria, such as those involving FDA-regulated drug or device products, studies with U.S. sites, or trials using products exported from the U.S. for research. The dataset includes key details about each trial, including its registration information, current status, sponsor, intervention type, and, when available, outcomes. Additionally, it tracks whether a study is completed, terminated prematurely, or still ongoing. With over half of millions of entries, ClinicalTrials.gov offers valuable data for predictive modeling, providing insights into the success, failure, and potential impact of clinical trials across a wide range of medical disciplines.

This research builds on existing work that highlights the growing complexity and regulatory demands of clinical trials, particularly in chronic disease research, where non-novel drug interventions require larger participant groups and more stringent oversight. Identifying early signals that indicate trial success or failure could help optimize resource allocation and reduce the financial burden on sponsors and investigators.

2 Literature Review

2.1 General Landscape of Clinical Trials

Clinical trials are a cornerstone of medical research, aiming to evaluate the safety and efficacy of new treatments, interventions, and diagnostics. They provide critical data for advancing healthcare but face significant challenges such as long timelines, high costs, and recruitment difficulties. The design and dissemination of clinical trials vary across funders, with industry-funded trials displaying higher dissemination rates compared to non-profit ones [1]. These trials are often more robust, utilizing randomization, blinding, and multi-national sites, contributing to their higher success rates. However, the landscape remains complex, with large portions of trials not yielding immediate publications, despite legal mandates like the FDA Amendments Act (FDAAA) of 2007, which requires certain trials to be registered on ClinicalTrials.gov database [2].

2.2 Trial Termination

The issue of trial termination is critical, as early termination results in wasted resources, disrupted patient care, and delayed scientific advancements. Several studies have sought to understand the multifaceted reasons behind early termination, which range from logistical challenges to scientific and ethical concerns.

Zhang, Ellen(2023)[3] conducted a cross-sectional analysis of oncology trials and found that 22.7% of trials were terminated early, with the leading causes being poor accrual (34.5%) and ambiguous or complex reasons. Early terminations were more prevalent in Phase 2 trials, especially those conducted solely in the U.S., as these trials often lacked the scale or infrastructure of larger, multi-national studies. Non-industry-funded trials also faced higher termination rates, highlighting the vulnerability of trials dependent on public or non-profit funding. Pak et al. (2015) [4] added further insight by creating an ontology to classify termination reasons. They found that insufficient enrollment was the single most common cause of termination, accounting for 33.7% of cases. Funding issues (7.6%) and business decisions (7.3%) were also significant contributors, reflecting the financial instability that can affect long-term clinical studies. Surprisingly, efficacy concerns accounted for only 6.8% of terminations, indicating that logistical and financial challenges often outweigh scientific failures in causing early trial closure.

Beyond these critical factors, other studies have highlighted the complexity of trial termination. Regulatory and ethical challenges play a substantial role, especially in trials involving vulnerable populations or cutting-edge therapies like gene or stem cell treatments, which require more stringent regulatory oversight. Trials terminated due to safety concerns are relatively rare, accounting for about 6.7% of cases, but there are instances where trials were halted due to overwhelmingly positive results during interim analysis, accounting for 0.5%. Moreover, trials with complex eligibility criteria or requiring highly specific patient subgroups often face difficulties in recruitment, increasing the likelihood of termination. For instance, studies focused on rare conditions, such as Mycosis

Fungoides or Sézary syndrome, often fail to enroll enough participants, contributing to early closure.

Geographic factors also play a role, with trials conducted at single sites or limited to national settings being more prone to early termination compared to multi-site, international trials. Limited geographic scope can lead to recruitment challenges, especially when trials are located in regions with smaller patient populations or fewer specialized medical centers. Multi-country trials, by contrast, benefit from broader patient access and a more diversified regulatory environment, making them more resilient to termination. Additionally, operational and administrative delays such as slow site activation, prolonged IRB approval processes, or extended contract negotiations can also disrupt trial timelines, exhausting financial resources before recruitment targets are met. Another significant factor is the sponsorship of the trial. Industry-sponsored trials, especially those run by large pharmaceutical companies, tend to be more financially stable and adaptable to unforeseen challenges, such as recruitment difficulties or regulatory hurdles. On the other hand, non-profit and academic trials are more susceptible to funding shortfalls, and their reliance on limited grants makes them more vulnerable to termination. These trials often lack the financial flexibility needed to overcome mid-trial challenges, particularly in situations where additional recruitment efforts or extended timelines are required. Furthermore, engagement issues between investigators and patients can also lead to termination. Trials that fail to effectively engage investigators or participants often see higher dropout rates and lower protocol adherence, which disrupt the integrity of the study. Investigator-led trials, particularly those at academic institutions, may face the challenge of shifting priorities or investigator turnover, further complicating the trial's continuation. External market forces, such as the introduction of a new competing therapy, can also impact the relevance of ongoing trials, leading to their early closure, especially in cases where the newer therapy offers superior efficacy.

2.3 Machine Learning for Predicting Trial Outcomes

Recent advancements in machine learning for healthcare applications demonstrate the increasing role of supervised and unsupervised algorithms, like deep learning and support vector machines, in predicting clinical outcomes by analyzing both structured and unstructured medical data. This approach, as highlighted by Habehh and Gohel (2021)[5], can be leveraged to predict trial outcomes, enhance recruitment strategies, and improve trial completion rates. In recent years, machine learning (ML) has become increasingly relevant in predicting the outcomes of clinical trials, including their completion status. Elkin and Zhu (2021)[6] applied machine learning models to predict clinical trial terminations using a combination of feature engineering and embedding learning. The study focused on a subset of 68,999 trials data derived from ClinicalTrials.gov and generated 640 features, including statistical, keyword, and embedding features which created by Doc2Vec model based on trial descriptions. They employed models such as XGBoost, Random Forest and neural network to identify key predictors of termination, such as strict eligibility criteria, sponsor types, and complex study designs. This work highlights how factors like oncology trials, especially those focusing on

rare cancers such as Mycosis Fungoides, are particularly prone to termination due to recruitment challenges. Similarly, Kavalci and Hartshorn (2023) [7] applied machine learning models to predict the success or failure of clinical trials using registered clinical trial data from the AACT database. By incorporating features like study characteristics, eligibility criteria, and disease categories, it identified XGBoost as the best-performing model with an ROC-AUC of 80%. The study used SHAP values to explain the model's predictions, highlighting that strict inclusion/exclusion criteria and recruitment failure were key predictors of trial termination, providing valuable insights for improving trial designs. Another notable study by Geletta et al. (2019) [8]used Latent Dirichlet Allocation (LDA) to predict trial termination by modeling the unstructured descriptions of clinical trials. Their approach showed that topics such as surgical procedures and complex study designs were key predictors of early termination. By combining structured data with topic modeling, they achieved higher sensitivity in predicting trial outcomes compared to traditional models.

2.4 Predicting Publication and Outcomes of Clinical Trials

Moving beyond predicting trial completion, some studies have explored the likelihood of trials being published and the outcomes of such publications. Wang et al. (2022)[9] developed a model to predict whether a trial would result in a published paper, incorporating both structured trial data and unstructured trial descriptions. Their findings demonstrated that larger trials and those with more complex designs had higher publication rates, with textual features (e.g., trial descriptions) significantly improving predictive accuracy. However, this study stopped short of predicting the nature of the publication outcomes—whether the results were positive or negative. In a more recent retrospective analysis conducted by White and Parsons [10], they found despite the large number of clinical prediction model studies registered, a significant proportion remains unpublished, indicating a high degree of research waste. This also study highlights the potential for publication bias in the field, where only successful or promising models might be published, inflating expectations about model performance in practice. The publication Bias were also supported by W Jones [11], who also found the issue of non-publication was more prevalent in industry-funded trials. This suggests that industry sponsorship may influence the decision not to publish, particularly when outcomes are unfavorable or lack commercial value. Further exacerbating the problem, most unpublished trials, despite involving large numbers of participants, also lacked results on ClinicalTrials.gov, limiting the availability of trial data even in the registry, therefore do not contribute to the scientific community due to non-publication.

More recently, Gao et al. (2024) [12] introduced the Clinical Trial Outcome (CTO) benchmark, a large-scale, open-source dataset containing over 125,000 drug and biologics trials with systematically labeled outcomes. Their work addresses a major bottleneck in the field: the scarcity of high-quality, reproducible outcome labels for clinical trials. CTO aggregates weak supervision signals—including LLM-generated interpretations of PubMed abstracts, sentiment from news headlines, trial phase transitions, and p-values—to generate pseudo-labels, which are further refined using a label model.

Importantly, they manually annotated over 2,500 trials to validate labeling quality and achieved high F1 scores (up to 94% in Phase 3). Their work represents a key advancement in scalable outcome labeling and sets a new standard for benchmarking predictive models in drug development. This framework highlights the potential of combining structured trial metadata with unstructured textual evidence for more reliable outcome inference.

2.5 Large Language models for Tokenization and sentiments analysis

Early methods for generating word embeddings laid the foundation for capturing semantic relationships in text data. Among these, Word2Vec, developed by Mikolov et al. [13], was one of the most influential. Using shallow neural networks, Word2Vec creates embeddings by learning from word co-occurrence patterns, producing vector representations that capture basic semantic similarities. However, as a context-independent model, Word2Vec has limitations in handling complex, contextual meanings, which newer transformer-based models address more effectively. With advancements of deep transformer architecture, models like GPT and BERT have demonstrated superior performance in tokenizing and processing textual data compared to traditional models like Word2Vec or TF-IDF. BERT (Bidirectional Encoder Representations from Transformers) [14] introduced a powerful bidirectional approach to capture word context, significantly advancing NLP tasks like text classification and sentiment analysis. Building on BERT, ClinicalBERT is fine-tuned on clinical notes from electronic health records, making it well-suited for patient-level healthcare applications [15]. In contrast, BioLinkBERT is trained on biomedical literature and document link structures, enabling it to capture richer semantic relationships across scientific publications—particularly useful for research-level tasks such as trial description embedding [16]. In recent studies, these embedding models were used to generate embeddings for clinical trial descriptions, capturing nuanced medical terminology and enhancing the contextual relevance of trial-specific details. One of the most cutting-edge models, GPT-4 by OpenAI, represents a significant leap from earlier LLMs. It excels in tokenizing large bodies of text while capturing subtle, domain-specific nuances, which is critical when processing clinical trial data. GPT-4's ability to understand context deeply allows it to identify relationships between complex medical terms and their corresponding sentiment (e.g., positive, negative, neutral). This makes it particularly suitable for identifying the sentiment of research publications related to clinical trials, a task where older models like BERT might struggle due to limitations in capturing broader contextual relationships. In Gao et al.'s paper [12], one of their key label-generation methods involves the use of ChatGPT to summarize clinical trial outcomes from associated publication abstracts. They manually evaluated ChatGPT's outputs on 100 trials and reported an accuracy of 89%, indicating high but imperfect reliability for outcome labeling.

These models, while not inherently designed to directly resolve issues like selection bias or publication bias, offer powerful tools to detect and address these biases. Through context-aware tokenization and sentiment labeling, LLMs can flag patterns in trial data that suggest skewed

participant demographics or restrictive eligibility criteria, which are often indicative of selection bias. Moreover, by performing sentiment analysis on published results, LLMs can help identify the extent of publication bias, where trials with positive outcomes are disproportionately reported. In summary, the field of tokenization and sentiment analysis in healthcare has been revolutionized by recent LLMs. These models bring superior contextual understanding, improved accuracy in labeling positive or negative sentiment in clinical publications, and more efficient handling of large datasets compared to their predecessors, making them the ideal tools for this research on clinical trials.

2.6 Research Contribution

This research offers several key contributions to the field of clinical trial outcome prediction. First, this study introduces a **two-task framework** that jointly addresses the operational feasibility (completion status) and scientific success (achievement of primary endpoints) of clinical trials. While prior studies have largely focused on predicting trial completion or publication likelihood in isolation, this work expands the scope by directly evaluating whether trials achieve their intended scientific objectives, offering a more holistic view of trial performance. In doing so, it also contributes to the broader research infrastructure by demonstrating how to **systematically retrieve, clean, and operationalize data from ClinicalTrials.gov**—a resource often underutilized due to its complexity—into structured inputs suitable for machine learning applications.

Second, this study demonstrates the value of **integrating structured trial features with unstructured textual information** through modern NLP-based embeddings. By generating contextualized representations of trial descriptions using BioLinkBERT and combining them with structured tabular features, the models can capture both explicit design factors (e.g., enrollment size, sponsor type) and implicit, semantically rich information present in trial narratives. This hybrid representation substantially improves model performance, especially for predicting trial completion. The successful implementation of a **dual-tower neural network** further highlights the advantage of explicitly modeling interactions between textual and tabular data, which has been underexplored in previous literature on clinical trial prediction.

Third, the study proposes a novel use of **large language models (LLMs)**—specifically GPT-4—to automatically label the success or failure of clinical trials based on their associated publications. This addresses a major challenge in prior work, where trial success is often inferred through indirect signals such as publication status or trial progression. A recent large-scale effort by Gao et al. (2024) [12]introduced the Clinical Trial Outcome (CTO) benchmark, which constructs pseudo-outcome labels using multiple weak supervision sources, including statistical significance (p-values), phase transitions, and LLM-derived paper summaries. While this is a commendable effort in dataset scale and reproducibility, it relies heavily on structured indicators (e.g., p-values) and string similarity for linking trials to publications, which may result in label noise. For instance, some trials report multiple endpoints or have nuanced success criteria that a single p-value may

not fully capture. Moreover, relying on title similarity to link papers risks incorporating retrospective reviews or unrelated studies that merely mention the trial's intervention. By contrast, our approach leverages GPT-4 as a *domain-aware agent* in a two-step pipeline: it first verifies whether a publication directly reports the primary results of a given trial, and then assesses whether the trial met its primary endpoint based on the abstract content. This method improves both the precision of publication-trial linkage and the fidelity of outcome interpretation, yielding cleaner labels that better reflect trial-specific contexts. Previous report shows that GPT-4 exhibits human-level performance on various professional and academic benchmarks [17], offering a promising path for building high-quality labeled datasets with minimal manual intervention.

Lastly, the framework presented here is **modular and scalable**, making it suitable for a variety of downstream applications. It separates the data retrieval, preprocessing, labeling, and modeling phases, allowing individual modules to be updated or extended without requiring complete redesign. For instance, the labeling component can incorporate alternative LLMs, or the feature engineering pipeline can be adapted to new datasets or domains. This modularity not only facilitates replication but also enables customization across different therapeutic areas or prediction tasks. Additionally, the use of automated LLM-based labeling and scalable machine learning models like XGBoost allows the framework to efficiently process large datasets without substantial increases in manual labor or computational complexity, making it suitable for high-volume applications in clinical trial analytics, decision support, and resource planning.

3 Data Preparation

3.1 Data Collection

The primary data source for this study is the publicly available dataset from **ClinicalTrials.gov**(CT.gov), accessed directly through its API. ClinicalTrials.gov provides detailed protocol and result elements for all registered clinical studies. As of October 6, 2024, the dataset contained a total of 511,843 clinical studies conducted between 1999 and 2024, of which 109,361 were classified as cancer-related trials, while the remaining 402,482 were non-cancer trials. Among all registered studies, 208,444 trials were linked to at least one associated publication.

The ClinicalTrials.gov data is organized into several distinct sections, including protocol information, results, annotations, and associated documents. To extract the relevant information, we developed customized Python scripts to interact with the ClinicalTrials.gov API. The retrieved data was initially collected in JSON format. After flattening the nested structures, we compiled a total of 160 candidate features, encompassing a broad spectrum of information such as trial design characteristics, demographic information, and outcome-related documentation.

Subsequently, we applied a design-type filter to focus exclusively on interventional trials, removing all observational studies from the dataset. This filtering step was motivated by the objective of predicting clinical trial success, a task that is inherently challenging for observational studies due to their lack of controlled interventions, absence of predefined primary endpoints, and limited use of formalized outcome reporting. Observational studies typically focus on exploratory analyses, cohort descriptions, or hypothesis generation, often without explicitly defined success or failure criteria. In contrast, interventional trials are characterized by structured designs, including randomization, treatment arms, and pre-specified endpoints, which provide measurable and standardized criteria for trial evaluation. These features make interventional trials better suited for predictive modeling tasks aiming to classify or forecast trial success. After applying this criterion, 263,448 interventional trials remained for further analysis.

To assess clinical trial outcomes, we further integrated the CT.gov dataset with publication data from PubMed. By leveraging the PubMed ID (PMID) associated with each trial, we matched trials to their corresponding scientific publications. PubMed is a comprehensive database of biomedical literature, encompassing MEDLINE, life science journals, and online books. We accessed PubMed records via the PubMed API and retrieved detailed information for each linked publication, including abstracts, authorship, institutional affiliations, and publication dates. The publication data was originally provided in XML format, which we parsed and processed using custom-built extraction functions. Among the cancer-related interventional trials, 24,910 were successfully linked to at least one published article indexed in PubMed.

3.2 Data Preprocessing

Given the complexity and heterogeneity of the ClinicalTrials.gov dataset, we began by systematically reviewing the definitions, data types, and interpretability of all available features. The selection process was guided by a combination of domain relevance, data availability, and potential predictive value. In collaboration with a clinical research expert, we qualitatively assessed whether each feature was likely to contribute meaningfully to the two prediction tasks—trial completion and trial success—based on its clinical significance and representation in the dataset. Features that were overly sparse, poorly defined, or redundant were excluded. Ultimately, we retained 28 features that were determined to be informative and appropriate for inclusion in the modeling pipeline.

3.2.1 General Features

To prepare the data for modeling, several preprocessing steps were undertaken. For categorical variables, such as trial phase, sampling methods, patient demographic categories (e.g., age groups, sex eligibility), and lead sponsor types, we employed **one-hot encoding** to convert each category into binary dummy variables. This transformation enabled the models to effectively capture non-numeric information without imposing any artificial ordering on inherently unordered categories. For numerical variables — including minimum and maximum eligible patient age, planned enrollment size, and trial duration (calculated as the number of days between the study start date and the study completion date) — missing values were imputed using **mean imputation**. This simple yet robust approach addresses missingness while minimizing bias. Subsequently, **standardization** using the `StandardScaler` was applied to ensure all numerical features had a mean of zero and a standard deviation of one. This step is particularly important for models sensitive to feature scales, such as neural networks and support vector machines, to avoid any single feature disproportionately influencing the model's optimization process. For location-based features, including the country where the trial was conducted and the number of recruiting sites, as well as eligibility and exclusion criteria text fields, we derived **quantitative representations** by calculating occurrence counts (e.g., number of listed countries, number of inclusion/exclusion criteria). These transformations provided interpretable proxies for a trial's geographic scope and recruitment complexity.

The number of completed clinical trials has steadily increased, reflecting expanded global research activity and improved trial reporting practices(Figure 1). The average trial duration was approximately 2.78 years, with a standard deviation of 2.66 years (Table 1). Planned patient enrollment exhibited a highly skewed distribution, likely corresponding to large registry-based or public health interventions). The majority of trials operated with a small number of facilities (median = 1), indicating a predominance of single-center studies. Regarding eligibility complexity, trials listed a median of three inclusion and four exclusion criteria, reflecting moderate eligibility requirements overall. Phase 2 trials constituted the largest proportion of the dataset, followed by Phase 1 and Phase 3 trials. Phase 4 trials were less common, consistent with their primary role in post-marketing surveillance rather than initial therapeutic evaluation. Early Phase 1 trials and combination-phase

studies (e.g., Phase 1/2, Phase 2/3) were also present but represented a relatively small fraction of the sample. Overall, this distribution reflects the typical structure of the clinical development pipeline, with a heavier concentration of trials in early- and mid-stage phases where investigational therapies undergo preliminary and intermediate efficacy testing (Figure 2).

Table 1: Descriptive Statistics of Key Variables

Variable	**Mean**	**Std. Dev.**	**Min**	**25%**	**Median**	**75%**	**Max**
Duration of Trial (years)	2.78	2.66	-0.04	1.00	2.09	3.67	105.65
Enrollment Count	1204.77	230,187.00	0.00	27.00	60.00	148.00	100,000,000.00
Number of Facilities	7.18	30.51	0.00	1.00	1.00	2.00	1746.00
Inclusion Criteria Count	3.70	4.27	0.00	0.00	3.00	5.00	103.00
Exclusion Criteria Count	5.65	6.61	0.00	0.00	4.00	8.00	176.00

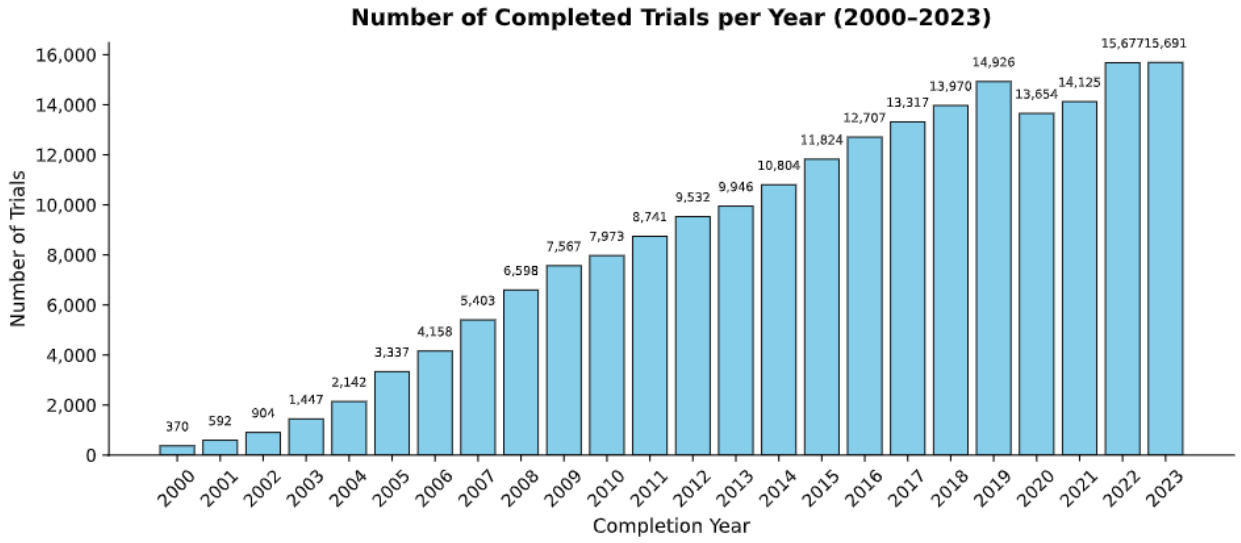

Figure 1: Number of Trials Completed by Year

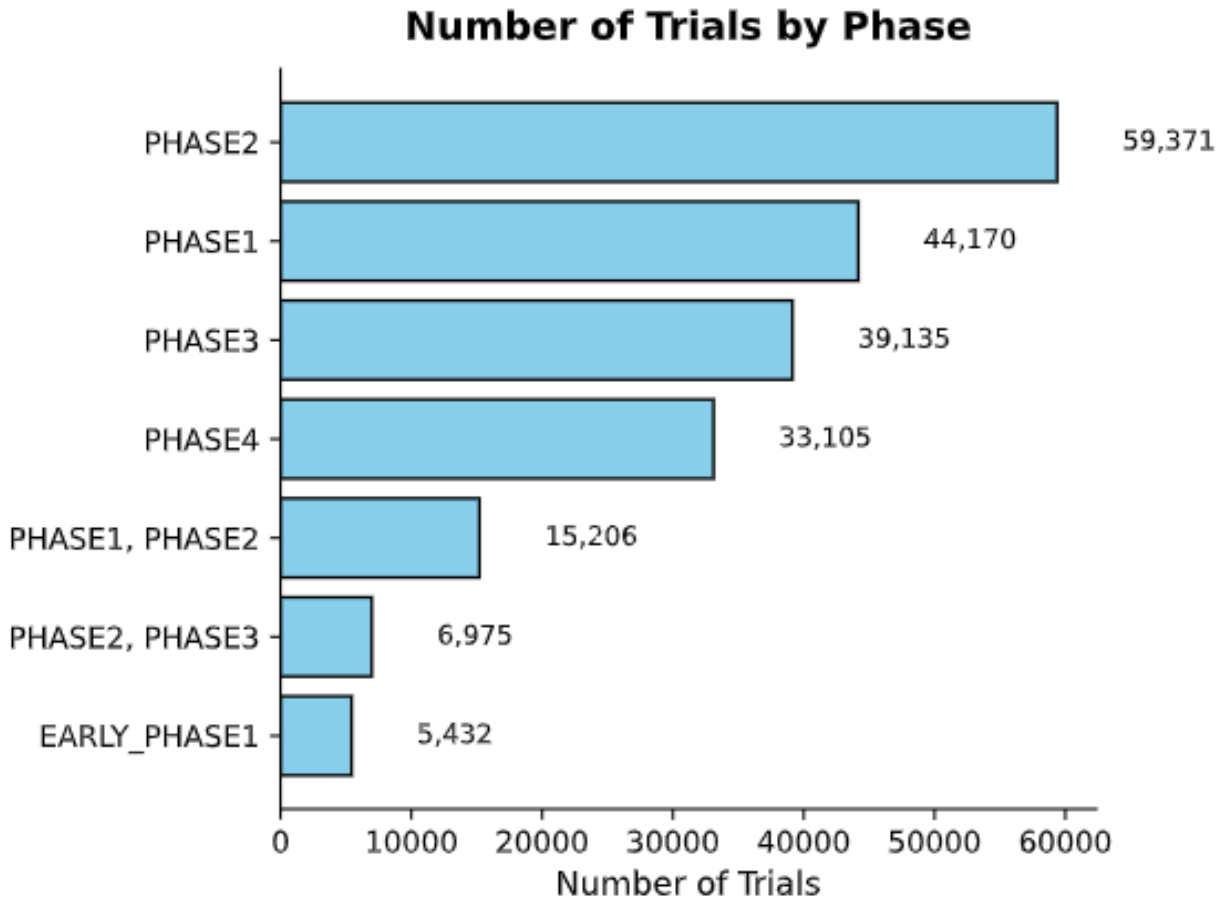

Figure 2: Number of Trials Conducted by Phase

3.2.2 Mesh Term

Given the substantial variability in the condition and intervention fields of the dataset, we performed additional preprocessing to standardize these variables using biomedical ontologies. For disease categorization, we utilized MeSH (Medical Subject Headings) ancestor terms provided in the dataset, mapping each trial to a predefined set of 22 broad disease categories (e.g., Neoplasms, Cardiovascular Diseases, Immune System Diseases). A custom extraction procedure was implemented to match available MeSH terms to this controlled vocabulary, resulting in a multi-label representation of each trial's disease types. The multi-label outputs were then converted into binary dummy variables using multi-label binarization, providing interpretable features for downstream modeling. For intervention types, we adopted a more involved process. First, MeSH identifiers linked to each intervention were used to query the official MeSH API to retrieve their corresponding tree numbers, which reflect the hierarchical structure of biomedical interventions. To address inconsistencies such as outdated or incomplete MeSH identifiers, we implemented heuristics to reformat and re-query failed cases. Extracted tree numbers were subsequently mapped to a set of broad intervention classes (e.g., Pharmaceutical Preparations, Biological Factors, Hormones and Hormone Antagonists). Finally, dummy variables were generated for these classes, yielding standardized and interpretable representations of intervention types, improving model comparability across trials.

Figure 3 and Figure 4 summarize the distribution of clinical trials by intervention and disease categories, respectively. The most common intervention types include organic chemicals, heterocyclic compounds, and amino acids or peptides, reflecting the pharmaceutical focus of interventional trials. On the disease side, neoplasms (cancers), nervous system diseases, and cardiovascular conditions dominate the trial landscape, highlighting areas of sustained medical research investment and clinical activity.

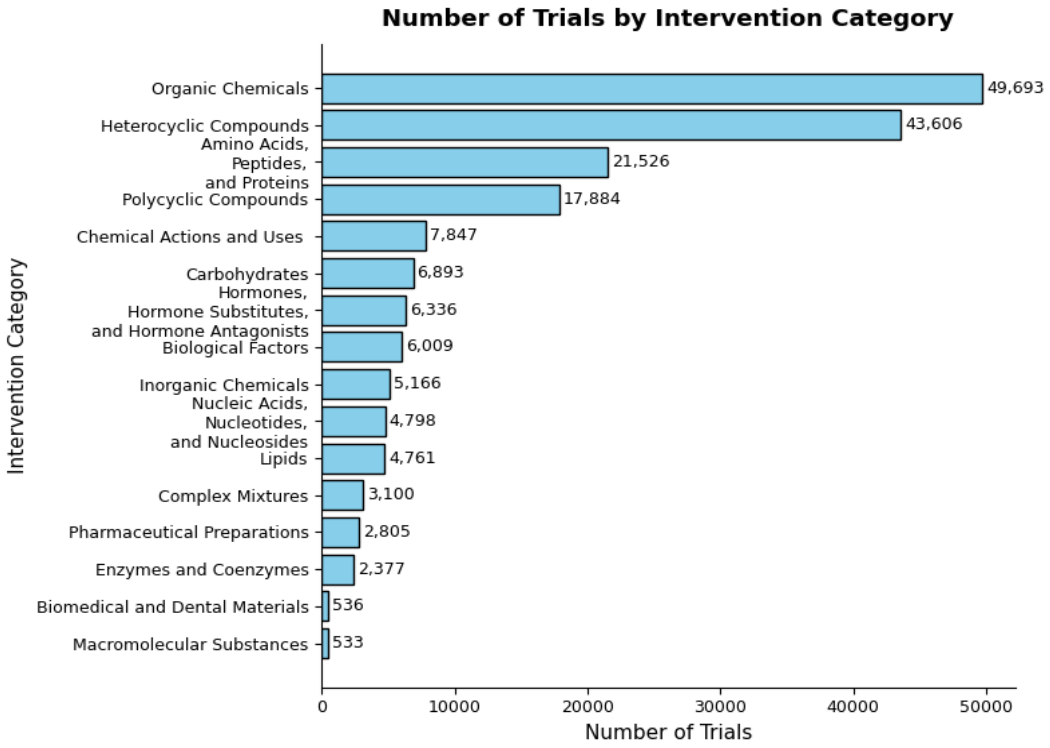

Figure 3: Number of Trials by Intervention Category

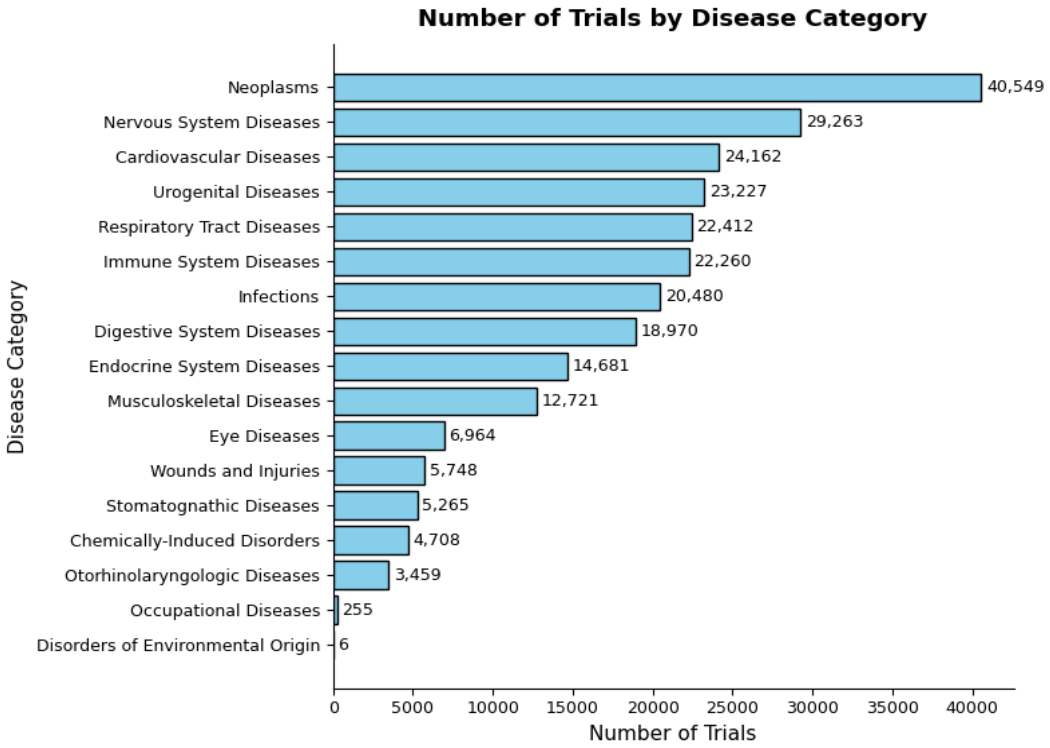

Figure 4: Number of Trials by Disease Category

3.2.3 Embeddings for trials descriptions

While structured features such as trial phase, sponsor type, and enrollment size provide valuable information for predictive modeling, they may not fully capture the complexity of clinical trial design, objectives, and context. Much of this nuanced information is often embedded in the detailed textual descriptions accompanying each trial, which elaborate on specific eligibility criteria, intervention details, study rationales, and anticipated outcomes. To leverage this rich source of information, we incorporated semantic representations of these descriptions into our modeling pipeline.

We experimented with two domain-specific language models: PubMedBERT and BioLinkBERT, both pre-trained on large-scale biomedical corpora. PubMedBERT is trained exclusively on PubMed abstracts, making it highly specialized for biomedical publication text. In contrast, BioLinkBERT is trained on a broader range of biomedical sources, including PubMed abstracts, full-text articles from PubMed Central, and additional clinical narratives, enabling it to capture a wider variety of biomedical linguistic patterns [18]. To select the more suitable model, we conducted an internal validation procedure: for a subset of trials with linked publications, we computed cosine similarities between embeddings generated from trial descriptions and the corresponding publication abstracts. This evaluation served to quantify each model's ability to semantically align trial descriptions with relevant scientific literature. The results showed that BioLinkBERT consistently achieved higher average similarity scores, indicating superior representation capacity for our task.

For each clinical trial, the chosen model generated a fixed-length embedding vector of 728 dimensions, representing the semantic content of the trial description. To ensure numerical stability and prevent the embeddings from disproportionately influencing the model due to their scale, we applied standardization to the embedding vectors using the StandardScaler. This step transformed each dimension to have a zero mean and unit variance, facilitating smoother model convergence and more balanced integration with other structured features during training.

3.2.4 Target Variable 1: Clinical Trial Completion Status

The first prediction task focuses on modeling the *completion status* of clinical trials. The `Overall Status` field in ClinicalTrials.gov provides a standardized classification of trial recruitment and follow-up status, based on the activity of all participating sites. To construct a binary classification target, we focused on trials that had reached a conclusive operational status by retaining only those labeled as `Completed`, `Terminated`, `Suspended`, or `Withdrawn`. Trials still in progress, actively recruiting, or with unclear status (e.g., `Not Yet Recruiting`, `Recruiting`, `Active but Not Recruiting`, `Enrolling by Invitation`) were excluded from the analysis.

For the binary target definition:

- Trials labeled as `Completed`—defined as trials that concluded all planned activities normally—were classified as **1 (completed)**.

- Trials labeled as `Terminated` (halted prematurely and will not resume), `Suspended` (halted temporarily but not yet resumed), or `Withdrawn` (stopped before enrollment) were classified as **0 (not completed)**.

This resulted in a dataset containing **263,448** interventional trials, of which **223,519** trials were labeled as completed and **39,929** trials were labeled as not completed. This outcome is directly related to the feasibility and operational success of trials, independent of their scientific findings, and serves as a key indicator for understanding trial performance patterns.

3.2.5 Target Variable 2: Clinical Trial Outcome Classification via Large Language Models

The second prediction task aims to determine the *outcome* of clinical trials based on the content of related scientific publications. Unlike the *Overall Status*, which reflects whether a trial was completed, this target focuses on whether the trial successfully met its pre-specified *primary endpoint* as documented in published results.

To maintain domain consistency and clinical relevance, this task focused exclusively on the subset of **109,361 cancer-related trials** within the ClinicalTrials.gov database. Cancer trials were selected because they are associated with a relatively rich body of published literature, clear endpoint reporting (e.g., overall survival, progression-free survival), and a higher likelihood of formal result dissemination, making them well-suited for outcome extraction tasks. Among these trials, **24,910** had associated publications recorded in PubMed.

We further refined the cohort to focus exclusively on trials whose intervention types involved either Biological or Drug therapies, excluding those involving devices, procedures, behavioral interventions, or other less standardized categories. This decision was grounded in several considerations. First, drug and biologic trials are subject to more rigorous regulatory standards (e.g., FDA or EMA oversight) and typically adhere to well-established protocols for efficacy assessment, including clearly defined primary endpoints and statistical evaluation criteria [19]. In contrast, trials involving devices or behavioral interventions often exhibit greater heterogeneity in both design and outcome reporting, which introduces additional noise and subjectivity into automated outcome classification. This filtering resulted in a final set of **16,738** trials for the outcome analysis.

To automatically classify trial outcomes based on the abstracts of associated publications, we employed **GPT-4o-mini**, a state-of-the-art large language model (LLM). Recognizing the critical importance of prompt engineering in optimizing LLM performance, we iteratively refined and validated the prompt designs. An additional challenge was the presence of multiple publications per trial, including interim analyses, subgroup analyses, or unrelated citations. To address this complexity, we implemented a **two-step classification pipeline**:

1. **Publication Type Identification**: The first step determined whether a publication abstract corresponded to a direct report of the trial's primary results. Abstracts not deemed primary

were excluded from further classification. The prompt used for this step was:

```
You are a clinical trial analyst.
Determine whether the following publication abstract describes the same
clinical trial as the one described below.
Only consider it a match if the interventions, patient population, and
study objective clearly align.
Respond with only one word: YES or NO.
```

2. **Outcome Classification**: For abstracts confirmed as direct publications, the model then classified the trial outcome using the following prompt:

```
You are given an abstract that reports results from a clinical trial.
Based on the abstract, label the trial outcome as:
- Positive: trial met its primary objective or showed favorable results
- Negative: trial did not meet its primary objective or results were
unfavorable
- Unknown: outcome is unclear or not explicitly reported
Respond using only one word: Positive, Negative, or Unknown.
```

This two-step procedure was designed to minimize misclassification arising from secondary or unrelated publications and to ensure that only primary outcome information was used for labeling. In total, GPT-4o-mini was asked to label 103,223 papers' abstracts, 84,917 abstracts are classified as 'Not a clinical trial's publication', 11,648 abstracts are classified as 'Positive' outcome, 4,000 abstracts are classified as 'Negative' outcome and 2,658 abstracts are classified as 'Outcome Unknown'.

To establish a performance benchmark, we manually labeled a random sample of 100 trial publications according to four possible categories: *Positive* (trial succeeded), *Negative* (trial failed), *Unknown* (outcome not clearly reported), and *Not a Trial's Publication*. Cross-validation results showed that **GPT-4o-mini** achieved an accuracy of **94%** against human annotations, indicating strong reliability in outcome labeling.

Finally, we aggregated multiple publication-level predictions into a single trial-level label using the following hierarchical rule:

- If **any** associated publication was classified as *Negative*, the trial was labeled as *Negative*.
- If no negative publications were found but **any** publication was classified as *Positive*, the trial was labeled as *Positive*.
- Trials with only *Unknown* or *Not a Publication* labels were excluded from further modeling.

This conservative approach prioritized capturing trial failures whenever evidence of negative results was present, thereby reducing optimistic bias in outcome classification.

After applying these rules, we identified a final cohort of **8,363 clinical trials** for downstream modeling, with **5,707** trials classified as having *positive outcomes* and **2,656** trials classified as *negative outcomes*.

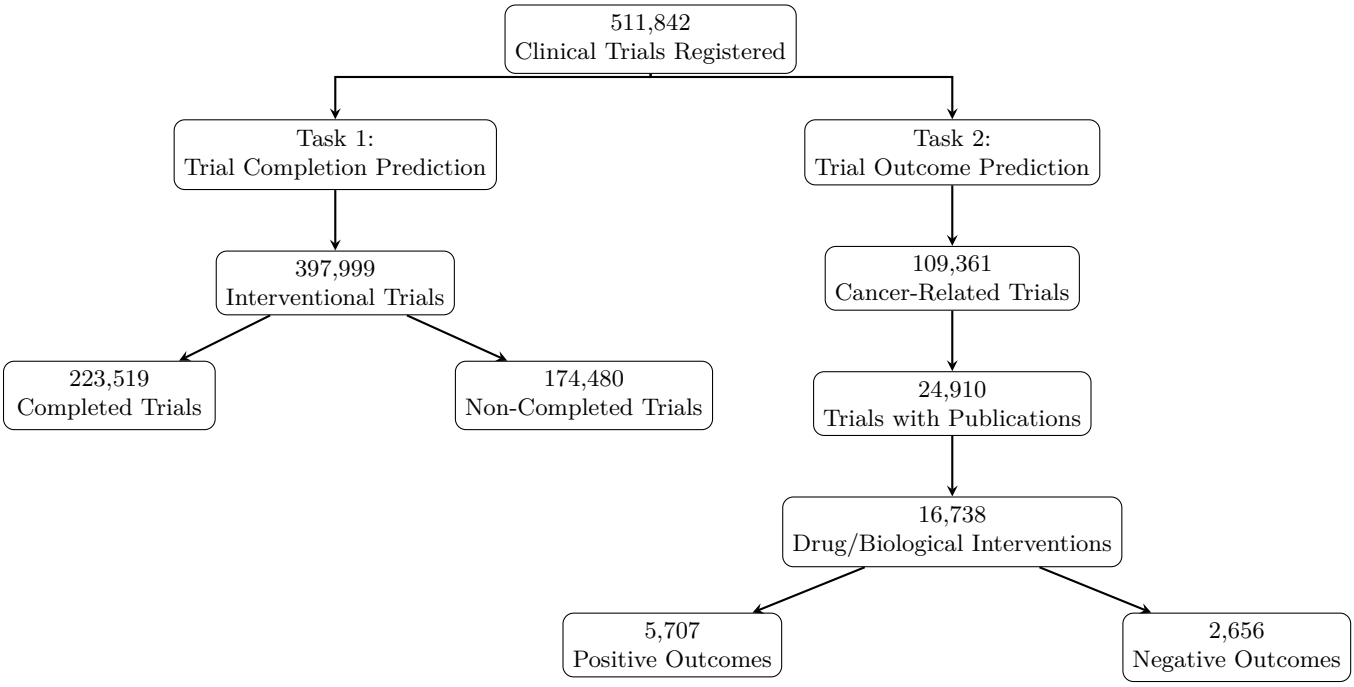

Figure 5: Data selection pipeline for clinical trial completion and outcome prediction tasks

3.2.6 Summarization for Feature Engineering

Table 2: Summary of Feature Engineering and Target Construction

Feature Type	**Preprocessing Method**	**Final Representation**
Numerical Variables	Mean imputation for missing values; standardization using StandardScaler	Scaled continuous variables
Categorical Variables	One-hot encoding (e.g., phase, sponsor type, sampling method, demographic categories)	Dummy variables
Location Variables	Count number of recruiting countries and sites	Numeric counts
Eligibility and Exclusion Criteria	Count number of eligibility/ineligibility criteria listed	Numeric counts
Disease Type (Condition)	Matched MeSH ancestor terms to 22 broad disease categories	Multi-label binary indicators
Intervention Type	Retrieved MeSH tree numbers and mapped to broad intervention classes	Multi-label binary indicators
Trial Descriptions (Textual)	Embedded using BioLinkBERT; vectors standardized using StandardScaler	728-dimensional dense vector
Target Variable 1 (Completion)	Labeled `Completed` as 1; `Terminated`, `Suspended`, or `Withdrawn` as 0	Binary label
Target Variable 2 (Outcome)	Labeled using GPT-4o-mini via a two-step prompt (publication identification + outcome classification)	Categorical label: Positive / Negative

4 Methods

4.1 Comparison Framework: Embedding-Enhanced vs. Tabular-Only Models

To assess the added value of unstructured textual information, we trained models under two distinct settings:

1. A baseline model trained solely on structured (tabular) data.
2. An embedding-enhanced model trained on a combination of structured features and BioLinkBERT-generated textual embeddings.

This comparative setup allows us to evaluate whether trial descriptions, encoded as semantic embeddings, contain predictive signals beyond conventional structured variables. Since trial descriptions often provide information not explicitly captured by structured fields (e.g., detailed design, intervention rationales, nuanced eligibility criteria), incorporating these embeddings may improve the model's ability to predict trial outcomes and completion status.

4.2 Dimensionality Reduction of Textual Embeddings

The BioLinkBERT embeddings generated for each trial description originally consisted of 728 dimensions. To address the high dimensionality and potential multicollinearity among embedding features, we applied **Principal Component Analysis (PCA)**, a widely used linear dimensionality reduction technique. PCA seeks to identify orthogonal directions (principal components) that capture the maximal variance in the data while reducing redundancy among variables.

We retained enough components to explain 95% of the total variance, resulting in a compressed yet informative representation of the textual embeddings. The reduced embedding vectors were then concatenated with the tabular features, forming the feature set for the embedding-enhanced model.

4.3 Feature Selection

To further reduce noise, improve model interpretability, and prevent overfitting, we performed **feature selection** on both the structured-only and embedding-enhanced datasets. Feature selection serves several purposes in predictive modeling: (i) it removes irrelevant or redundant features that may introduce noise and degrade model performance; (ii) it reduces the dimensionality of the feature space, thereby decreasing model complexity and training time; and (iii) it enhances generalization by preventing the model from fitting spurious patterns present only in the training data.

We adopted a model-based selection approach using an **XGBoost classifier** as the selector model. This procedure involved:

1. Fitting an XGBoost model on the training data.

2. Computing feature importance scores based on the trained model.
3. Retaining only features with importance scores exceeding the mean importance across all features.

This automatic selection process allowed us to filter out uninformative variables, focusing the model on features with genuine predictive power. Feature selection was performed separately on both datasets (with and without textual embeddings) to ensure that the reduced feature sets were tailored to the characteristics of each setting.

4.4 Handling Class Imbalance via Downsampling

Both prediction tasks exhibited class imbalance; completed trials significantly outnumbered non-completed trials, and successful trials (based on publications) outnumbered unsuccessful ones. To address this, we applied a **downsampling** technique to balance the datasets. Specifically, we randomly sampled the majority class to create datasets with class ratios of 1:1, 1:1.5, and 1:2 between the minority and majority classes.

Unlike oversampling methods (e.g., SMOTE), which synthetically generate new data points, downsampling avoids introducing potentially unrealistic samples that may not reflect real-world distributions. Since clinical trial data is highly structured and often limited in variability, generating synthetic samples could distort the underlying patterns. Downsampling, while reducing the size of the training data, preserves the integrity of real observations.

To enhance model robustness, we repeated the downsampling process across multiple iterations, each time selecting a different random subset of the majority class. This strategy mitigates the risk of overfitting to a specific sampled subset and promotes generalization by exposing the model to diverse realizations of the majority class.

4.5 Evaluation Metrics and Aggregation Strategy

The final model outputs were evaluated using four standard classification metrics: **accuracy**, **precision**, **recall**, and the **area under the ROC curve (AUC)**. To ensure robustness of the evaluation given the stochastic nature of the downsampling procedure, we performed the entire model training and evaluation process over **10 independent iterations** for each task and for both dataset configurations (structured-only vs. embedding-enhanced).

For each iteration, we randomly resampled the majority class and trained the XGBoost model under the specified downsampling ratio. The final reported results were obtained by computing the **average** of each metric across the 10 iterations. This aggregation helps to mitigate the variability introduced by random sampling and provides a more reliable estimate of the model's generalization performance. By adopting this strategy, we aim to avoid drawing conclusions based on any single random realization of the data.

4.6 Model Selection

We evaluated four classifiers for trial outcome prediction: Support Vector Machine (SVM), Random Forest, XGBoost, and a dual-tower neural network. This diverse set of models was selected to balance baseline benchmarking, interpretability, nonlinear modeling capability, and multi-modal integration.

SVM was included as a baseline due to its simplicity and robustness on high-dimensional, small-to-medium-scale datasets. By excluding embeddings and training solely on tabular features, it provided a reference point to assess whether more complex models and feature modalities offered meaningful improvements.

Random Forest and XGBoost were chosen for their strong empirical performance in structured healthcare data settings. These ensemble tree-based models can naturally handle feature interactions and nonlinear relationships while offering built-in feature importance measures, making them useful for both prediction and interpretability. In particular, XGBoost has demonstrated top-tier performance in many machine learning competitions and healthcare applications due to its gradient-boosting optimization and regularization capabilities[20].

Lastly, we explored the use of a dual-tower neural network to examine whether deep learning methods could effectively capture interactions between structured and unstructured features. It enables separate processing pipelines for tabular features and textual embeddings, which are then integrated into a shared latent representation. Given the increasing availability of unstructured trial descriptions, incorporating such models allows us to investigate their potential for capturing complex patterns that may be less accessible to tree-based or linear classifiers.

4.6.1 Support Vector Machine (SVM)

The SVM model served as a **baseline classifier** and was trained solely on structured (tabular) features. Given the relatively small feature space and the sensitivity of SVMs to high-dimensional inputs, embeddings were not incorporated for this model. The SVM was implemented with a linear kernel, probability estimates enabled, and class weighting set to "balanced" to compensate for class imbalance (SVC(kernel='linear', probability=True, class_weight='balanced', random_`state=42`)). No hyperparameter tuning was performed for the SVM, consistent with its role as a baseline.

4.6.2 Random Forest

Similarly, the Random Forest model was evaluated under both tabular-only and embedding-enhanced settings. To optimize model performance, a **grid search** was performed separately for tabular-only and embedding-enhanced models. Hyperparameters such as number of estimators, maximum depth and maximum features were systematically tuned based on 5-fold cross-validation performance. This separate tuning ensured that differences in performance were attributable to input features rather than suboptimal hyperparameters.

Three different imbalance-handling strategies were also explored:

1. Training on the original imbalanced dataset with `class_weight='balanced'`.
2. Training on a downsampled, fully balanced dataset (1:1) without class weighting.
3. Training on a partially downsampled dataset (1:2) with `class_weight='balanced'`.

4.6.3 XGBoost

The XGBoost model was evaluated under two input configurations: (i) structured (tabular) features only, and (ii) structured features concatenated with PCA-reduced textual embeddings. XGBoost was configured for binary classification using a logistic objective and log-loss evaluation metric (XGBClassifier(objective='binary:logistic', eval_metric='logloss', use_label_`encoder=False`)).

Grid Search was performed separately for tabular-only and embedding-enhanced models. Hyperparameters such as the number of estimators, maximum depth, learning rate, subsample ratio, and column sample ratio were systematically tuned based on 5-fold cross-validation performance.

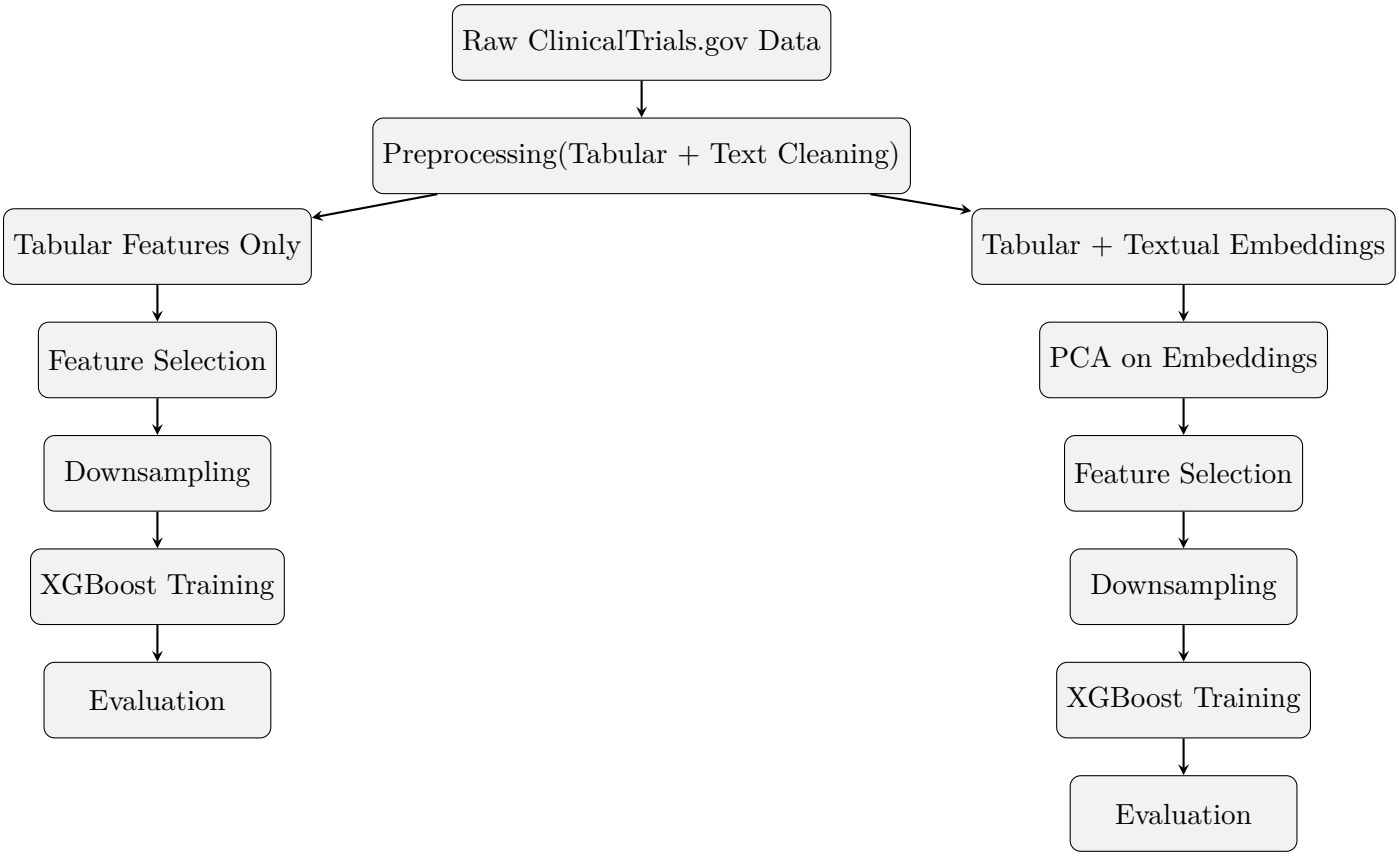

Figure 6: Modeling pipeline for tabular-only and embedding-enhanced models

4.6.4 Dual-Tower Neural Network with Embedding and Tabular Features

In addition to tree-based models, we implemented a neural network approach that integrates both structured and unstructured data using a **dual-tower architecture**. Neural networks, particularly multilayer perceptrons (MLPs), are well-suited for capturing complex, non-linear relationships among heterogeneous features, making them appropriate for modeling interactions between clinical trial characteristics and narrative descriptions.

To address class imbalance in the trial completion prediction task, we applied **random oversampling** via the `RandomOverSampler` technique. Oversampling ratios of 0.6, 0.8, and 1.0 were tested. Oversampling was preferred over undersampling to preserve majority class information and stabilize neural network training, which typically requires larger sample sizes for effective convergence and generalization.

For the input features, structured tabular data were preprocessed by converting boolean variables to numerical format and applying **MinMaxScaler** normalization. For unstructured data, BioLinkBERT-generated embeddings from trial descriptions were reduced via **Principal Component Analysis (PCA)**, retaining 95% of the variance, and further standardized via **StandardScaler**.

The dual-tower neural network consists of two subnetworks: a *tabular tower* and an *embedding tower*. The tabular tower processes structured inputs through a two-layer MLP (32 units $\rightarrow$ 64 units) with ReLU activations, batch normalization, and dropout (rate = 0.3). The embedding tower applies a parallel two-layer MLP (64 units $\rightarrow$ 64 units) with similar regularization. Outputs from both towers are concatenated into a 128-dimensional representation, followed by a classification head consisting of a hidden layer (64 units, ReLU, dropout) and a final output node for binary classification.

Model training was conducted using the **Adam optimizer** (learning rate = 0.001, weight decay = 10^{-4}) and **ReduceLROnPlateau** learning rate scheduling, with **early stopping** applied based on validation AUC (patience = 5 epochs). To optimize decision thresholds, we selected the **optimal threshold** that maximized the F1 score on the validation set rather than defaulting to 0.5, balancing precision and recall in the presence of residual class imbalance.

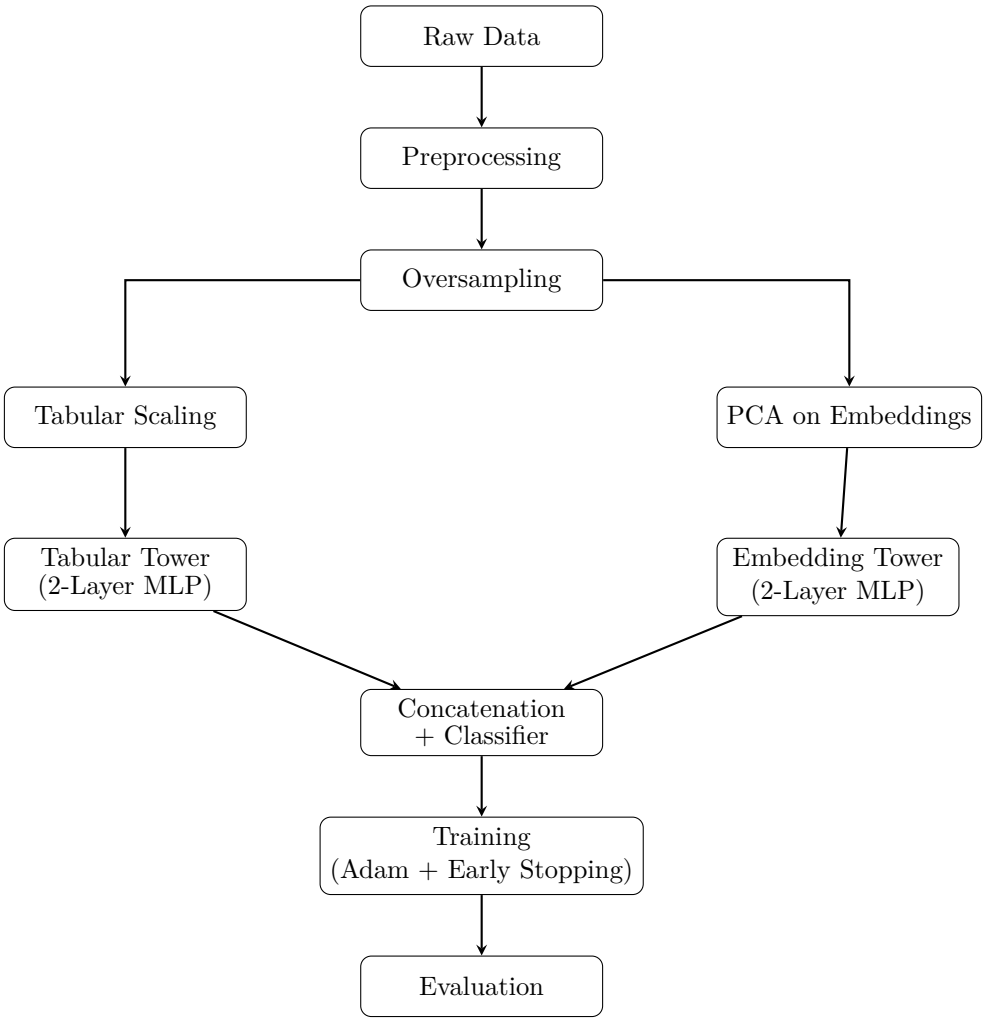

Figure 7: Detailed modeling pipeline for the Dual-Tower Neural Network (Tabular + Embedding) approach

4.6.5 Training and Evaluation

Each model was trained and evaluated over **10 independent iterations** to account for variability introduced by random downsampling. Performance metrics including accuracy, precision, recall, and AUC were averaged across the 10 iterations for each sampling configuration to ensure robust and reliable evaluation.

5 Results

5.1 Trial Completion

5.1.1 Baseline Model Comparison

To provide a comprehensive evaluation of different modeling strategies, we compared the performance of four classifiers: **Support Vector Machine (SVM)**, **Random Forest**, **XGBoost**, and a **dual-tower neural network**, all trained solely on the structured (tabular) features without incorporating textual embeddings. The results, summarized in Table 3, reflect the models' predictive performance averaged over multiple iterations on the downsampled dataset.

Among the models, **XGBoost** achieved the highest scores across all metrics, **Random Forest** also demonstrated competitive performance, though it lagged slightly behind XGBoost in precision and overall AUC. Training on a fully downsampled balanced dataset (1:1), without additional class weighting yielded the best results. This suggests that Random Forest models benefited more from a clean class balance through resampling rather than relying on internal weighting adjustments.**SVM** exhibited notably lower performance, suggesting that linear decision boundaries may not adequately capture the complexities of clinical trial features. The **dual-tower neural network**, despite its flexibility, achieved modest results, likely attributable to the relatively small tabular-only dataset and the higher capacity of deep models, which may have led to overfitting even with regularization strategies in place.

These findings justify the subsequent focus on **XGBoost** for deeper analysis, including the evaluation of textual embeddings and the impact of different class balancing ratios.

Model	AUC	Precision	Recall	Accuracy	F1 Score
Support Vector Machine (SVM)	0.7153	0.6737	0.5997	0.6539	0.6341
Random Forest	0.8669	0.7415	0.8734	0.7845	0.8021
XGBoost	**0.8830**	**0.8048**	**0.9288**	**0.8221**	**0.8623**
Neural Network (Dual-Tower)	0.7379	0.6668	0.5797	0.6400	0.6442

Table 3: Performance Comparison across Different Models (Tabular Features Only)

5.1.2 Detailed Evaluation of XGBoost Models

Following the overall model comparison, we conducted a more detailed analysis focusing on the **XGBoost** models, which achieved the best performance across all classifiers. We evaluated the XGBoost models under three different downsampling ratios (1:1, 1:1.5, and 1:2) and compared results between the *tabular-only model* and the *embedding-enhanced model*. On average, approximately 21–25 tabular features (from an initial 109 features) and 98–107 reduced embedding components (from the original 728-dimensional BioLinkBERT vectors) were retained across different sampling

ratios. All performance metrics were aggregated over **10 independent iterations** for each setting to ensure robustness, with results reported as mean ± standard deviation.

Table 4: XGBoost Performance under Different Downsampling Ratios (Mean ± Std Dev)

Metric	Class	Ratio	Tabular-only	With Embedding
AUC	-	1:1	0.8826 ± 0.0023	**0.8927** ± 0.0032
	-	1:1.5	0.8830 ± 0.0012	**0.8940** ± 0.0021
	-	1:2	0.8810 ± 0.0011	**0.8921** ± 0.0017
Precision	Class 0	1:1	0.852 ± 0.003	**0.852** ± 0.003
	Class 1	1:1	0.761 ± 0.003	**0.767** ± 0.003
	Class 0	1:1.5	0.862 ± 0.003	**0.871** ± 0.003
	Class 1	1:1.5	0.803 ± 0.002	**0.812** ± 0.002
	Class 0	1:2	0.862 ± 0.002	**0.873** ± 0.002
	Class 1	1:2	0.833 ± 0.002	**0.841** ± 0.002
Recall	Class 0	1:1	0.728 ± 0.003	**0.743** ± 0.003
	Class 1	1:1	0.871 ± 0.002	**0.872** ± 0.002
	Class 0	1:1.5	0.665 ± 0.010	**0.681** ± 0.013
	Class 1	1:1.5	0.932 ± 0.002	**0.932** ± 0.002
	Class 0	1:2	0.621 ± 0.012	**0.633** ± 0.002
	Class 1	1:2	0.952 ± 0.001	**0.954** ± 0.001
F1 Score	Class 0	1:1	0.778 ± 0.003	**0.791** ± 0.003
	Class 1	1:1	0.811 ± 0.002	**0.820** ± 0.002
	Class 0	1:1.5	0.754 ± 0.003	**0.761** ± 0.003
	Class 1	1:1.5	0.861 ± 0.002	**0.875** ± 0.002
	Class 0	1:2	0.724 ± 0.002	**0.732** ± 0.002
	Class 1	1:2	0.891 ± 0.001	**0.895** ± 0.001
Accuracy	-	1:1	0.7994 ± 0.0028	**0.8081** ± 0.0030
	-	1:1.5	0.8221 ± 0.0022	**0.8279** ± 0.0019
	-	1:2	0.8391 ± 0.0019	**0.8442** ± 0.0022

5.1.3 Interpretations

The experimental results collectively highlight the value of integrating textual embeddings alongside structured features for clinical trial completion prediction. Across all settings, the embedding-enhanced models consistently outperformed the tabular-only models, with improvements observed in AUC, accuracy, and especially in the precision, recall, and F1 scores for both completed (Class 1) and non-completed (Class 0) trials.

The incorporation of semantic embeddings notably strengthened the model's ability to correctly identify non-completed trials, a historically more challenging class. Embeddings captured nuanced trial design details—such as intervention complexities or eligibility criteria strictness—that were not explicitly available in the structured variables. This richer representation helped the model reduce

false positives without significantly sacrificing recall, thereby improving precision-driven metrics and yielding better balanced predictions overall.

Furthermore, the experiments demonstrated that the degree of downsampling had a meaningful impact on model behavior. Training under a 1:2 downsampling ratio maximized recall and F1 scores for completed trials but weakened detection of non-completed trials. In contrast, a 1:1 ratio improved Class 0 sensitivity at the expense of slightly lower overall discrimination. The 1:1.5 ratio provided a practical balance, maintaining high AUC and accuracy while balancing performance across classes. These findings suggest that moderately balancing the dataset, rather than fully equalizing or preserving original imbalance, may be optimal when combined with embedding augmentation.

Finally, although the absolute improvements in AUC and precision appeared modest (typically around +0.5% to +1%), such margins are significant in large clinical datasets, where even small gains can translate into a substantial number of better-classified trials. Overall, the results affirm that trial descriptions contain predictive signals beyond conventional metadata, and that combining structured and unstructured information enhances both the robustness and interpretability of clinical trial outcome prediction models.

5.1.4 Feature Importance and SHAP

The analysis of feature importance provides insights into the key factors influencing clinical trial completion predictions. To ensure interpretability and robustness, we examined both traditional feature importance scores from XGBoost in Figure 8and SHAP (SHapley Additive exPlanations) values in Figure 9, which offer a more nuanced, model-agnostic view of each feature's marginal contribution to the model output. SHAP values quantify the contribution of each feature to the model's prediction for individual instances, offering a unified and interpretable framework for understanding feature impact across complex models.

Across all downsampling ratios and modeling runs, enrollment size consistently emerged as the most influential predictor. The SHAP summary plot confirms this dominance, showing a strong positive relationship between higher enrollment counts and the likelihood of trial completion. This aligns with domain expectations—larger trials are often better funded, more visible to sponsors and regulators, and have stronger recruitment infrastructures, all of which enhance completion probability.

Trial duration also ranked among the top predictors in SHAP, showing a complex effect: longer durations were often associated with lower completion probabilities, possibly due to increased risk of logistical failure, patient attrition, or external disruptions over extended periods. Interestingly, this feature did not rank as highly in raw importance scores, underscoring SHAP's ability to detect subtle nonlinear relationships that might otherwise be underweighted.

Other consistently important features included:

- Trial design variables such as interventionModel_`PARALLEL` and interventionModel_`CROSSOVER`,

reflecting operational complexity. Parallel models generally showed positive contributions, while crossover models displayed more negative SHAP impacts—likely due to their logistical demands and higher dropout risks.

- `facility_count`, which showed positive SHAP values, indicating that multicenter trials with broader geographic reach may enjoy better recruitment and completion odds.
- Age eligibility criteria (eligibilityModule_`maximumAge`) and healthy volunteer status (healthyVolunteers_`yes/no`) were also informative. Trials enrolling older or more vulnerable populations may be more difficult to complete due to higher adverse event rates or recruitment challenges.
- Regulatory variables (FdaRegulatedDrug_`yes`, FdaRegulatedDevice_`no`) played a notable role. While FDA regulation ensures rigor, it may also introduce administrative burdens that affect trial feasibility.
- Disease area (`Neoplasms`) appeared as a moderately important predictor. Oncology trials are often complex and resource-intensive, possibly leading to a higher risk of early termination or delays.
- Organizational and sponsor characteristics, such as organization_class_`INDUSTRY` and collaborator_`NIH`, had visible SHAP contributions. Industry-sponsored trials typically benefit from clearer incentives and resource allocation, increasing their likelihood of completion.

Finally, the SHAP summary illustrates the value of interaction-aware explanations: features like `oversightHasDmc_yes`, `PHASE2`, and `PHASE3` may exert moderate average effects but become critical within specific contextual combinations (e.g., phase $\times$ sponsor type or masking strategy). This supports the idea that trial completion is a multifactorial outcome influenced by both additive and interaction effects.

In summary, the feature importance results are largely consistent with operational knowledge in clinical trial management. Factors related to trial scale, sponsor type, regulatory status, trial design, and disease area are among the most influential predictors of trial completion, providing validation for both the modeling approach and the interpretation of results. The SHAP-based interpretation reinforces the validity of our model's learned patterns. Features related to scale, complexity, regulatory oversight, and sponsor attributes dominate the completion landscape, and the SHAP framework provides transparent, interpretable reasoning behind model decisions. This further validates the modeling approach and highlights pathways for improving trial feasibility assessments in practice.

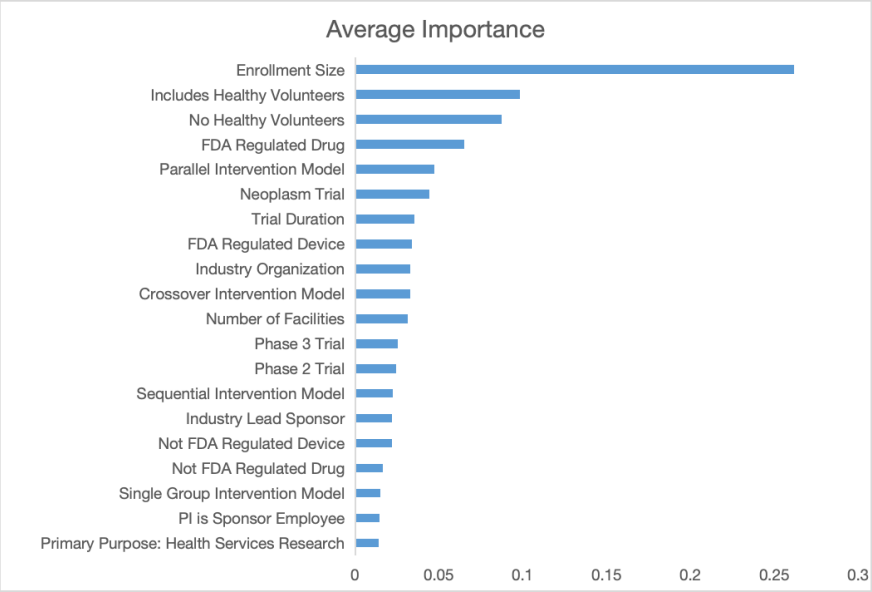

Figure 8: Top feature importances averaged across downsampling ratios.

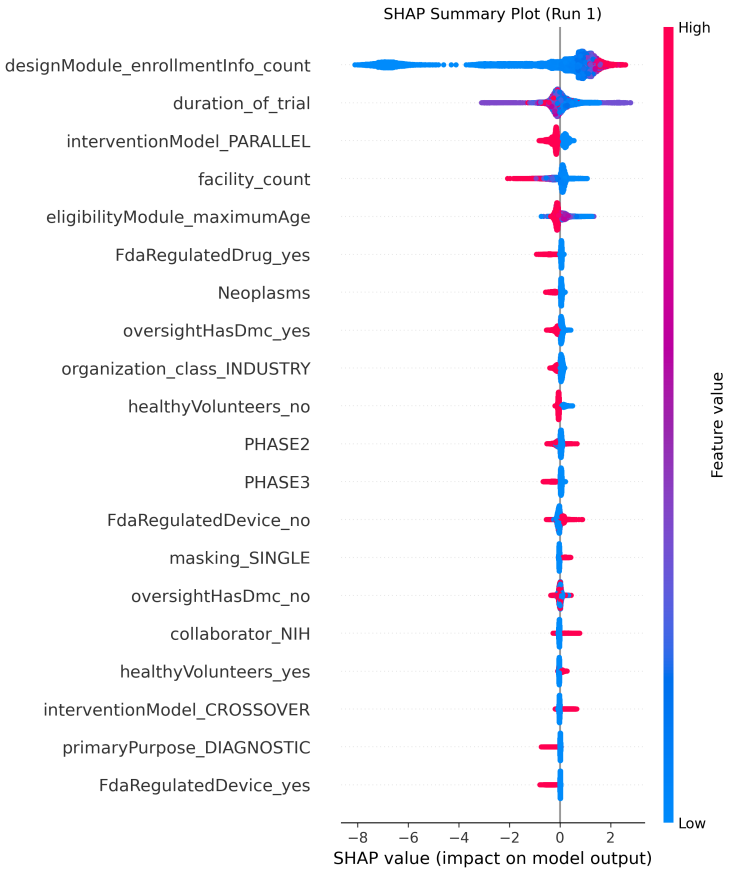

Figure 9: Top SHAP value for XGBoost

5.2 Trial Outcome Prediction

We report the results for the second prediction task: determining clinical trial outcomes (positive vs. negative) based on structured trial features. Unlike the completion prediction task, here we **did not incorporate textual embeddings** of trial descriptions into the modeling process. Preliminary experiments showed that the addition of embeddings did not improve model performance. This is likely because trial descriptions primarily detail study designs and eligibility criteria, which are more relevant to trial feasibility and completion rather than to the biological effectiveness of interventions or the likelihood of achieving a positive clinical outcome.

Three models were evaluated: Support Vector Machine (SVM), Random Forest, and XGBoost. All models were trained using tabular features after feature selection and class balancing procedures as described earlier.

The performance summary is shown in Table 5. Among the three models, XGBoost achieved the best overall performance, with an average AUC of 0.717, an accuracy of 66.1%, and an F1 score of 0.657. Random Forest followed closely, while SVM achieved comparatively lower scores across all metrics.

Table 5: Trial Outcome Prediction Performance (Structured Features Only)

Model	**AUC**	**Precision**	**Recall**	**Accuracy**	**F1 Score**
SVM	0.678	0.633	0.618	0.630	0.625
Random Forest	0.714	0.671	0.644	0.665	0.657
XGBoost	**0.717**	**0.663**	**0.651**	**0.661**	**0.687**

Overall, while the outcome prediction task proved to be more challenging than the trial completion prediction task (as reflected by generally lower AUCs and accuracies), tree-based models such as XGBoost and Random Forest still demonstrated reasonable predictive ability. These findings suggest that structured trial design characteristics, although less directly related to biological efficacy, still contain useful signals for anticipating the likelihood of clinical success.

To further investigate heterogeneity in prediction performance, we conducted a stratified analysis by clinical trial phase. Specifically, we trained and evaluated separate XGBoost models on trials categorized into Phase 1, Phase 2, and Phase 3. This stratification is motivated by the distinct objectives, designs, and success rates inherent to each trial phase, which can significantly influence predictive modeling. Phase 1 trials primarily focus on assessing safety, tolerability, and pharmacokinetics in a small cohort of healthy volunteers or patients. Due to their emphasis on safety rather than efficacy, these trials often report higher success rates, with approximately 63% transitioning to Phase 2. [21] The limited sample sizes and the nature of endpoints in Phase 1 trials may result in different predictive patterns compared to later phases. Phase 2 trials aim to evaluate preliminary efficacy and further assess safety in a larger patient population. This phase serves as a critical decision point in drug development, often referred to as the "valley of death,"

due to its historically low success rates—only about 31% of Phase 2 trials progress to Phase 3. The increased complexity and variability in trial designs at this stage can introduce additional challenges for predictive modeling. Phase 3 trials are large-scale studies designed to confirm efficacy, monitor side effects, and collect information that will allow the drug or treatment to be used safely. These trials typically have higher success rates than Phase 2, with approximately 58% advancing to regulatory submission. The rigorous design and larger sample sizes of Phase 3 trials may contribute to more stable and generalizable predictive models. Phase 4 trials were omitted from this analysis due to their limited sample size in our dataset and their distinct objective, which primarily focuses on post-marketing surveillance rather than initial efficacy evaluation By stratifying our analysis, we aim to account for these phase-specific differences, allowing for more tailored and potentially accurate predictive modeling. This approach acknowledges that a one-size-fits-all model may not be appropriate across all trial phases due to the inherent heterogeneity in trial objectives, designs, and success probabilities.

The distribution of labeled outcomes across phases was as follows:

- **Phase 1:** 1,763 Positive Outcomes vs. 422 Negative Outcomes
- **Phase 2:** 2,906 Positive Outcomes vs. 1,347 Negative Outcomes
- **Phase 3:** 1,227 Positive Outcomes vs. 913 Negative Outcomes

As expected, Phase 1 trials demonstrated a notably higher success rate, which reflects the typical design goals of early-stage studies focused on feasibility and safety rather than efficacy endpoints.

The model performance for each phase is summarized in Table 6. Across all phases, predictive performance remained modest, with AUCs ranging between 0.67 and 0.70. Interestingly, Phase 2 and Phase 3 trials achieved slightly higher AUCs compared to Phase 1, despite more balanced outcome distributions. This suggests that predicting efficacy-related outcomes remains a challenging task even with structured trial design features.

Despite the clinical and structural differences across trial phases, we did not observe substantial variation in predictive performance. One potential explanation is that the structured features used for modeling may not sufficiently capture the nuanced determinants of trial success that differ by phase. Furthermore, publication bias and inconsistent reporting practices across phases could also introduce noise, thereby limiting the model's ability to exploit any phase-specific signal. This suggests that additional features—such as biomarker usage, protocol deviations, or quality of endpoint definitions—might be necessary to better differentiate predictive patterns by phase.

Overall, these results highlight that while trial phase captures some variance in outcome prediction, structured design features alone provide limited predictive power for clinical trial success, particularly in later-phase studies where biological complexity and trial design variability are greater.

Table 6: Trial Outcome Prediction Performance by Phase (XGBoost Model)

Phase	AUC	Precision	Recall	Accuracy	F1 Score
Phase 1	0.669	0.627	0.622	0.609	0.628
Phase 2	0.691	0.636	0.628	0.634	0.630
Phase 3	0.698	0.635	0.632	0.630	0.643

6 Conclusion

This research investigated two critical prediction tasks in the clinical trial landscape: (1) forecasting the operational completion of trials and (2) predicting the scientific success of trials based on their published outcomes.

For the first task, we built and compared multiple machine learning models—including Support Vector Machine (SVM), Random Forest, XGBoost, and a dual-tower neural network—to predict whether a clinical trial would be completed or prematurely terminated. Across all models, XGBoost consistently achieved the best performance, demonstrating high AUC, accuracy, and F1 scores. Importantly, integrating semantic embeddings of trial descriptions (generated by BioLinkBERT and reduced via PCA) alongside structured tabular features significantly boosted model performance, particularly for detecting non-completed trials. These findings highlight that unstructured textual data contain valuable predictive signals beyond traditional trial metadata. Moderate downsampling ratios (e.g., 1:1.5) combined with embedding augmentation yielded the best overall results, striking a balance between sensitivity and specificity. Feature importance analysis further confirmed the operational relevance of key factors such as enrollment size, sponsor type, trial phase, and study design characteristics.

For the second task, we focused on predicting whether trials ultimately achieved their primary clinical endpoints, as inferred from scientific publications. To generate high-quality outcome labels at scale, we developed a two-step GPT-4o-mini-based classification pipeline that first identified primary publications and then classified trial outcomes as positive or negative. The LLM-based annotation approach achieved a 94% accuracy against human-labeled data, providing a reliable foundation for downstream modeling.

Unlike the completion prediction task, adding embeddings did not improve model performance in outcome prediction. This is likely because trial descriptions focus more on design logistics and eligibility criteria rather than biological efficacy signals. As a result, tabular features alone—such as phase, sponsor type, and intervention model—remained the primary drivers of predictive performance. Among the models tested, XGBoost again achieved the best overall performance, although the AUCs and accuracies were notably lower compared to the completion task, reflecting the greater inherent difficulty of predicting clinical success. Stratified analyses across Phase 1, 2, and 3 trials further illustrated phase-dependent variations in success rates and model performance, with Phase 1 trials exhibiting higher success rates due to their safety-focused design, while Phase 2 and 3 trials,

centered on efficacy testing, posed greater prediction challenges.

Overall, the results of this study demonstrate that combining structured clinical trial features with unstructured textual data (when appropriate) and leveraging modern large language models for scalable annotation can meaningfully enhance the prediction of clinical trial outcomes. While predicting operational feasibility (completion) appears more tractable with available metadata, predicting scientific success (outcome achievement) remains challenging, suggesting the need for richer biological and trial-specific information in future models.

By offering a modular, scalable, and interpretable framework for clinical trial outcome prediction, this research lays a foundation for future studies aiming to improve trial design, optimize resource allocation, and ultimately enhance the success rates of clinical research efforts.

7 Discussion

While this study presents a comprehensive framework for predicting clinical trial completion and outcome success, several limitations and areas for future improvement warrant discussion.

First, while using publication abstracts to infer clinical trial outcomes is a viable and scalable proxy, it remains an imperfect approach. Abstracts may omit nuanced or secondary outcomes, selectively emphasize positive findings, or lack sufficient detail about the success or failure of primary endpoints. Although our two-step LLM-based labeling pipeline mitigates some of these concerns by filtering for primary result publications, the approach ultimately depends on what is reported and how comprehensively it is conveyed. Additionally, while the use of the GPT-4o-mini model helped keep costs relatively manageable in this study, labeling full-text articles — or scaling the labeling process to even larger datasets — would substantially increase computational costs due to the significantly higher token counts involved. Furthermore, although cheaper alternatives such as LLaMA models or DeepSeek LLMs could reduce costs, they currently tend to underperform GPT-4-family models in specialized biomedical tasks, posing a trade-off between scalability and label quality. Future work could benefit from incorporating full-text article analysis, exploring semi-supervised learning to reduce reliance on labeled data, or triangulating outcomes with regulatory approval statuses and trial registries to improve coverage and accuracy.

Second, important predictors of trial success—such as the number of serious adverse events, mortality rates, or participant dropout rates—were unavailable in the dataset used. These clinical safety and feasibility signals are critical, as high adverse event rates or substantial dropout often predict both early trial termination and negative study outcomes. Although the AACT database offers more granular adverse event reporting, it lacks precise temporal resolution (i.e., whether adverse events are cumulative or assessed at a specific interim time point), limiting its utility. Future research could prioritize integrating more temporally-annotated datasets or trial sponsor reports to capture dynamic trial health indicators over time.

Third, interim analysis reports represent another vital information source. Interim results often

determine whether trials are continued, modified, or stopped early due to futility or overwhelming success. However, interim data are rarely publicly available and are typically confined to internal documents, Data Monitoring Committee (DMC) reviews, or high-level press releases. Without systematic access to interim analyses, early prediction of trial trajectory remains challenging. Collaborations with regulatory agencies or trial sponsors, where interim reports are mandated, could provide a pathway to overcoming this limitation in future research.

Fourth, publication coverage and quality gaps represent a significant structural limitation. Not all trials have corresponding PubMed-indexed publications; some results may be disseminated through alternative platforms such as conference abstracts, regulatory filings, clinical trial registries, or institutional repositories. Moreover, among the linked publications, the quality of association is often low: a large proportion of PubMed abstracts are not direct trial result publications but rather retrospective reviews, secondary analyses, or unrelated studies mentioning the trial identifier (NCT number). This suggests that the ClinicalTrials.gov linkage process may have been largely automated, likely relying on keyword or NCT number matching, without systematic human verification to ensure that the associated publications truly report the trial's primary outcomes. Consequently, publication missingness and misclassification are both likely non-random, with negative or inconclusive results underrepresented and genuine trial results inconsistently captured. This selection and linkage bias could skew model training and evaluation, adding more labor and spend into the verification. Future directions could involve mining broader corpora, including Google Scholar, Microsoft Academic Graph, or even targeted news media sentiment analysis, to capture a more complete and accurate picture of trial dissemination and outcomes.

Beyond these specific points, several broader challenges and opportunities remain. For instance, while this study treated trial descriptions as fixed textual inputs, advances in retrieval-augmented generation (RAG) could dynamically incorporate additional context such as intervention details, comparator arms, and evolving standards of care. Likewise, future models could employ multimodal learning strategies that combine textual, tabular, and imaging (e.g., radiology endpoints) data. Incorporating longitudinal modeling techniques—where trials are modeled dynamically over time rather than as static snapshots—could better mirror real-world decision-making processes.

Finally, while this work focused on binary classification tasks (completion vs. non-completion; success vs. failure), future work could explore ordinal or multi-class prediction frameworks, capturing the nuanced spectrum of trial outcomes such as early termination for futility, early success, or protocol amendments, thereby providing richer actionable insights for trial planning and investment decisions.

In summary, while the present framework represents a significant step toward more predictive and interpretable models of clinical trial trajectories, addressing these limitations and expanding the data sources, modeling strategies, and outcome granularity will be critical for advancing toward a truly comprehensive and practical trial prediction system.

Data and Code Availability Statement

GitHub repo: Master Thesis Repository
ClinicalTrials.gov API: API Documentation
BioLinkBERT: Hugging Face Model Page
PubMed API: Example Query
MeSH API: Lookup Endpoint

References

[1] Magdalena Zwierzyna, Mark Davies, Aroon D Hingorani, and Jackie Hunter. Clinical trial design and dissemination: comprehensive analysis of clinicaltrials.gov and PubMed data since 2005. BMJ, page k2130, June 2018.

[2] Gillian Gresham, Jill L. Meinert, Arthur G. Gresham, Steven Piantadosi, and Curtis L. Meinert. Update on the clinical trial landscape: analysis of ClinicalTrials.gov registration data, 2000–2020. Trials, 23(1):858, October 2022.

[3] Ellen Zhang and Steven G. DuBois. Early Termination of Oncology Clinical Trials in the United States. Cancer Medicine, 12(5):5517–5525, March 2023.

[4] Theodore R. Pak, Maria D. Rodriguez, and Frederick P. Roth. Why clinical trials are terminated, July 2015.

[5] Hafsa Habehh and Suril Gohel. Machine Learning in Healthcare. Current Genomics, 22(4):291–300, December 2021.

[6] Magdalyn E. Elkin and Xingquan Zhu. Predictive modeling of clinical trial terminations using feature engineering and embedding learning. Scientific Reports, 11(1):3446, February 2021.

[7] Ece Kavalci and Anthony Hartshorn. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. Scientific Reports, 13(1):121, January 2023.

[8] Simon Geletta, Lendie Follett, and Marcia Laugerman. Latent Dirichlet Allocation in predicting clinical trial terminations. BMC Medical Informatics and Decision Making, 19(1):242, December 2019.

[9] Siyang Wang, Simon Šuster, Timothy Baldwin, and Karin Verspoor. Predicting Publication of Clinical Trials Using Structured and Unstructured Data: Model Development and Validation Study. Journal of Medical Internet Research, 24(12):e38859, December 2022.

[10] Nicole White, Rex Parsons, David Borg, Gary Collins, and Adrian Barnett. Planned but ever published? A retrospective analysis of clinical prediction model studies registered on clinicaltrials.gov since 2000. Journal of Clinical Epidemiology, 173:111433, September 2024.

[11] C. W. Jones, L. Handler, K. E. Crowell, L. G. Keil, M. A. Weaver, and T. F. Platts-Mills. Non-publication of large randomized clinical trials: cross sectional analysis. BMJ, 347(oct28 9):f6104–f6104, October 2013.

[12] Chufan Gao, Jathurshan Pradeepkumar, Trisha Das, Shivashankar Thati, and Jimeng Sun. Automatically Labeling Clinical Trial Outcomes: A Large-Scale Benchmark for Drug Development, March 2025. arXiv:2406.10292 [cs].

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, September 2013. arXiv:1301.3781 [cs].

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs].

[15] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, November 2020. arXiv:1904.05342 [cs].

[16] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8003–8016. Association for Computational Linguistics, 2022.

[17] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David

Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. arXiv:2303.08774 [cs].

[18] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. LinkBERT: Pretraining Language Models with Document Links, March 2022. arXiv:2203.15827 [cs].

[19] U.S. Food and Drug Administration. Providing clinical evidence of effectiveness for human drug and biological products. Technical Report Clinical 6, Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), May 1998. Guidance for Industry.

[20] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, San Francisco California USA, August 2016. ACM.

[21] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimation of clinical trial success rates and related parameters. Biostatistics, 20(2):273–286, April 2019.