Deep Learning-based Prediction of Clinical Trial Enrollment with Uncertainty Estimates

Tien Huu Do[*]
Pfizer

Antoine Masquelier[*]
Pfizer

Nae Eoun Lee
Pfizer

Jonathan Crowther[†]
Merck

Abstract

Clinical trials are a systematic endeavor to assess the safety and efficacy of new drugs or treatments. Conducting such trials typically demands significant financial investment and meticulous planning, highlighting the need for accurate predictions of trial outcomes. Accurately predicting patient enrollment, a key factor in trial success, is one of the primary challenges during the planning phase. In this work, we propose a novel deep learning-based method to address this critical challenge. Our method, implemented as a neural network model, leverages pre-trained language models (PLMs) to capture the complexities and nuances of clinical documents, transforming them into expressive representations. These representations are then combined with encoded tabular features via an attention mechanism. To account for uncertainties in enrollment prediction, we enhance the model with a probabilistic layer based on the Gamma distribution, which enables range estimation. We apply the proposed model to predict clinical trial duration, assuming site-level enrollment follows a Poisson-Gamma process. We carry out extensive experiments on real-world clinical trial data, and show that the proposed method can effectively predict the number of patients enrolled at a number of sites for a given clinical trial, outperforming established baseline models.

1 Introduction

Drug development is a complex process with multiple stages, including drug discovery, pre-clinical research, clinical research, review, and post-market safety monitoring. Each of these stages may include multiple steps. For instance, the clinical research evaluates potential treatments in different phases, which are referred to as *clinical trials*. Depending on the disease that is being addressed, a clinical trial can span months or even years, involving human participants. Additionally, clinical trials may be conducted in various sites distributed in multiple countries. Given the complexity of clinical trials, substantial investments and efforts are required for their successful execution, rendering the necessity of careful design of the trials. Even with meticulous planning, many clinical trials have failed to produce anticipated clinical outcomes. Recent analyses on different clinical indications have revealed that the success rate of clinical trials is approximately 7.9%, which is a high risk for pharmaceutical companies [23]. One of the primary reasons for the failure of clinical trials is the insufficient enrollment of patients. Research shows that 19% of the trials are terminated due to insufficient enrollment [8] and 80% of the trials do not meet initial enrollment goals, resulting in significant financial loss of up to $8 million in revenue per day [20]. Therefore, it is of utmost

---

[*]Equal contribution
[†]Work performed while at Pfizer

importance to be able to predict the clinical trial success in terms of enrollment before the trials start. This work focuses on enrollment prediction at planning phase.

Many factors may affect the enrollment outcome of clinical trials such as therapeutic area, patient population, phase, length of treatment, and the geographical distribution of clinical trial sites. For instance, it is expected that a Vaccine trial has hundreds to thousands of patients while a rare disease trial often has much smaller scale (i.e., tens of patients). Additionally, multi-center trials recruiting patients in numerous countries are likely able to enroll more patients compared to single-center trials. In other words, clinical trial design characteristics provide early indicators of potential enrollment success and can be leveraged for predictive modeling.

Many approaches have been proposed for clinical trial enrollment prediction, and they usually fall into two categories: *deterministic* and *stochastic* approaches. The former addresses point estimation while the latter aims at giving an estimation with some levels of uncertainty. Early deterministic works rely on fixed enrollment rates derived from historical clinical trials [12]. However, given the non-linear dependencies of the enrollment process on a large number of influential factors, the prediction of enrollment using this approach does not align with reality in most cases. Recent deterministic works leverage advanced machine learning (ML) techniques, producing promising results [27, 6, 1, 37, 36]. The ML-based methods can model complex relationships and learn from large datasets, enabling more realistic and data-driven prediction thanks to the advance in ML algorithms and innovated hardware capabilities. However, most ML models still rely heavily on structured features and cannot fully capture the nuanced information contained in unstructured clinical text, such as inclusion and exclusion criteria. On the other hand, stochastic approach leverages statistical modeling by assuming the observed outcomes follow a single distribution or a mixture of distributions. Although theoretically sound, the stochastic approach often struggles with scalability due to the large volume and heterogeneity of clinical trial data [9, 10, 3, 2, 38].

Our work takes a further step by combining two approaches in a unified model: we propose a novel learning architecture capable of effectively predicting patient enrollment and accounting for prediction uncertainties. Our main contributions are as follows:

- We propose a novel model capable of effectively predicting patient enrollment for a given clinical trial design using both structured attributes and unstructured text. The model leverages pre-trained language models (PLMs) to encode textual information and integrates it with structured data using a multi-head attention mechanism, obtaining an expressive representation of the input trial. Subsequently, the representation is used for the downstream task of patient enrollment prediction.

- To address uncertainties, we have incorporated a probabilistic component into the proposed architecture through the use of the Gamma distribution. The model learns to predict the parameters of this distribution, and hence the prediction uncertainty can be measured with a confidence interval. By doing so, we demonstrate the flexibility of the proposed method in both settings of point and range estimation.

- We apply the proposed stochastic model to predict the duration of clinical trials, leveraging the Poisson-Gamma mixture process. We show that the proposed model can be used in different trial-related contexts.

- We empirically show the superior performance of the proposed model compared to strong baselines via extensive experiments and analyses on a large-scale dataset.

The rest of the paper is organized as follows. Section 2 positions our work concerning existing works. Section 3 describes the proposed method in detail. The experiments and discussions are presented in Section 4 and we draw our conclusions in Section 5.

2 Related Work

2.1 Enrollment Modeling

Given potential benefit of enrollment prediction, substantial efforts have been dedicated to the modeling of the patient enrollment. Existing works can be broadly categorized into two approaches: deterministic and stochastic. The deterministic approach focuses on point prediction, relying on fixed enrollment rates or ML algorithms. In contrast, the stochastic approach looks at the uncertainty of the

enrollment process and assumes that the enrollment follows some prior distributions. The prediction is then produced through simulation, such as the Monte Carlo method [16].

Early deterministic methods rely on the assumption of fixed enrollment rates (i.e., patient per site per month) and site initiation rates [10, 12] to predict trial duration. More precisely, these methods assume a linear or quadratic relationship between the patient enrollment duration and the total number of patients, enabling the possibility of calculating the enrollment duration based on closed-form solutions. While these method are straightforward to implement, relying on fixed rates seem to be a strong assumption. In addition, estimating the fixed rates from historical data is another challenge given the complexity of clinical trials and the lack of a reliable estimation method. Our work is different from these methods in that we focus on predicting the total number of patients instead of estimating the time needed for the enrollment. In addition, we do not rely on a pre-defined relationship. We instead learn the relationship between number of patients and the characteristics of input trial from a big dataset leveraging the power of advanced deep learning techniques.

The recent advancement in AI and machine learning has given the rise to a multitude of methods for patient enrollment modeling. In [27, 6], gradient boosting based models are used to predict number of enrolled patients and enrollment rate at trial-level. Different from these methods, in [1], enrollment rate prediction is made via a tensor factorization technique with promising results. However, the tensor factorization method is similar to gradient boosting methods in that they both require extensive feature engineering and hence they struggle to process unstructured clinical text, which plays a critical role in determining enrollment feasibility. Wang *et al.* [33] proposed *trial2vec*, based on BioBERT [24], to generate general-purpose embeddings for clinical trials, which can be used for various downstream tasks. Leveraging *trial2vec*, latent topics from complex clinical documents can be found for trial clustering. The found clusters are then used to formulate a sequence feeding a deep recurrent model for predicting trial enrollment outcome [34]. BioBERT is also used in [36, 37] to generate embeddings from eligibility criteria and trial descriptions for downstream predictive tasks. Notably, large language models (LLMs), models with emergent capabilities of generalization to diverse tasks, are leveraged in [36] to capture contextual information about trial drug candidates. Similar to these language model based methods, our method relies on deep learning and uses a pre-trained language model, namely Longformer [26], to generate clinical text embeddings. It is worth noting that we select Longformer owing to its longer context window compared to BERT-based models. In addition, we generate the embeddings on serialized text, i.e., the text created by concatenating multiple attributes of a clinical trial. By doing so, we are able to preserve representativeness and learn contextual embeddings tailored to the prediction task; this differentiates our method from prior approaches that aggregate sentence-level embeddings or simply truncate text. Lastly, we enhance our method by utilizing an extended set of features, and we use the multi-head attention mechanism to effectively combine multi-modal embeddings.

On the other hand, there exist patient enrollment modeling works that account for the uncertainties of recruitment process. Broadly speaking, methods of this approach assume that the enrollment of patients follows some random distributions, e.g., Poisson distribution. The parameters of these distributions are then found using historical data, and the enrollment scenarios are generated through simulation, e.g., Monte Carlo simulation. In [9], the arrivals of patients within a trial are modeled with a Poisson process. The empirical distribution of trial duration is then found via simulation. While this method is simple and capable of giving a confidence level, it requires a fixed enrollment rate estimated from historical data. However, there are no evidences on how accurate the rate is. Extending this method, a Poisson-Gamma process is used to model the patient enrollment in [2, 3, 4], where the site-level enrollment rates are samples from a Gamma distribution. More recently, a Poisson regression model is used to predict the enrollment duration and number of patients at site and trial levels [38]. Although results are promising, the method relies on a linear relationship between therapeutic area, trial duration, and some other randomness effects, which might be suboptimal given the complexity of clinical trials. Similar to these methods, our approach models the uncertainty of enrollment outcomes using a Gamma distribution. Nevertheless, we learn the parameters of the distribution directly from data leveraging an advanced deep learning algorithm, offering both scalability and generalization across diverse trial types.

2.2 Large Language Models

Large language models (LLMs) refer to state-of-the-art models with the capabilities of dealing with complex language understanding tasks such as question answering, machine translation, summarization, and text generation. LLMs typically possess an extensive number of parameters and necessitate huge computational resources for training and inference. The popularity of LLMs can be dated back to the introduction of GPT-3 [7] in 2022 when end-users were able to interact with LLMs via a chat interface. Since then, many LLMs including LLama [14], Gemini [30], and Claude [22] have been introduced with emergent capabilities of solving diverse reasoning tasks and processing multi-modality data.

In the medical domain, given the complexity of clinical documents, there is an utmost need of understanding these documents for downstream tasks. There exist approaches that rely on early pre-trained language models (PLMs) such as BioBERT [24] and ClinicalBERT [19] for medical text understanding. In these works, BERT [13] is adapted to specific domains, namely clinical or biomedical domains. Similarly, our work leverages the expressiveness power of PLMs, namely the Longformer model [26] to obtain domain-specific embeddings. However, we do not rely on fine-tuning directly Longformer for downstream tasks. Instead, the embeddings generated by the model are fed into a deep neural network specially designed for patient enrollment prediction. Recent works have leveraged emergent abilities of advanced LLMs for clinical text understanding [29, 25]. While these works have shown positive results on some tasks such as classification and summarization, little attention has been paid to regression task. In this work, we employ LLM-based fine-tuning as a baseline using Llama2 [31] and show empirically that the proposed method has superior performance compared the LLM fine-tuning based approach on enrollment prediction task.

3 Method

3.1 Problem Statement

Denote a clinical trial by C_i, our goal is to predict the total number of patients enrolled to the trial $y_i \in \mathbb{N}$. This prediction task can be formalized as a supervised regression problem

$$\hat{y}_i = \mathcal{F}(C_i; \Theta), \tag{1}$$

where $\mathcal{F}$ is the model parameterized by Θ; $\mathcal{F}$ can be thought of as a deterministic function. One might expect the variation of the enrollment given that clinical trials are often complex, involving many factors and random events as discussed in Section 1. Therefore, sometimes it is desirable to account for uncertainties with a certain level of confidence instead of point estimation. Towards this goal, we can modify model $\mathcal{F}$ to yield a distribution of the number of patients:

$$\Phi_i = \mathcal{F}(C_i; \Theta) \tag{2}$$
$$Y_i \sim \mathcal{D}(\Phi_i), \tag{3}$$

where Y_i is a random variable representing number of patients and $\mathcal{D}(\Phi_i)$ is a certain distribution parameterized by Φ_i.

3.2 Deterministic Modeling

Our first approach is based on a deep neural network model, specially designed for the enrollment regression problem. This approach is referred to as *deterministic modeling* since the proposed model produces a single value given one input trial.

As discussed in Section 1, a clinical trial is generally complex, and has multiple attributes of miscellaneous types (see Table 1). In order to handle these attributes, similar to [33], we first group the set of trial attributes into two subsets, namely *Key* and *Context*. The *Key* set contains important attributes that allow the positioning of the trial of interest among reference clinical trials. These attributes are: phase, country, therapeutic area (TA), sponsor, planned number of participants, and planned number of sites. We then apply pre-processing operations to these attributes using MultiLabelBinarizer[3] and standard transformation (a.k.a. Z-score) to obtain two embeddings $\mathbf{x}^{cat} \in$

---

[3]`https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html`

$\mathbb{R}^{d_1}$ and $\mathbf{x}^{num} \in \mathbb{R}^{d_2}$. Conversely, the *Context* set is a mixture of textual and categorical attributes such as title, objective, mechanism of action, indication (disease), inclusion, and exclusion criteria. To effectively process this set, we first apply a simple text serialization technique [15] to form a unified text from attributes, that is we concatenate these attributes using a separator. We then generate an embedding $\mathbf{x}^{emb} \in \mathbb{R}^{d_3}$ for the unified text using *Clinical Longformer* [26]. We select Clinical Longformer because it has the context length of 4096 tokens, which eliminates the issue of context being truncated given that the concatenated text may have thousands of tokens. More importantly, this model has been fine-tuned on clinical corpora, which is capable of capturing nuances in clinical text.
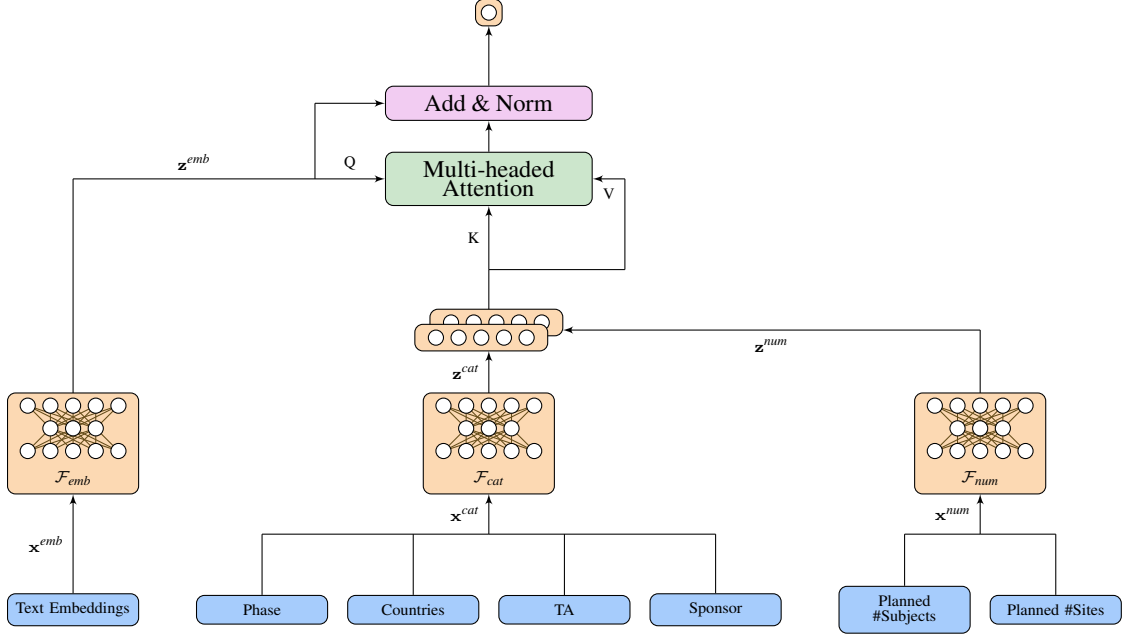

Figure 1: Deterministic model architecture for enrollment prediction. Two sets of attributes are *Key* and *Context*, where the latter is encoded into text *embeddings* using Clinical Longformer. $\mathcal{F}_{emb}, \mathcal{F}_{cat}$, and $\mathcal{F}_{num}$ are fully connected layers.

Our model is presented in Figure 1, where it takes as inputs three embeddings $\mathbf{x}^{emb}$, $\mathbf{x}^{cat}$, and $\mathbf{x}^{num}$. Using several fully connected (FC) layers with activations, we obtain intermediate representations:

$$\mathbf{z}^{emb} = \mathcal{F}_{emb}(\mathbf{x}^{emb}; \Theta_1), \tag{4a}$$

$$\mathbf{z}^{cat} = \mathcal{F}_{cat}(\mathbf{x}^{cat}; \Theta_2), \tag{4b}$$

$$\mathbf{z}^{num} = \mathcal{F}_{num}(\mathbf{x}^{num}; \Theta_3), \tag{4c}$$

where $\mathbf{z}^{emb}, \mathbf{z}^{cat}, \mathbf{z}^{num} \in \mathbb{R}^D$. These vectors are then combined using the multi-head attention mechanism [35] where

$$Q = \text{unsqueeze}_1(\mathbf{z}^{emb}) \tag{5}$$

$$K = V = (\mathbf{z}^{cat}, \mathbf{z}^{num}) \tag{6}$$

with $\text{unsqueeze}_1 : \mathbb{R}^{B \times 32} \to \mathbb{R}^{B \times 1 \times d_3}$ and B being the batch size.

This operation yields $\mathbf{z}^{att} = \text{MultiHead}(Q, K, V) \in \mathbb{R}^D$. In the end, we add a skip connection [17] and a layer-norm operation [5] to produce the final representation:

$$\mathbf{h} = \text{LayerNorm}(\mathbf{z}^{att} + \mathbf{z}^{emb}) \tag{7}$$

This final embedding $\mathbf{h}$ is passed to a regression head to produce the predicted number of patients.
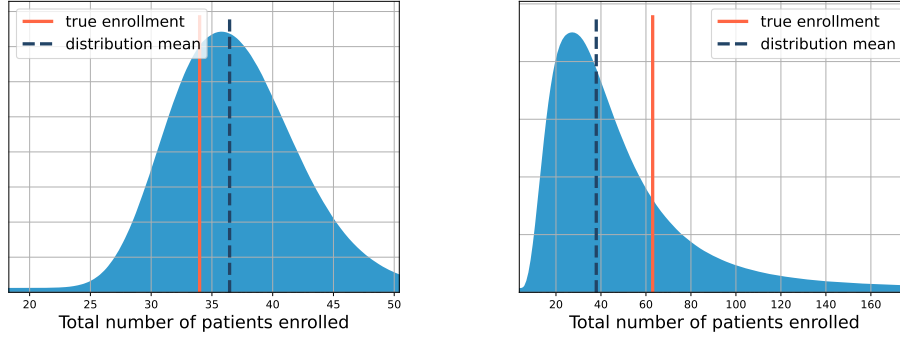
Figure 2: Examples of output of the stochastic model for two different studies. The dashed line represents the mean of the output distribution and the orange line represents the true total number of patients enrolled. The horizontal axis represents the number of patients enrolled and the vertical axis represents its probability density.

We train the proposed model by minimizing a L_1 loss, defined by:

$$\mathcal{L}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} |\ln y_i - \ln \hat{y}_i|, \tag{8}$$

where n is the number of trials in our dataset. It is worth noting that $\mathcal{L}(\Theta)$ is calculated on logarithmic scale since we found empirically that utilizing the logarithmic scale loss function brings better prediction performance.

3.3 Stochastic Modeling

It is desirable to design clinical trials with predictable outcomes, e.g., in terms of patient enrollment. However, clinical trials in reality often include some sorts of uncertainty due to factors like patient availability, site activation delays, and external disruptions. In order to account for uncertainties, we extend the proposed deterministic model in Section 3.2 by predicting the distribution of the number of enrolled patients instead of a single number. More precisely, the model predicts the parameters of the distribution, which in our case is chosen to be a Gamma distribution due to its flexibility to fit a wide variety of shapes for right-skewed data. This way, the model can not only predict the number of participants via sampling but it can also give a range estimation with a confidence level. Examples of such probabilistic outputs are shown in Figure 2.

The stochastic model, depicted in Figure 3, employs the same backbone as the architecture outlined in Section 3.2, but its final layer branches into two outputs that parametrize the Gamma distribution. The logits are passed through an exponential function to ensure that the shape α and rate λ are strictly positive

$$\alpha = \exp(\text{FC}_\alpha(\mathbf{h})) \tag{9}$$
$$\lambda = \exp(\text{FC}_\lambda(\mathbf{h})), \tag{10}$$

where FC refers to a fully connected layer with a Leaky ReLU function applied at the end of each layer.

Since the PDF of the distribution is known, the loss can be defined by:

$$\mathcal{L}(\Theta) = -\frac{1}{n} \sum_{i=1}^{n} \ln p(\ln \hat{y}_i \mid C_i) \tag{11}$$

3.4 Deep Poisson-Gamma

The stochastic approach described in Section 3.3 is more challenging to train due to its variable nature. However, its ability to capture variability can be beneficial in cases where a same study design leads
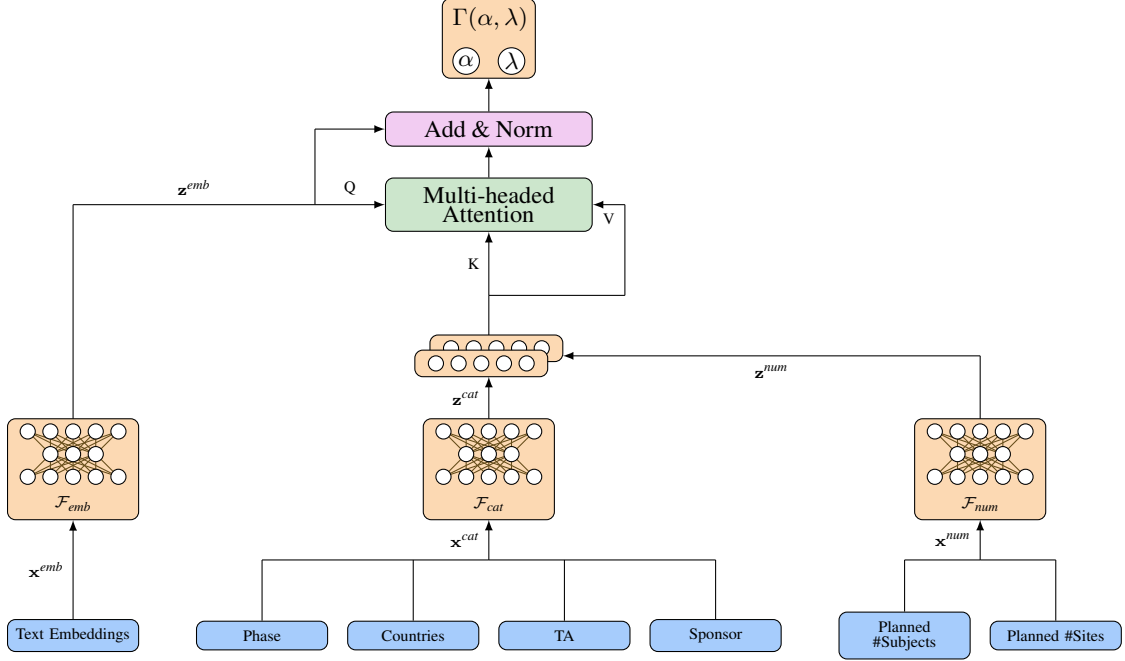
6

Figure 3: Stochastic model architecture for patient enrollment prediction. This model shares the architecture with the deterministic model except the last layer where two parameters of the Gamma distribution are predicted.

to a wide variety of outcomes. Let us consider the Poisson-Gamma enrollment forecasting framework proposed by Anisimov [3], which models the enrollment of a trial as a mixture of Poisson processes, where each Poisson process represents the enrollment of a site. The rates of these Poisson processes follow a Gamma distribution $\Gamma_\mu(\alpha, \lambda)$. Additionally, for each site, the Poisson process starts at a random startup time following a distribution $\Gamma_\theta(\alpha, \lambda)$. In this particular use case, a same input leads to a very wide range of rates and startup times combinations as we are predicting site-level output based on trial-level information. The work suggests that, at planning stage, the parameters α and λ of the enrollment rates can be evaluated by using historical data from similar trials and that the startup times could be modeled as a random variable. Inspired by Anisimov's work, we explore two different approaches to evaluate the parameters of the Poisson-Gamma model that are used to predict the enrollment *duration* of a trial at planning phase based on its attributes.

We model the enrollment process of a trial C_i with a set of sites S_i as a mixture of Poisson processes as described in Anisimov [3]. The recruitment $\rho_{i,s,t}$ of each site $s \in S_i$ at time t is modeled as a simple Poisson process with a rate $\mu_{i,s}$ and a startup time $\theta_{i,s}$. The *Target enrollment* from C_i is also used in the modeling and will be denoted as π_i.

The model is defined by

$$\rho_{i,s,t} = \begin{cases} \sim Poisson(\mu_{i,s}) & \text{if } t \geq \theta_{i,s} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$R_{i,T} = \sum_{t=0}^{T} \sum_{s \in S_i} \rho_{i,s,t} \tag{13}$$

$$\hat{\tau}_i = min\{t \in \mathbb{Z} \mid R_{i,t} \geq \pi_i\}, \tag{14}$$

where $\rho_{i,s,t}$ is the number of enrolled patients of a site s at time t, $R_{i,T}$ the total number of patients for trial C_i at time T. The predicted enrollment duration of a trial $\hat{\tau}_i$ is defined as the first time step at which the total enrollment $R_{i,t}$ reaches the target enrollment π_i.

For the first approach, we train a slightly different version of our stochastic model from Section 3.3. The outputs are now the parameters of the distributions of the enrollment rate and the startup time of a site $\Gamma_{\mu,i}$ and $\Gamma_{\theta,i}$. These distributions are then described by the following equations:

$$((\alpha_{\mu,i}, \lambda_{\mu,i}), (\alpha_{\theta,i}, \lambda_{\theta,i})) = \mathcal{F}_{(\mu,\theta)}(C_i; \Theta) \tag{15}$$

$$\Gamma_{\mu,i} \equiv \Gamma(\alpha_{\mu,i}, \lambda_{\mu,i}) \tag{16}$$

$$\Gamma_{\theta,i} \equiv \Gamma(\alpha_{\theta,i}, \lambda_{\theta,i}) \tag{17}$$

The target outputs in our model are vectors of enrollment rates and startup times from sites S_i in their original scale. The architecture used is the same as the one described in Section 3.3. However, due to the different nature and scale of the target output, the loss function as described in equations (19) have been modified accordingly.

$$\mathcal{L}_\mu(\Theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{|S_i|} \sum_{i,s}^{S_i} \ln p(\hat{\mu}_{i,s} \mid C_i) \tag{18}$$

$$\mathcal{L}_\theta(\Theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{|S_i|} \sum_{i,s}^{S_i} \ln p(\hat{\theta}_{i,s} \mid C_i) \tag{19}$$

$$\mathcal{L}(\Theta) = \mathcal{L}_\mu(\Theta) + \mathcal{L}_\theta(\Theta) \tag{20}$$

For the second approach, we implement a filtering-and-fitting baseline inspired by Anisimov [3] and Zhong *et al.* [38]. The method identifies, for each trial C_i, historical trials with similar characteristics (based on the categorical features listed in Table 1) and fits Gamma distributions $\Gamma_{\mu,i}$ and $\Gamma_{\theta,i}$ to, respectively, their historical site-level enrollment rates and startup times using maximum likelihood estimation (MLE).

Let the feature set $\mathcal{F}$ include attributes such as phase, country, and therapeutic area. Each feature $f \in \mathcal{F}$ can have multiple values for a given trial. Let f_i be the set of values for feature f in trial C_i. We consider trial C_d to be part of the set of similar trials if:

$$\forall f : f_d \cap f_i \neq \emptyset \tag{21}$$

This ensures that the historical trials used to fit the distributions $\Gamma_{\mu,i}$ and $\Gamma_{\theta,i}$ are sufficiently similar in design.

4 Experiments

4.1 Dataset

We collect trial-level data from two primary sources, namely IQVIA Data Query System (DQS[4]) and Citeline[5]. DQS contains proprietary operational data about clinical trials, principal investigators, and facilities. The data in DQS are provided by pharmaceutical companies or collected from shared data collaborations such as Investigator Registry or the Investigator Databank. Thanks to its volume and high quality, DQS enables sponsors and contract research organizations (CROs) to find the best matched investigators and facilities, and hence effectively support clinical trial planning. Data from Citeline come from more than 40k sources[6] of monitored information, and are curated by their

---

[4]https://www.iqvia.com/solutions/technologies/orchestrated-clinical-trials/
planning-suite/data-query-system
[5]https://www.citeline.com/en
[6]https://www.informa.com/globalassets/documents/investor-relations/2019/
investorday/citeline.pdf

Table 1: The list of attributes and label in our dataset collected from Citeline and DQS. The target enrollment is the planned number of patients set by clinical trial managers. The number of enrolled patients (No. enrolled) is the actual number of patients when the corresponding trial completes, which is our prediction target. It should be noted that the target enrollment number in some cases is close to the actual enrolled number but in many cases there are wide discrepancies between those numbers. " - " stands for attributes that are not applicable to the corresponding columns and TA stands for Therapeutic Area. Categorical attributes may be single-valued or multi-valued; those marked with a dagger are multi-valued. Attribute *Phase* has eight different values because some trials are multi-phase (e.g., II/III).

Attribute	Source	Type	Min	Median	Mean	Max	Cardinality
Title	Citeline	Text	-	-	-	-	-
Objective	Citeline	Text	-	-	-	-	-
TA	Citeline	Categorical	-	-	-	-	9
Indication[†]	Citeline	Categorical	-	-	-	-	970
Mechanism[†]	Citeline	Categorical	-	-	-	-	995
Sponsor[†]	Citeline	Categorical	-	-	-	-	962
Sponsor Type[†]	Citeline	Categorical	-	-	-	-	10
Country[†]	Citeline	Categorical	-	-	-	-	147
Target enrollment	Citeline	Numeric	3	126	393.03	84965	-
Drug name	Citeline	Categorical	-	-	-	-	3418
Inclusion	Citeline	Text	-	-	-	-	-
Exclusion	Citeline	Text	-	-	-	-	-
Phase[†]	DQS	Categorical	-	-	-	-	8
No. site	DQS	Numeric	1	15	42.56	1140	-
No. enrolled*	DQS	Numeric	1	85	292.24	11719	-

scientists. Similar to DQS, Citeline data contain information about trials, investigators, and facilities. Citeline data lack the detailed site-level performance metrics that DQS offers, however, Citeline provides extra features and metrics at trial and drug levels. Hence, we collect the information about number of sites, trial duration, and number of enrolled patients from DQS while other structured and textual attributes are taken from Citeline. Another data source is *ClinicalTrials.gov*, which is a publicly available database established by National Institute of Health (NIH). It should be noted that a large portion of Citeline trial data can be found in this database.

Our dataset contains more than 11.4k clinical trials of different therapeutic areas such as Oncology, Inflammation, and Cardiovascular. Since our method is based on supervised learning, we only consider trials with known outcome. For this reason, the status of a trial in our dataset must be either *Completed* or *Closed*. The number of enrolled patients may vary greatly across trials depending on the nature of therapeutic areas, ranging between 1 and 11.7k. A comprehensive list of attributes along with their corresponding summary descriptions is presented in Table 1.

For the Deep Poisson-Gamma modeling use case described in Section 3.4, we collect the trial-site-level data from IQVIA DQS, which is a more granular version of the data presented previously. There are three additional attributes needed for this use case compared to Section 4.1, namely, the *total number of patients enrolled at trial-site level*, *trial duration*, and *enrollment rate on trial-site level*. The *enrollment duration* is derived from the *trial start date* and *trial completion date* attributes from DQS, while the *enrollment rate on trial-site level* calculated by taking the ratio between the *the total number of patients enrolled at trial-site level* and the difference between the *trial duration* and the *site startup-time*.

4.2 Experimental Settings

4.2.1 Patient Enrollment Prediction Setting

In line with standard procedures, we split the data into three subsets, the training set (9410 trials), the development set (1000 trials), and the test set(1000 trials). The dataset is split in such a way that the distributions of number of enrolled patients in these training, development and test sets

Table 2: Trial selection criteria for the Poisson-Gamma modeling.

Attribute	Value
Trial completion date	Between 2015 and 2025
Number of sites	More than 10 sites
Trial duration	Between 6 and 36 months
Trial phase	II and III

are similar. We then optimize the model parameters using the training set and select the model exhibiting the best performance on the development set. Subsequently, the evaluation metrics are calculated on the test set to obtain an unbiased assessment of the proposed model. We utilize two widely used metrics for measuring regression performance: the Mean Absolute Error (MAE) and the Coefficient of Determination (R^2). R^2 is popular metric for regression, however, its reliability can be compromised in case the data contain outliers. Hence, it is beneficial to complement R^2 with MAE to quantify the average magnitude of errors. It is important to note that although the model is trained on the logarithmic scale (see Section 3), MAE and R^2 are calculated on the original scale (i.e., number of participants) of data to maintain interpretability.

We employ both classical machine learning models and advanced language models (PLMs/LLMs) as baselines for performance comparison. For classical models, we use the established gradient boosting models, including XGBoost [11] and LightGBM [21]. These models have been known for very high performance on tabular datasets. In order to use these models, encoded attributes are concatenated to create a unified input embedding. For PLMs, we fine-tune BioBERT [24], ClinicalBERT [32], and Clinical Longformer [26] for regression task; these models take as input the serialized text as discussed in Section 3.2. For LLMs, we employ the fine-tuning approach, leveraging Llama2 [31]. Similar to Longformer, Llama2 features an extensive context length of 4096 tokens, which are sufficiently ample to process long clinical text. Given the limited computational resources, we fine-tune the smallest version of Llama2 (7b) with the following instruction:

" Analyze the subject enrollment of the following clinical trial enclosed in square brackets. Determine the number of subjects as follows:
{Clinical Trial Text}: {Number of subjects}. " Using this format, we prepare the training dataset for LLama2 and fine-tune the model using LoRA technique [18]. The implementation is based on library PEFT[7].

For training the proposed model, we leverage AdamW optimization algorithm [28]. We use a small dropout rate of 0.3 for the category branch but we do not apply dropout regularization on other branches. The batch size is set to 256. We also use two different learning rates: the same learning rate of 0.0001 for the input modules, while the learning rate for the rest of the model is 0.001.

4.2.2 Deep Poisson-Gamma Setting

To maintain the statistical stability and avoid irregularities of enrollment data, we choose to filter the data similar to the setting in Zhong *et al.* [38]. Specifically, we keep only trials with more than 10 sites and a duration between 6 and 36 months. We then use a subset of the training, development, and test sets described in Section 4.1 by filtering the trials according to the criteria described in Table 2. After filtering, the training, development, and test sets contain 1641, 152, and 174 trials, respectively.

We follow the same procedure outlined in Section 4.2.1, and introduce two additional metrics, namely, the Median Average Error (MedAE) and the 6-month coverage. The latter represents the percentage of trials for which the actual duration falls within a 6-month window centered around our prediction. For the first approach leveraging our stochastic model described in Section 3.4, we train our stochastic model with a learning rate of 0.0001 and a batch-size of 32 for 256 epochs. After training our model $\mathcal{F}_{(\mu,\theta)}$, we use it to simulate the enrollment 1024 times with the Poisson-Gamma model. Each of these simulations produces a predicted duration $\hat{\tau}$. Subsequently, the predicted durations are averaged to obtain our final prediction.

---

[7]https://github.com/huggingface/peft

Table 3: The performance of the proposed method in comparison with baseline models. R^2 is the coefficient of determination. MAE is the mean absolute error. ($\uparrow$) indicates that higher values reflect better performance while ($\downarrow$) indicates that smaller values correspond to better performance. For XGBoost and LightGBM, * indicates that the embeddings of textual attributes are used. Baseline (first row) shows the MAE and R^2 calculated between planned and actual numbers.

Method	R^2 ($\uparrow$)	MAE ($\downarrow$)
Baseline (Target number of patients)	0.39	71.42
XGBoost* [11]	0.70	66.74
XGBoost [11]	0.72	62.22
LightGBM* [21]	0.71	62.61
LightGBM [21]	0.76	56.84
Llama2 [31]	0.07	110.98
ClinicalBERT* [32]	0.72	81.08
BioBERT* [24]	**0.77**	69.36
Clinical-Longformer-1536* [26]	0.74	59.88
Clinical-Longformer-2048* [26]	0.73	59.64
Our Deterministic Model*	0.76	**51.99**
Our Probabilistic Model*	**0.77**	55.92

For the "filtering-and-fitting" approach, we use an RMSprop optimizer to find the parameters α and λ of our distributions $\Gamma_{\mu,i}$ and $\Gamma_{\theta,i}$ for each trial. The parameters of the optimizer are a learning rate of 0.01, a batch-size of 128 and the optimization is carried out across 512 epochs. The final prediction is obtained in the same way as for the first approach. Moreover, it should be noted that, for the two different approaches, the maximum duration allowed for the prediction is two times the trial duration upper-limit described in Table 2 to avoid unnecessary computation while keeping enough margin for potential errors.

4.3 Result

4.3.1 Patient Enrollment Prediction

The performance of the proposed method compared to baseline models is presented in Table 3. Classical ML models perform reasonably good, surpassing the naive baseline of target number of patients, which achieves an R^2 of 0.39 and MAE of 71.42. It is also worth noting that simply concatenating textual embeddings to attribute embeddings does not generally improve performance. Notably, LightGBM without textual embeddings achieves top performance among classical ML models with an R^2 of 0.76 and MAE of 56.84. The PLM fine-tuning approach consistently achieves good R^2 scores. Nonetheless, ClinicalBERT and BioBERT do not have good performance on MAE, which is not the case for Clinical Longformer with an MAE of 59.64. This can be explained by the fact that Longformer has much longer context window compared to ClinicalBERT and BioBERT, which allows it to capture the details of the input clinical text.

Compared to the baselines, our models demonstrate superior performance in both metrics. Specifically, the deterministic model performs impressively, achieving an R^2 score of 0.76, which places it in the second-highest position among the models evaluated. The model excels in terms of MAE, delivering the best performance with a value of 51.99 - an improvement of approximately 9% over the leading baseline model. The stochastic model, despite its inherent variability, secures the second-best position in MAE and outperforms other models in R^2, achieving a score of 0.77. The superior performance of the proposed deterministic and stochastic models can be attributed to its architecture, which can effectively capture and represent multi-modality information.

The proposed stochastic model can be used for range estimation, in addition to point estimation. This ability enables the generation of confidence intervals around enrollment estimates, offering planners a range of plausible outcomes. Specifically, it is possible to predict an interval containing the number of patients with a confidence level of $(1 - \alpha)100\%$, where α is the significance level. For instance, choosing $\alpha = 0.1$ leads to a confidence level of 90%, a widely used confidence interval. Upon comparison with the actual number of patients, it is observed that 78.73% of ground-truth values fall into the predicted 90% confidence intervals, of which the median value of the interval widths is 99.35.
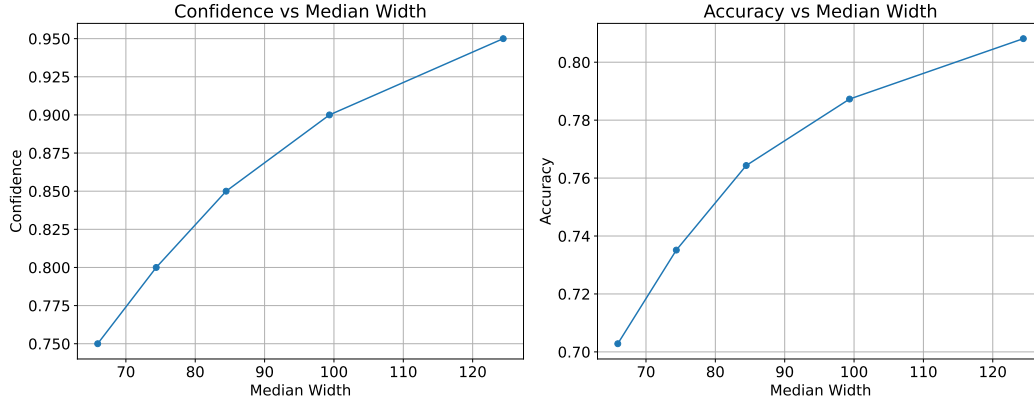
Figure 4: Relationships between interval width, confidence level, and interval accuracy of the stochastic model described in Section 3.3. Interval accuracy is the percentage of intervals that contain the actual value of number of patients.

Table 4: The performance of the Deep Poisson-Gamma model in comparison with the "Filtering-and-Fitting" approach in terms of trial duration prediction. The unit of trial duration is *month*.

	Filtering-and-Fitting	Deep Poisson-Gamma
MAE ($\downarrow$)	10.55	**7.52**
MedAE ($\downarrow$)	8.70	**5.59**
6-months coverage [%] ($\uparrow$)	14.94	**32.18**
Inference time [s] ($\downarrow$)	8.74	**0.07**

Adjusting the significance level directly influences the confidence level, interval width, and interval accuracy. This demonstrates the flexibility of the stochastic approach over the deterministic approach. Figure 4 illustrates the variation of confidence level and interval accuracy relative to interval width.

4.3.2 Deep Poisson-Gamma

As it can be seen in Table 4, our Deep Poisson-Gamma model outperforms the filtering-and-fitting approach on all metrics. More precisely, the former achieves an MAE of 7.52 months, a three-month error reduction from the latter. In terms of 6-month coverage accuracy, the proposed model performs twice as well as the filtering-and-fitting approach. Moreover, it can be seen that by leveraging machine learning paradigm, we can significantly reduce inference time. This is due to the fact that for the filtering-and-fitting approach, new distributions need to be fit on similar studies for each trial. This process can be time-consuming as what can be seen as "training" is part of the inference process. Our approach leveraging machine learning, allows to separate the training and inference processes and thus greatly reduce the inference time.

5 Conclusion

Accurately modeling the clinical trial enrollment process, especially in terms of number of patients and enrollment duration, is crucial for clinical trial planning given its potential benefits. In this work, we propose a novel approach to address this challenge. First, we create a novel model to predict the number of patients using trial characteristics. Leveraging the advances in large language models and deep learning, the proposed model is capable of capturing the complexities and nuances of clinical trial documents to effectively make enrollment prediction. We further extend the model by accounting for uncertainties with the Gamma distribution; this way, we naturally connect the statistical approach with the modern machine learning based approach in a unified stochastic model. Using the unified model, it is possible to estimate the trial duration via a Poisson-Gamma model, which covers site-level uncertainties.

Our experiments show that the proposed method offers superior performance compared to existing approaches in both accuracy and scalability. Our future work will focus on exploring an end-to-end architecture that incorporates large language models directly. In addition, our method is designed to be used at planning phase, when no specific information about sites is available. Hence, we plan to extend this method to consider ongoing trials where the enrollment of patients can be effectively predicted in a near real-time setting to dynamically adjust enrollment forecasts and support adaptive trial management.

Acknowledgements

References

[1] M. Amiridi, C. Qian, N. D. Sidiropoulos, and L. M. Glass. Enrollment rate prediction in clinical trials based on cdf sketching and tensor factorization tools. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[2] V. Anisimov. Predictive modelling of recruitment and drug supply in multicenter clinical trials. In *Proc. of joint statistical meeting*, pages 1248–1259, 2009.

[3] V. V. Anisimov. Using mixed poisson models in patient recruitment in multicentre clinical trials. In *Proceedings of the World Congress on Engineering*, volume 2, pages 1046–1049, 2008.

[4] V. V. Anisimov and V. V. Fedorov. Design of multicentre clinical trials with random enrolment. *Advances in Statistical Methods for the Health Sciences: Applications to Cancer and Aids Studies, Genome Sequence Analysis, and Survival Analysis*, pages 387–400, 2007.

[5] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[6] C. Bieganek, C. Aliferis, and S. Ma. Prediction of clinical trial enrollment rates. *Plos one*, 17(2):e0263193, 2022.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] B. Carlisle, J. Kimmelman, T. Ramsay, and N. MacKinnon. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. *Clinical trials*, 12(1):77–83, 2015.

[9] R. E. Carter. Application of stochastic processes to participant recruitment in clinical trials. *Controlled clinical trials*, 25(5):429–436, 2004.

[10] R. E. Carter, S. C. Sonne, and K. T. Brady. Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC medical research methodology*, 5:1–5, 2005.

[11] T. Chen and C. 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, 2016.

[12] S. Comfort. Improving clinical trial enrollment forecasts using sorm. *Applied Clinical Trials*, 22, 2013.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[14] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[15] X. Fang, W. Xu, F. A. Tan, Z. Hu, J. Zhang, Y. Qi, S. H. Sengamedu, and C. Faloutsos. Large language models (LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*, 2024.

[16] R. L. Harrison. Introduction to monte carlo simulation. In *AIP conference proceedings*, volume 1204, page 17, 2010.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[19] K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[20] O. Johnson. An evidence-based approach to conducting clinical trial feasibility assessments. *Clinical Investigation*, 5:491–499, 2015.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[22] D. Kevian, U. Syed, X. Guo, A. Havens, G. Dullerud, P. Seiler, L. Qin, and B. Hu. Capabilities of large language models in control engineering: A benchmark study on gpt-4, claude 3 opus, and gemini 1.0 ultra. *arXiv preprint arXiv:2404.03647*, 2024.

[23] E. Kim, J. Yang, S. Park, and K. Shin. Factors affecting success of new drug clinical trials. *Therapeutic Innovation & Regulatory Science*, 57(4):737–750, 2023.

[24] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[25] R. Li, X. Wang, and H. Yu. Llamacare: An instruction fine-tuned large language model for clinical nlp. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10632–10641, 2024.

[26] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022.

[27] J. Liu, P. J. Allen, L. Benz, D. Blickstein, E. Okidi, and X. Shi. A machine learning approach for recruitment prediction in clinical trial design. *arXiv preprint arXiv:2111.07407*, 2021.

[28] I. Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[29] T. Savage, S. Ma, A. Boukil, V. Patel, E. Rangan, I. Rodriguez, and J. H. Chen. Fine tuning large language models for medicine: The role and importance of direct preference optimization. *arXiv preprint arXiv:2409.12741*, 2024.

[30] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[31] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[32] G. Wang, X. Liu, Z. Ying, G. Yang, Z. Chen, Z. Liu, M. Zhang, H. Yan, Y. Lu, Y. Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642, 2023.

[33] Z. Wang and J. Sun. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. *arXiv preprint arXiv:2206.14719*, 2022.

[34] Z. Wang, C. Xiao, and J. Sun. Spot: Sequential predictive modeling of clinical trial outcome with meta-learning. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–11, 2023.

[35] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

[36] L. Yue, J. Li, S. Xing, M. Z. Islam, B. Xia, T. Fu, and J. Chen. Trialdura: Hierarchical attention transformer for interpretable clinical trial duration prediction. *arXiv preprint arXiv:2404.13235*, 2024.

[37] L. Yue, S. Xing, J. Chen, and T. Fu. Trialenroll: Predicting clinical trial enrollment success with deep & cross network and large language models. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–9, 2024.

[38] S. Zhong, Y. Xing, M. Yu, and L. Wang. Enrollment forecast for clinical trials at the portfolio planning phase based on site-level historical data. *Pharmaceutical Statistics*, 23(2):151–167, 2024.