LIFTED: Multimodal Clinical Trial Outcome Prediction via Large Language Models and Mixture-of-Experts

**Wenhao Zheng[†,1], Liaoyaqi Wang[†,2], Dongshen Peng[1], Hongxia Xu[3],
Yun Li[1], Hongtu Zhu[1], Tianfan Fu[4], Huaxiu Yao[*,1]**

[1]The University of North Carolina at Chapel Hill, [2]Johns Hopkins University,
[3]Zhejiang University, [4]Rensselaer Polytechnic Institute

[†]Equal Contribution [*]Correspondence: wenhao@unc.edu, huaxiu@unc.edu

Abstract

Clinical trials are pivotal yet costly processes, often spanning multiple years and requiring substantial expenses, motivating predictive models to identify likely-to-fail drugs early and save resources. Recent approaches leverage deep learning to integrate multimodal data for clinical outcome prediction; however, they rely heavily on manually designed modality-specific encoders, limiting their adaptability to new modalities and ability to effectively share information across modalities. To address these challenges, we propose a multimodal mixture-of-experts (LIFTED) framework. Specifically, LIFTED transforms modality-specific data into natural language descriptions, encoded via unified, noise-resilient encoders. A sparse Mixture-of-Experts mechanism then identifies shared patterns across modalities, extracting consistent representations. Finally, another mixture-of-experts module dynamically integrates these modality representations, emphasizing critical information. Experiments show that LIFTED significantly outperforms baseline methods in predicting clinical trial outcomes across all phases, highlighting the effectiveness of our proposed approach.

1 Introduction

The clinical trial is a crucial step in the development of new treatments to demonstrate the safety and efficacy of the drug. Drugs must pass three trial phases involving human participants with target diseases before approval for manufacturing. However, the clinical trial is time-consuming and experiments expensive, taking multiple years and costing up to hundreds of millions of dollars (Martin et al., 2017). In addition, the success rate of clinical trials is exceedingly low and many drugs fail to pass these clinical trials (Wu et al., 2022b; Huang et al., 2020). Therefore, the ability to predict clinical trial outcomes beforehand, allowing the exclusion of drugs with a high likelihood of failure, holds the potential to yield significant cost savings. Given the increasing accumulation of clinical trial data over the past decade (e.g., drug descriptions, and patient criteria), we can now leverage this wealth of data for the prediction of clinical trial outcomes.

Early attempts aim to improve the clinical trial outcome prediction results by modeling the components of the drugs (e.g., drug toxicity (Gayvert et al., 2016), modeled the pharmacokinetics (Qi and Tang, 2019)). Recently, deep learning methods have been proposed for trial outcome predictions. For instance, Lo et al. (2019) predicted drug approvals for 15 different disease groups by incorporating drug and clinical trial features into machine learning models. Fu et al. (2022) proposed an interaction network leveraging multimodal data (e.g., molecule information, trial documents) to capture correlations for trial outcome predictions. However, this approach relies on modal-specific encoders to extract representations from different modal data, which require manually designed encoder structures and limit their extensibility when new modal data becomes available for use.

To address these issues, we aim to design a unified encoder to extract representations from various modalities, but it poses the following challenges:

- **How to extract representations from different modalities with a unified encoder?** Different modalities are represented in various data formats. For instance, molecule information is typically depicted as graphs, while disease names rely on relationships between diseases. Therefore, a unified encoder structure should be capable of unifying these different formats to effectively extract information.

- **How to effectively utilize both the modality-independent information patterns and the modality-specific patterns to enhance the extracted representations?** Information across different modalities can be presented in both similar
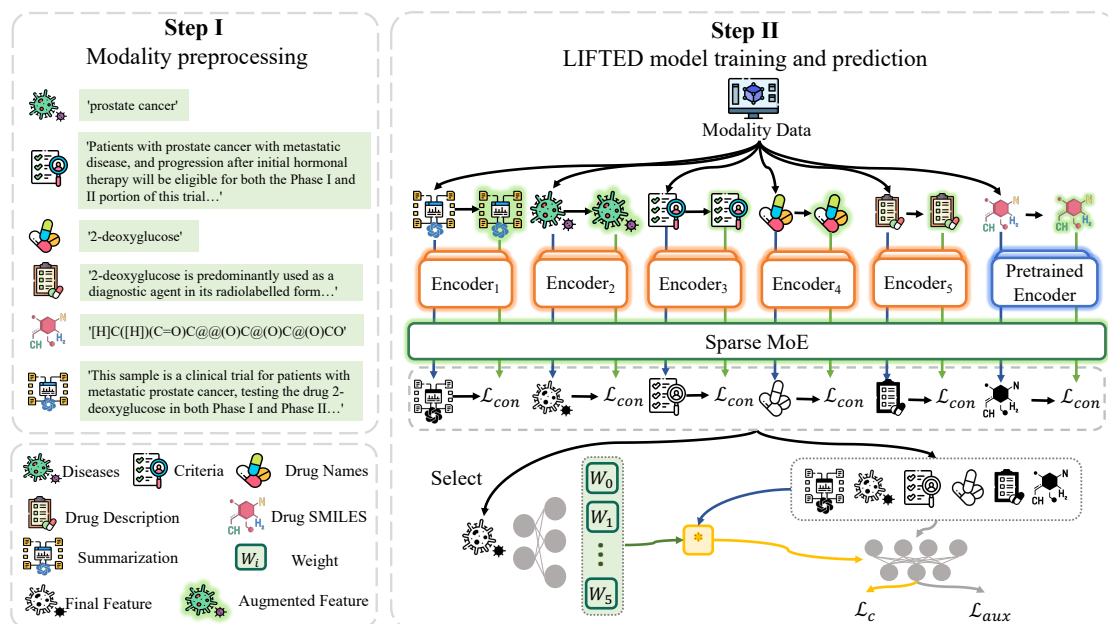
Figure 1: An overview of LIFTED. **Step 1**: Transforming multimodal data into natural language descriptions, where all modalities are converted into natural language descriptions to facilitate the representation extraction process of the transformer encoders. **Step 2**: Extract and combine representations from different modalities, where representations are extracted by the noise-resilient unified encoders and integrated by a Mixture-of-Experts (MoE) framework to make the final predictions.

and different forms. For example, descriptions of a disease and corresponding drugs may mention the same symptoms, which can be extracted similarly. However, molecules and drug names represent information differently and should be extracted using distinct methods. Therefore, a method to dynamically identify similar information patterns across different modalities and direct them to the same encoder is also required.

- **How to integrate extracted information from different modalities?** Extracted representations from different modalities must be integrated for accurate predictions, yet the importance of each modality can vary greatly among patients. For example, one patient's outcome might heavily depend on a challenging disease like type 2 diabetes mellitus (Wu et al., 2022a), while another's outcome could be primarily influenced by effective medications. Therefore, automatically weighting modality-specific representations is essential.

To address those challenges, we propose an approach called Multimodal Clinical Trial Outcome Prediction via Large Language Models and Mixture-of-Experts (LIFTED), which extracts information from different modalities with a transformer based unified encoder, enhances the extracted features by a Sparse Mixture-of-Experts

(SMoE) framework and integrates multimodal information with Mixture-of-Experts (MoE). Specifically, LIFTED unifies diverse multimodal features, even those in different formats, by converting them into natural language descriptions. Subsequently, we build a unified transformer-based encoder to extract representations from these modal-specific language descriptions and refine the representations with an SMoE framework. Here, the representations from different modalities are dynamically routed by a noisy top-k gating network to a portion of shared expert models, facilitating the extraction of similar information patterns. In addition, we introduce representation augmentation to enhance the resilience of transform-based encoders and the SMoE framework to potential data noise introduced during the data collection process. Furthermore, LIFTED treats the extracted representations from various modalities as distinct experts and utilizes a Mixture-of-Experts module to dynamically combine these multimodal representations for each example. This dynamic combination allows for the automatic assignment of higher weights to more crucial modalities. Finally, we evaluate LIFTED on the HINT benchmark (Fu et al., 2022) and the CTOD benchmark (Gao et al., 2024) to demonstrate the effectiveness of LIFTED and the effectiveness of our proposed components.

2 Multimodal Clinical Trial Outcome Prediction via Large Language Models and Mixture-of-Experts

This section introduces our proposed Multimodal Clinical Trial Outcome Prediction via Large Language Models and Mixture-of-Experts (LIFTED) approach. The objective of LIFTED is to integrate multimodal clinical trial data using natural language descriptions within a Mixture-of-Experts (MoE) framework, as illustrated in Figure 1. Our method involves two main steps: firstly, multimodal data is converted into natural language descriptions using a Large Language Model (LLM). Secondly, these descriptions along with their corresponding labels are utilized to train our predictive model. Specifically, we employ transformer-based encoders to derive meaningful representations from the textual data. To improve the robustness of these embeddings, we introduce perturbations and apply a consistency loss. Finally, the representations from different modalities are integrated through an MoE framework to produce the predictions. Each of these components is explained comprehensively in the following sections.

2.1 Transforming Multimodal Data into Natural Language Descriptions

To build a unified encoder, the key challenge is how to unify multimodal data, which often have different structures for different modalities. For instance, molecule information is typically depicted as a graph, while disease names rely on relationships between different diseases (Wu et al., 2022a). In LIFTED, we unify these different modality data by converting them into natural language descriptions. Specifically, we first format the input features into a key-value pair. After that, we use a prompt coupled with the corresponding key-value pair to ask an LLM to generate a natural language description for our input. Subsequently, these descriptions will be fed into a unified tokenizer for further encoding, except the SMILES string modality, which is tokenized by a specifically designed tokenizer to enhance the representation of molecule information. The first two steps, linearization and prompting, are detailed below:

Linearization. In linearization, we format each data point $x_{i,k}$ of trial i and modality k into a key-value pair. In this pair, the key of each element represents the feature name $c_{i,k}$, and the corresponding
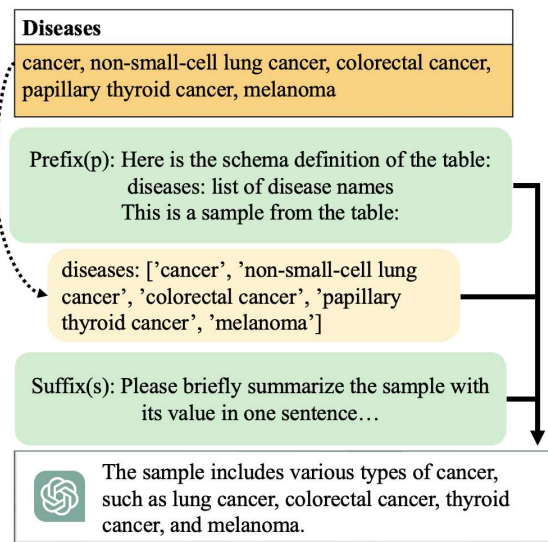

Figure 2: Processes of the linearization and the prompting.

value is $x_{i,k}$. This can be formulated as follows:

$$\text{Linearize}(x_{i,k}) = \{c_{i,k} : x_{i,k}\}. \qquad (1)$$

Prompting. As depicted in Figure 2, the prompts we use to communicate with the LLM consist of three components: a prefix p to describe the schema of the input features, the linearization and a suffix s to instruct the LLM on how to describe the input data point in natural language. Given the prompts, the LLM will generate a readable and concise natural description $z_{i,k}$, which can be formulated as:

$$z_{i,k} = \text{LLM}(p, \text{Linearize}(x_{i,k}), s). \qquad (2)$$

For instance, given the linearization of disease modality, "diseases: ['cancer', 'non-small-cell lung cancer', 'colorectal cancer', 'papillary thyroid cancer', 'melanoma']", the LLM will generate the natural language description like "The sample includes various types of cancer, such as lung cancer, colorectal cancer, thyroid cancer, and melanoma."

In addition to transforming existing modality data to natural language descriptions, we also generate a new summarization modality to provide an overall description of the whole trial in a similar way. The only difference for generating the summarization is that we concatenate the linearization of all modalities as input to provide the information of the whole trial as

$$z_{i,0} = \text{LLM}(p, \text{Linearize}(x_i), s),$$
$$\text{where} \quad \text{Linearize}(x_i) = \{c_{i,k} : x_{i,k}\}_{k=1}^{K}. \qquad (3)$$

By this, all data is summarized in text format.

2.2 Representation Learning and Refinement

After transforming multimodal data into natural language descriptions, we build $K+1$ transformer-based encoders on top of these descriptions. Specifically, each modality description $z_{i,k}$ is tokenized into a sequence of tokens $\{z_{i,k}^t\}_{t=1}^T$ with length T by a tokenizer $\mathcal{T}$ and embedded into a sequence of embeddings $\{u_{i,k}^t\}_{t=1}^T$ by a modal-specific embedding layer $\mathcal{E}_k$ first, and then they are added by the position embeddings pos^t and fed into the correponding modal-specific transformer encoder $\mathcal{F}_k$ coupled with a learnable token $[cls]_k$ to get encoded representation $U_{i,k}$. The encoding process can be formulated as follows:

$$
\begin{aligned}
\{z_{i,k}^t\}_{t=1}^T &= \mathcal{T}(z_{i,k}) \\
u_{i,k}^t &= \mathcal{E}_k(z_{i,k}^t) \\
U_{i,k} &= \mathcal{F}_k(\{u_{i,k}^t + \text{pos}^t\}_{t=0}^T),
\end{aligned} \tag{4}
$$

where, we define $u_{i,k}^0 \equiv [cls]_k$.

Furthermore, to equip LIFTED with the capability to dynamically identify similar information patterns across different modalities and route them to the same encoder, we employ a Sparse Mixture-of-Experts (SMoE) framework to further refine the extracted representations. The encoded representations $U_{i,k}$ from different modalities will be dynamically routed by a modality-independent noisy top-k gating network $\mathcal{G}$ to a subset of shared expert models $\{\mathcal{R}^r\}_{r=1}^R$ to facilitate the extraction of similar information patterns, following the original design of SMoE (Shazeer et al., 2017). The whole process can be formulated as follows:

$$
\begin{aligned}
\mathcal{G}(U_{i,k}) &= \text{Softmax}(\text{TopK}(\mathcal{P}(U_{i,k}), k)) \\
\mathcal{P}(U_{i,k}) &= U_{i,k} \cdot W_g + \mu \text{Softplus}(U_{i,k} \cdot W_{\text{noise}}) \\
\text{TopK}(v, k)_j &= \begin{cases} v_j, \text{if } v_j \text{ in the top } k \text{ elements of } v \\ -\infty, \text{otherwise} \end{cases}
\end{aligned} \tag{5}
$$

where the μ is random noise sampled from a standard normal distribution, W_g is a learnable weight matrix shared through different modalities and W_{noise} is another learnable noise matrix to control the amount of noise per component. Subsequently, the encoded representations $U_{i,k}$ will be routed only to the shared expert models $\{\mathcal{R}^r\}_{r=1}^R$ with top-k gating scores generated by the gating network $\mathcal{G}$. The refined representations $\tilde{U}_{i,k}$ can then be calculated by combining the encoding results from the top-k expert models with their corresponding gating scores. The whole process can be

formulated as follows:

$$
\begin{aligned}
\tilde{U}_{i,k} &= \mathcal{G}(U_{i,k}) \cdot \mathcal{R}(U_{i,k}) \\
&= \sum_{r=1}^R \mathcal{G}^r(U_{i,k}) \mathcal{R}^r(U_{i,k}).
\end{aligned} \tag{6}
$$

2.3 Consistent Representation Augmentation

However, building informative modal-specific encoders and the SMoE framework solely from these modal-specific natural language descriptions remains challenging, primarily due to potential noise introduced during the data collection process. To enhance robustness against such noise, we augment embeddings $u_{i,k}^t$ with small perturbations to obtain a perturbed repersentation $v_{i,k}^t$, and introduce a consistency loss, encouraging the encoders and SMoE framework to produce stable outputs despite minor data variations.

Representation Augmentation. To perform representation augmentation, we begin by considering each embedding vector $u_{i,k}^t \in \mathbb{R}^L$, where L represents the number of elements $\{m_l\}_{l=1}^L$. We randomly select a subset of these elements from $u_{i,k}^t$ with a probability p for perturbation, while leaving the remaining elements unchanged. For simplicity, we will omit the subscript and superscript for m_l specific to the embedding $u_{i,k}^t$. Next, we proceed to sample a small value α_l from a uniform distribution $\text{Uniform}(-\lambda, \lambda)$ for each selected element. Here, λ serves as a hyperparameter that controls the magnitude of the minor perturbation. Following this, each selected element is multiplied by $\exp(\alpha_l)$ to apply the perturbation. This process can be expressed as:

$$
\hat{m}_l = \begin{cases} \exp(\alpha_l) * m_l, & m_l \text{ is selected} \\ m_l, & \text{otherwise} \end{cases} \tag{7}
$$

The perturbed vector is defined as $v_{i,k}^t = \{\hat{m}_l\}_{l=1}^L$.

Consistency Loss. After the representation augmentation step, we obtain a sequence of perturbed embeddings $\{v_{i,k}^t\}_{t=1}^T$ derived from the original embeddings $\{u_{i,k}^t\}_{t=1}^T$. These perturbed embeddings are then input into the encoder $\mathcal{F}_k$ and the SMoE framework to generate the encoded representation $\tilde{V}_{i,k}$. In order to ensure the robustness of the encoded embeddings, we introduce a consistency loss $\mathcal{L}_{con}$ to control the disparity between the encoded representation of the original embeddings

and the augmented embeddings. This consistency loss can be formulated as follows

$$\mathcal{L}_{con} = \frac{1}{N(K+1)} \sum_{k=0}^{K} \sum_{i=1}^{N} \left\| \tilde{U}_{i,k} - \tilde{V}_{i,k} \right\|_F^2,$$

where $\tilde{V}_{i,k} = \mathcal{G}(V_{i,k}) \cdot \mathcal{R}(V_{i,k}),$

$$V_{i,k} = \mathcal{F}_k(\{v_{i,k}^t + pos^t\}_{t=0}^T), \tag{8}$$

where we define $v_{i,k}^0 \equiv [cls]_k$.

2.4 Integrating Multimodal Information

After obtaining representations from different modalities, the model's ability to integrate these multimodal representations and discern the patient-specific importance of each modality becomes crucial for precise predictions. As illustrated in Figure 1, we employ a Mixture-of-Experts (MoE) framework to dynamically integrate multimodal representations. In this framework, we treat the extracted representations from various modalities as distinct experts.

Concretely, for each example i, we start by concatenating the extracted representations from the selected modalities and then feed them into a fully connected layer denoted as $\mathcal{C}$ to calculate the modality importance weights $W_{i,k}$ for each modality. In our implementation, we exclusively utilize the disease modality to generate these importance weights, as knowing the patient's disease allows us to determine which modality should receive emphasis. Subsequently, we multiply these weights by their corresponding representations $\{U_{i,k}\}_{k=0}^K$ and aggregate them to obtain the integrated representation U_i. This can be formulated as follows:

$$W_{i,k} = \text{Softmax}(\mathcal{C}(\oplus_{j \in \mathcal{J}} U_{i,j}) * \gamma_k)$$

$$U_i = \sum_{k=0}^{K} W_{i,k} * \tilde{U}_{i,k}, \tag{9}$$

where the $\oplus$ is the concatenate operation along the representation dimension and the $\mathcal{J}$ is the set of selected modalities. γ_k is a learnable modal-specific temperature factor.

Following this, we make the prediction $\hat{y}_i$ by inputting the integrated representation U_i into the classifier $\mathcal{H}$. The classification loss $\mathcal{L}_c$ is defined as follows:

$$\mathcal{L}_c = \frac{1}{N} \sum_{i=1}^{N} \ell(\hat{y}_i, y_i), \quad \text{where } \hat{y}_i = \mathcal{H}(U_i) \tag{10}$$

where the y_i is the ground truth label for sample i and the loss term ℓ is the cross entropy loss.

To ensure that the unimodal representations are of high quality and consistently contribute to the final prediction, we introduce an auxiliary loss to align the representations from different modalities. Similar to the classification loss $\mathcal{L}_c$, the auxiliary loss $\mathcal{L}_{aux}$ is calculated as the sum of uni-modal prediction losses, which can be formulated as follows:

$$\hat{y}_{i,k} = \mathcal{H}(U_{i,k})$$

$$\mathcal{L}_{aux} = \frac{1}{N(K+1)} \sum_{k=0}^{K} \sum_{i=1}^{N} \ell(\hat{y}_{i,k}, y_i). \tag{11}$$

Finally, the overall loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_c + \eta_1 \mathcal{L}_{con} + \eta_2 \mathcal{L}_{aux}, \tag{12}$$

where η_1 and η_2 are hyperparameters to balance these loss terms. The whole algorithm is illustrated in Appendex B.

3 Experiments

In this section, we evaluate the performance of LIFTED aiming to answer the following questions: **Q1**: Compared to the existing methods with modal-specific encoders, can LIFTED achieve better performance with the unified transformer encoders? **Q2**: Do the key components, including the multi-modal data integration component and the representation augmentation component, of LIFTED boost the performance? **Q3**: Does the Sparse Mixture-of-Experts framework route similar information patterns in different modalities to the same expert models correctly? **Q4**: Does the Mixture-of-Experts approach precisely measure the importance of different modalities for each patient?

3.1 Experimental Setup

Dataset Descriptions. We evaluate our method and other baselines on the HINT dataset (Fu et al., 2022; Chen et al., 2024b) and CTOD dataset (Gao et al., 2024), covering Phases I, II and III trials. More details of the HINT dataset and the CTOD dataset are shown in Appendix C.

Baselines. We compare LIFTED with both machine learning methods (Siah et al., 2021; Yi et al., 2018; Rajpurkar et al., 2020; Fan et al., 2020; Lu, 2018; Chen et al., 2024a) and deep learning models (Zhang et al., 2020; Gao et al., 2020), such as

Table 1: The clinical trial outcome performance (%) of LIFTED and baselines on HINT dataset and CTOD dataset. The results are averaged on 30 independent runs with different random seeds. [†]: The results of the HINT and SPOT methods were obtained by running their released codes. The best results and second best results are **bold** and underlined, respectively. We observe that LIFTED consistently outperforms all other methods over all three phases.

| | HINT | | | | | | | | | CTOD | | | | | | | | |
| | Phase I Trials | | | Phase II Trials | | | Phase III Trials | | | Phase I Trials | | | Phase II Trials | | | Phase III Trials | | |
Method	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
LR	50.0	60.4	52.0	56.5	55.5	58.7	68.7	69.8	65.0	55.6	58.5	51.5	56.0	63.6	56.7	75.0	73.8	51.9
RF	51.8	62.1	52.5	57.8	56.3	58.8	69.2	68.6	66.3	58.3	70.0	54.0	60.8	70.9	56.0	75.1	85.1	50.7
XGBoost	51.3	62.1	51.8	58.6	57.0	60.0	69.7	69.6	66.7	57.4	68.0	50.5	61.2	70.6	61.3	78.7	84.9	57.1
AdaBoost	51.9	62.2	52.6	58.6	58.3	60.3	70.1	69.5	67.0	58.8	66.7	53.5	60.2	68.9	55.2	79.5	84.6	56.9
kNN+RF	53.1	62.5	53.8	59.4	59.0	59.7	70.7	69.8	67.8	58.4	70.3	54.7	61.2	70.9	56.6	76.8	85.2	52.5
FFNN	54.7	63.4	55.0	60.4	59.9	61.1	74.7	74.8	68.1	55.2	58.4	48.7	57.2	66.0	53.0	76.1	79.6	52.6
MMF (early)	60.6	59.4	54.4	60.2	62.6	60.7	85.5	81.5	70.6	61.2	64.2	56.2	64.0	70.1	59.0	83.3	84.4	64.6
MMF (late)	63.4	67.5	59.0	62.9	63.0	62.6	86.9	83.1	71.8	63.0	70.5	57.4	65.7	71.4	60.3	83.5	85.2	65.8
HINT[†]	58.4	68.2	62.1	59.1	63.9	62.8	85.9	80.9	70.8	63.4	71.0	57.6	64.6	71.2	60.6	81.8	85.7	60.8
SPOT[†]	69.8	68.4	64.6	62.6	64.3	63.0	81.7	81.0	71.0	66.8	70.0	62.3	64.5	71.8	58.8	83.2	77.6	67.5
LIFTED	**70.7**	**71.6**	**64.9**	**69.8**	**66.2**	**65.1**	**88.3**	**83.8**	**73.5**	**69.7**	**71.8**	**63.4**	**67.7**	**72.0**	**63.0**	**86.7**	**85.9**	**69.5**

Feedforward Neural Network (FFNN) (Shen et al., 2023), Multi Modal Fusion (MMF), HINT (Fu et al., 2022; Wang et al., 2024), SPOT (Wang et al., 2023b). More details of those baselines are presented in Appendix B.

Evaluation Metrics. Following Fu *et al.* (2022) and Chen *et al.* (2024b), we use F1 score, PR-AUC, and ROC-AUC to measure the performance of all methods. For all these metrics, higher scores indicate better performance.

Table 2: The clinical trial outcome prediction performance (%) of LIFTED and variants without certain key component. The best results are **bold**. LIFTED outperforms all variants, showcasing the effectiveness of our proposed components.

| HINT | | | | | | | | | |
| | Phase I Trials | | | Phase II Trials | | | Phase III Trials | | |
Method	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
LIFTED-aug	68.4	69.8	64.8	69.5	66.0	64.3	86.9	82.4	72.1
LIFTED-aux	69.0	71.2	63.7	69.6	64.5	64.6	87.4	82.8	71.1
LIFTED-LLM	68.5	70.8	64.0	69.7	64.9	65.0	86.7	82.7	70.8
LIFTED-gating	69.9	71.3	64.9	69.7	65.5	65.0	87.0	82.7	72.4
LIFTED	**70.7**	**71.6**	**64.9**	**69.8**	**66.2**	**65.1**	**88.3**	**83.8**	**73.5**
CTOD									
	Phase I Trials			Phase II Trials			Phase III Trials		
Method	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
LIFTED-aug	65.1	70.6	61.9	67.1	70.5	62.2	85.2	82.4	67.8
LIFTED-aux	64.4	71.1	60.0	60.0	71.0	54.9	83.8	85.5	64.6
LIFTED-LLM	65.1	71.4	58.5	66.3	71.1	61.2	83.9	85.4	65.3
LIFTED-gating	67.1	71.4	60.8	65.1	70.4	61.9	85.0	85.8	68.8
LIFTED	**67.7**	**71.6**	**62.3**	**67.7**	**72.0**	**63.0**	**86.7**	**85.9**	**69.5**

3.2 Overall Performance

We conduct experiments to evaluate the performance of LIFTED on all three phases trails, compared to our baselines. The trial outcome prediction results of all models are reported in Table 1. We

Table 3: Performance analysis of multimodal data integration. The best results and second best results are **bold** and underlined, respectively.

| HINT | | | | | | | | | |
| | Phase I Trials | | | Phase II Trials | | | Phase III Trials | | |
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
Summarization	63.2	69.9	57.8	66.1	61.2	61.2	85.1	80.7	66.6
Drugs	62.1	67.2	57.9	60.5	62.3	55.8	83.8	81.7	63.8
Disease	65.3	67.3	59.7	68.0	59.9	62.4	86.0	80.5	69.1
Description	55.5	71.3	50.2	55.5	0.0	50.0	74.9	85.7	49.7
SMILES	62.8	69.6	58.5	59.3	58.3	54.9	76.1	83.6	51.1
Criteria	68.0	70.5	63.1	67.6	64.4	63.0	83.7	82.7	65.0
All (LIFTED)	**70.7**	**71.6**	**64.9**	**69.8**	**66.2**	**65.1**	**88.3**	83.8	**73.5**
CTOD									
	Phase I Trials			Phase II Trials			Phase III Trials		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
Summarization	61.6	70.9	56.9	65.4	71.5	61.1	84.1	84.8	64.9
Drugs	60.8	70.6	57.6	58.0	71.3	53.8	77.3	85.2	54.5
Disease	65.2	70.2	61.3	68.3	71.1	62.6	85.3	85.5	68.4
Description	56.6	70.7	52.3	56.0	71.3	52.4	76.5	85.7	52.9
SMILES	61.6	70.6	56.0	58.0	71.5	53.1	75.1	85.6	50.8
Criteria	60.5	71.2	55.8	59.9	71.2	53.9	75.7	84.6	52.5
All (LIFTED)	**67.7**	**71.6**	**62.3**	**67.7**	**72.0**	**63.0**	**86.7**	**85.9**	**69.5**

first observed that the deep learning-based methods and the methods designed for clinical trial outcome prediction outperform the machine learning based methods with a significant performance gap, especially on the HINT dataset, showcasing the powerful ability to extract critical information from different modalities in various formats of the deep learning encoders and those encoders specifically designed to extract representation hidden in the clinical trial records. This observation is not surprising, since the critical information of different modalities is represented in different ways, which is hard to extract for those traditional machine learning methods or those deep learning encoders that are not designed for clinical trial outcome prediction. Nevertheless, LIFTED consistently outperforms all

other methods over all three phases, verifying its effectiveness in unifying different modalities and dynamically integrating them within the MoE.

3.3 Ablation Study

In this section, we perform comprehensive ablation studies to demonstrate the effectiveness of our key components, including the representation augmentation, the auxiliary loss and the modalities used to generate weights in the Mixture-of-Experts (MoE) framework.

- **LIFTED-aug:** In LIFTED-aug, the representation augmentation component and the consistency loss are removed. Representations from different modalities are directly fed into the multimodal data integration component without the constraint of robustness to the noise in data.

- **LIFTED-aux:** In LIFTED-aux, we remove the auxiliary loss component. Representations from different modalities are no longer required to make consistent predictions with the final representation.

- **LIFTED-LLM:** In LIFTED-LLM, we remove the transformation preprocessing step and utilize the linearization, instead of the natural language description, of each modality as input. In addition, the summarization modality is also removed, since it is generated by LLM.

- **LIFTED-gating:** In LIFTED-gating, we use all modalities instead of just disease modality to generate the weights for the multimodal data integration component.

The results are shown in Table 2, and the results of LIFTED are also reported for comparison. From those tables, we observe that: (1) LIFTED outperforms all the variants without certain components, including LIFTED-aug, LIFTED-aux and LIFTED-LLM, showcasing the effectiveness and complementary of the representation augmentation component, the auxiliary loss component and the LLM transformation preprocessing step; (2) LIFTED outperforms its variant, LIFTED-gating, with a slight advantage in performance. This suggests that determining the modality importance for each trial based solely on disease information is sufficient. Including additional modality information, even to a slight extent, appears to have a negative impact on performance.

Figure 3: The SMoE experts' importance weights of our model predicting the knee osteoarthritis patient. Experts 6 and 7 play a crucial role in extracting common information patterns across modalities, while other experts specialize in a single specific modality.

3.4 Analysis of Multimodal Data Integration

We further analyze how multimodal data integration contributes to clinical outcome prediction. Here, we compare the performance of models using data from only one modality with LIFTED that integrates all those modalities. We report the results in Table 3. The results indicate that LIFTED outperforms almost all unimodal models, demonstrating the effectiveness of multimodal integration. In addition, the results also demonstrate that the drug description and the criteria modalities are the least and the most important modality, respectively, which is expected since the quality of recruited patients plays a crucial role in trial success (Jin et al., 2017; Zhang et al., 2021).

3.5 Analysis of Sparse Mixture-of-Experts

In addition, we delve into an analysis of the Sparse MoE model to understand the performance enhancements obtained by it. Here, we select a knee osteoarthritis patient case. For each modality, the SMoE framework selects top-3 experts from a pool of 16 experts with the highest weights. The weights of these selected SMoE experts are visualized in Figure 3. As expected, certain experts, such as 6 and 7, are consistently chosen across multiple modalities, indicating their pivotal role in extracting similar information patterns among different modalities. Furthermore, other experts demonstrate a more focused expertise, concentrating on one or two modalities. This demonstrates the effectiveness of the SMoE framework in both extracting similar information patterns across different modal-
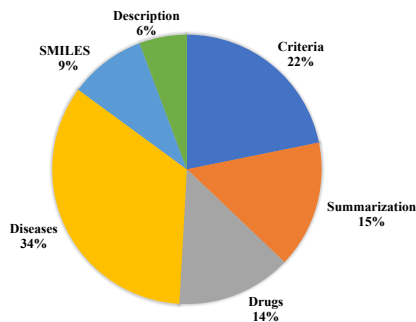
Figure 4: The modality importance weights of our model predicting the type 2 diabetes mellitus patient. LIFTED pay more attention to the disease modality as expected, since type 2 diabetes mellitus is hard to cure.

ities and capturing specialized information patterns within a single modality.

3.6 Case Study

In addition, we conduct a case study to analyze the contribution of each modality in clinical trial outcome prediction. Specifically, we analyze the result of a type 2 diabetes mellitus patient, who was inadequately controlled with metformin at the maximal effective and tolerated dose of metformin for at least 12 weeks. Since type 2 diabetes mellitus is hard to cure (Chang et al., 2019), the model should pay attention to the name of the disease and predict the trial as failed, which is consistent with the behavior of our model. The modality importance weights are shown in Figure 4. As we expected, the attention weights of the disease modality are much higher than other modalities, which demonstrates that our model pays attention to the disease modality and predicts the trial correctly.

4 Related Works

Clinical Trial Outcome Prediction. Machine learning methods have been proven efficient on diverse tabular data prediction tasks, especially the clinical trial outcome prediction task, resulting in profound performances (Chen et al., 2023; Yan et al., 2023). Recently, Fu et al. (2023) proposed a hierarchical interaction network employing different encoders to fuse multiple modal data and capture their correlations for trial outcome predictions; Wang et al. (2023b) clustered multi-sourced trial data into different topics, organizing trial embeddings for prediction. Wang et al. (2023a) converted clinical trial data into a format compatible description for prediction. However, converting all modalities into a single description poses significant challenges. This approach makes it difficult

for the model to distinguish the unique information of each modality and necessitates external data to aid in differentiating these modalities. In contrast, LIFTED extracts representations for each modality separately and dynamically integrates them, providing a more effective way to preserve distinct characteristics of each modality.

Mixture-of-Experts. Mixture-of-Experts (MoE) is a special type of neural network whose parameters are partitioned into a series of sub-modules, called experts, functioning in a conditional computation fashion (Jacobs et al., 1991; Jordan and Jacobs, 1993). Recently, Shazzer et al. (2017) simplified the MoE layer by selecting a sparse combination of the experts, instead of all experts, to process input data, significantly reducing the computational cost and improving the training stability. To encourage specialization and decrease redundancy among experts (Chen et al., 2022), Dai et al. (2022) predefined the expert assignment for different input categories, and Hazimeh et al. (2021) advocated multiple, diverse router policies, facilitating the intriguing goals of SMoE is to divide and conquer the learning task by solving each piece of the task with adaptively selected experts (Aoki et al., 2022; Mittal et al., 2022). To identify similar information patterns between different modalities and extract them with the same expert model, LIFTED follows the design of Sparse Mixture-of-Experts (Shazeer et al., 2017), routing inputs to a subset of experts, dynamically selecting the experts instead of using a pre-defined assignment.

5 Conclusion

We introduce LIFTED, an approach that unifies multimodal data using natural language descriptions and integrates this information within a Mixture-of-Experts (MoE) framework for clinical trial outcome prediction. We employ noise-resilient encoders to extract representations from each modality, utilize a Sparse MoE framework to further dig the similar information patterns in different modalities, and introduce an auxiliary loss to improve the quality of modal-specific representations. Empirically, our LIFTED method demonstrates superior performance compared to existing approaches across all three phases of clinical trials, underscoring the effectiveness and potent representational capacity of natural language and highlighting the potential for a unified text modality to supplant diverse information modalities.

Limitations

Though LIFTED outperforms the existing models, there still are several limitations to our work. As we mentioned in Section 2.1, different modality data is transformed into natural language descriptions by LLM, specifically, the GPT-4o. However, the output from LLM is unstable, which may affect the performance of LIFTED. In addition, since GPT-4o was not trained with clinical trial records specifically, LIFTED may achieve a better performance coupled with the natural language descriptions generated by a new LLM, which is specifically designed for clinical trial outcome prediction.

References

Raquel Aoki, Frederick Tung, and Gabriel L Oliveira. 2022. Heterogeneous multi-task learning with expert diversity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3093–3102.

Yi-Tan Chang, Eric P Hoffman, Guoqiang Yu, David M Herrington, Robert Clarke, Chiung-Ting Wu, Lulu Chen, and Yue Wang. 2019. Integrated identification of disease specific pathways using multi-omics data. *bioRxiv*, page 666065.

Jintai Chen, Yaojun Hu, Yue Wang, Xu Cao, Miao Lin, Hongxia Xu, Jian Wu, Cao Xiao, Jimeng Sun, and 1 others. 2024a. Trialbench: Multi-modal artificial intelligence-ready clinical trial datasets. *arXiv preprint arXiv:2407.00631*.

Jintai Chen, Jiahuan Yan, Danny Ziyi Chen, and Jian Wu. 2023. Excelformer: A neural network surpassing gbdts on tabular data. *arXiv preprint arXiv:2301.02819*.

Tianyi Chen, Nan Hao, Yingzhou Lu, and Capucine Van Rechem. 2024b. Uncertainty quantification on clinical trial outcome prediction. *arXiv preprint arXiv:2401.03482*.

Tianyu Chen, Shaohan Huang, Yuan Xie, Binxing Jiao, Daxin Jiang, Haoyi Zhou, Jianxin Li, and Furu Wei. 2022. Task-specific expert pruning for sparse mixture-of-experts. *arXiv preprint arXiv:2206.00277*.

Yong Dai, Duyu Tang, Liangxin Liu, Minghuan Tan, Cong Zhou, Jingquan Wang, Zhangyin Feng, Fan Zhang, Xueyu Hu, and Shuming Shi. 2022. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *arXiv preprint arXiv:2205.06126*.

Z. Fan, L. Wang, H. Jiang, Y. Lin, and Z. Wang. 2020. Platelet dysfunction and its role in the pathogenesis of psoriasis. *Dermatology*, page 1 – 10. Cited by: 1.

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

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

Tianfan Fu, Cao Xiao, Cheng Qian, Lucas M Glass, and Jimeng Sun. 2021. Probabilistic and dynamic molecule-disease interaction modeling for drug discovery. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 404–414.

Chufan Gao, Jathurshan Pradeepkumar, Trisha Das, Shivashankar Thati, and Jimeng Sun. 2024. Automatically labeling $200b life-saving datasets: A large clinical trial outcome benchmark. *Preprint*, arXiv:2406.10292.

Junyi Gao, Cao Xiao, Lucas M. Glass, and Jimeng Sun. 2020. Compose: Cross-modal pseudo-siamese network for patient trial matching. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 803–812, New York, NY, USA. Association for Computing Machinery.

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

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. 2021. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems*, 34:29335–29347.

Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. 2020. Deeppurpose: A deep learning library for drug-target interaction prediction. *Bioinformatics*, 36(22-23):5545 – 5547. Cited by: 128; All Open Access, Green Open Access, Hybrid Gold Open Access.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Susan Jin, Richard Pazdur, and Rajeshwari Sridhara. 2017. Re-evaluating eligibility criteria for oncology clinical trials: analysis of investigational new drug applications in 2015. *Journal of clinical oncology*, 35(33):3745.

M.I. Jordan and R.A. Jacobs. 1993. Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pages 1339–1344 vol.2.

Wen-Sheng Lo, Hong-Wen Chiou, Shih-Chieh Hsu, Yu-Min Lee, and Liang-Chia Cheng. 2019. Learning based mesh generation for thermal simulation in handheld devices with variable power consumption. In *2019 18th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 7–12.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yingzhou Lu. 2018. *Multi-omics Data Integration for Identifying Disease Specific Biological Pathways*. Ph.D. thesis, Virginia Tech.

Linda Martin, Melissa Hutchens, Conrad Hawkins, and Alaina Radnov. 2017. How much do clinical trials cost? *Nature Reviews Drug Discovery*, 16(6):381–382.

Sarthak Mittal, Yoshua Bengio, and Guillaume Lajoie. 2022. Is a modular architecture enough? *Advances in Neural Information Processing Systems*, 35:28747–28760.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Youran Qi and Qi Tang. 2019. Predicting phase 3 clinical trial results by modeling phase 2 clinical trial subject level data using deep learning. In *Machine Learning for Healthcare Conference*, pages 288–303. PMLR.

Pranav Rajpurkar, Jingbo Yang, Nathan Dass, Vinjai Vale, Arielle S. Keller, Jeremy Irvin, Zachary Taylor, Sanjay Basu, Andrew Ng, and Leanne M. Williams. 2020. Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: A prespecified secondary analysis of a randomized clinical trial. *JAMA Network Open*, 3(6):e206653–e206653.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Minjie Shen, Yue Zhao, Chenhao Li, Fan Meng, Xiao Wang, David Herrington, Yue Wang, Tim Fu, and Capucine Van Rechem. 2023. Genocraft: A comprehensive, user-friendly web-based platform for high-throughput omics data analysis and visualization. *arXiv preprint arXiv:2312.14249*.

Kien Wei Siah, Nicholas W. Kelley, Steffen Ballerstedt, Björn Holzhauer, Tianmeng Lyu, David Mettler, Sophie Sun, Simon Wandel, Yang Zhong, Bin Zhou, Shifeng Pan, Yingyao Zhou, and Andrew W. Lo. 2021. Predicting drug approvals: The novartis data science and artificial intelligence challenge. *Patterns*, 2(8):100312.

Léon-Charles Tranchevent, Francisco Azuaje, and Jagath C Rajapakse. 2019. A deep neural network approach to predicting clinical outcomes of neuroblastoma patients. *BMC medical genomics*, 12:1–11.

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

Zifeng Wang, Chufan Gao, Cao Xiao, and Jimeng Sun. 2023a. Anypredict: Foundation model for tabular prediction. *Preprint*, arxiv:2305.12081.

Zifeng Wang, Cao Xiao, and Jimeng Sun. 2023b. Spot: Sequential predictive modeling of clinical trial outcome with meta-learning. *Preprint*, arxiv:2304.05352.

Chiung-Ting Wu, Sarah J Parker, Zuolin Cheng, Georgia Saylor, Jennifer E Van Eyk, Guoqiang Yu, Robert Clarke, David M Herrington, and Yue Wang. 2022a. Cot: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinformatics Advances*, 2(1):vbac037.

Chiung-Ting Wu, Minjie Shen, Dongping Du, Zuolin Cheng, Sarah J Parker, Jennifer E Van Eyk, Guoqiang Yu, Robert Clarke, David M Herrington, and 1 others. 2022b. Cosbin: cosine score-based iterative normalization of biologically diverse samples. *Bioinformatics Advances*, 2(1):vbac076.

Jiahuan Yan, Jintai Chen, Yixuan Wu, Danny Z Chen, and Jian Wu. 2023. T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10720–10728.

Steven Yi, Adam Yee, John Harmon, Frank Meng, and Saurabh Hinduja. 2018. Enhance wound healing monitoring through a thermal imaging based smartphone app. In *Medical imaging 2018: Imaging informatics for healthcare, research, and applications*, volume 10579, pages 438–441. SPIE.

Bai Zhang, Yi Fu, Zhen Zhang, Robert Clarke, Jennifer E Van Eyk, David M Herrington, and Yue Wang. 2021. Ddn2. 0: R and python packages for differential dependency network analysis of biological systems. *bioRxiv*, pages 2021–04.

Xingyao Zhang, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2020. Deepenroll: patient-trial matching with deep embedding and entailment prediction. In *Proceedings of the web conference 2020*, pages 1029–1037.

A Prompt

The whole prompt, including the system message, is demonstrated in Table 4, and some examples are demonstrated in Table 5.

Table 4: Prompting.

System Message
You are a helpful assistant.

Prompting
Here is the schema definition of the table: $schema_definition This is a sample from the table: $linearization Please briefly summarize the sample with its value in one sentence. You should describe the important values, like drugs and diseases, instead of just the names of columns in the table. A brief summarization of another sample may look like: This study will test the ability of extended-release nifedipine (Procardia XL), a blood pressure medication, to permit a decrease in the dose of glucocorticoid medication children take to treat congenital adrenal hyperplasia (CAH). Note that the example is not the summarization of the sample you have to summarize.

Response
$summarization_of_the_sample

B Algorithm of LIFTED's training pipeline

The whole training pipeline of LIFTED is illustrated as algorithm 1.

C Baselines

Many methods have been selected as baselines in our experiments, including both statistical machine learning and deep learning models. We use the same setups in Fu *et al.* (2022) and Wang *et al.* (2023b) for most of them.

- **Logistic regression (LR)** (Lo et al., 2019; Siah et al., 2021): logistic regression with the default hyperparameters implemented by scikit-learn (Pedregosa et al., 2011).

Algorithm 1 Training Pipeline of LIFTED

Input : Training dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$

step 1. Transforming Multimodal Data into Natural Language Descriptions **for** $i \leftarrow 1$ **to** N **do**

  for $k \leftarrow 1$ **to** K **do**

    $x_{i,k} \leftarrow \text{LLM}(p, \text{Linearize}(x_{i,k}), s)$

  $x_{i,0} \leftarrow \text{LLM}(p, \text{Linearize}(x_i), s)$

step 2. Train LIFTED **foreach** *minibatch* $\mathcal{B}$ *in dataset* $\mathcal{D}$ **do**

  for $i \leftarrow 1$ **to** *batch size* b **do**

    for $k \leftarrow 0$ **to** K **do**

      $u_{i,k}^t \leftarrow \mathcal{E}_k(\mathcal{T}(z_{i,k}))$

      $v_{i,k}^t \leftarrow \hat{u}_{i,k}^t$ as Equation 7

      $U_{i,k} \leftarrow \mathcal{F}_k(\{u_{i,k}^t + pos^t\}_{t=0}^T)$,

      $\tilde{U}_{i,k} \leftarrow \mathcal{G}(U_{i,k}) \cdot \mathcal{R}(U_{i,k})$,

      $V_{i,k} \leftarrow \mathcal{F}_k(\{v_{i,k}^t + pos^t\}_{t=0}^T)$,

      $\tilde{V}_{i,k} \leftarrow \mathcal{G}(V_{i,k}) \cdot \mathcal{R}(V_{i,k})$

    Fuse representations with MoE method as (9)

  Compute the losses $\mathcal{L}_{con}$, $\mathcal{L}_c$, $\mathcal{L}_{aux}$ and $\mathcal{L}$

  Optimize the parameters θ of LIFTED

- **Random Forest (RF)** (Lo et al., 2019; Siah et al., 2021): similar to logistic regression, the random forest is also implemented by scikit-learn with the default hyperparameters (Pedregosa et al., 2011).

- **XGBoost** (Rajpurkar et al., 2020; Siah et al., 2021): An implementation of gradient-boosted decision trees optimized for speed and performance.

- **Adaptive boosting (AdaBoost)** (Fan et al., 2020): an adaptive boosting-based decision tree method implemented by scikit-learn (Pedregosa et al., 2011).

- **k Nearest Neighbor (kNN) + RF** (Lo et al., 2019): a combined model using kNN to impute missing data and predicting by random forests.

- **Feedforward Neural Network (FFNN)** (Tranchevent et al., 2019): a feedforward neural network that uses the same feature as HINT (Fu et al., 2023). The FFNN contains three fully-connected layers with hidden dimensions of 500 and 100, as well as a rectified linear unit (ReLU) activation layer to provide nonlinearity.

- **Multi-Modal Fusion (MMF)**: This technique amalgamates multi-modal data to arrive at a final prediction, employing both early fusion and late

7513

Table 5: Examples of Prompting.

Linearization	Summarization
phase: phase 1/phase 2; diseases: ['adenocarcinoma of the lung', 'non-small cell lung cancer']; icdcodes: ["['D02.20', 'D02.21', 'D02.22']", "['C78.00', 'C78.01', 'C78.02', 'D14.30', 'D14.31', 'D14.32', 'C34.2']"]; drugs: ['erlotinib hydrochloride', 'hsp90 inhibitor auy922']; criteria: \n Inclusion Criteria:\n - All patients must have pathologic evidence of advanced lung adenocarcinoma (stage II-IBor stage IV) confirmed histologically/cytologically at NU, MSKCC, or DFCI and EITHER previous RECIST-defined response . . .	This sample is a phase 1/phase 2 trial for patients with advanced lung adenocarcinoma, testing the efficacy of erlotinib hydrochloride and hsp90 inhibitor auy922 in patients who have previously responded to erlotinib or gefitinib or have a documented mutation in the EGFR gene. The study has specific inclusion and exclusion criteria, and patients must meet certain medical conditions and have negative pregnancy tests to be eligible.
phase: phase 2; diseases: ['multiple myeloma']; icdcodes: ["['C90.01', 'C90.02', 'C90.00']"]; drugs: ['dexamethasone', 'thalidomide', 'lenalidomide']; criteria: \n Inclusion Criteria:\n\n - Subject must voluntarily sign and understand written informed consent.\n\n - Age > 18 years at the time of signing the consent form.\n\n - Histologically confirmed Salmon-Durie stage II or III MM. Stage I MM patients will be\n eligible if they display poor prognostic factors (ß2M $\geq$ 5.5 mg/L, plasma cell\n proliferation index $\geq$ 5%, albumin of less than 3.0, and unfavorable cytogenetics). . . .	This sample is a phase 2 clinical trial for patients with relapsed or refractory multiple myeloma, testing the combination of dexamethasone, thalidomide, and lenalidomide as a treatment option. The eligibility criteria include specific disease stage, prior treatment history, and certain laboratory parameters. Exclusion criteria include non-secretory MM, prior history of other malignancies, and certain medical conditions.
phase: phase 3; diseases: ["Alzheimer's disease"]; icdcodes: ["['G30.8', 'G30.9', 'G30.0', 'G30.1']"]; drugs: ['rivastigmine 5 cm^2 transdermal patch', 'rivastigmine 10 cm^2 transdermal patch']; criteria: \n Inclusion Criteria:\n\n - Be at least 50 years of age;\n\n - Have a diagnosis of probable Alzheimer's Disease; \n\n - Have an MMSE score of $\geq$ 10 and $\leq$ 24;\n\n - Must have a caregiver who is able to attend all study visits;\n\n - Have received continuous treatment with donepezil for at least 6 months prior to\n screening, and received a stable dose of 5 mg/day or 10 mg/day for at least the last 3\n of these 6 months.\n\n . . .	This sample is a phase 3 clinical trial for Alzheimer's disease, testing the efficacy of rivastigmine transdermal patches in patients aged 50 and above with a diagnosis of probable Alzheimer's disease and an MMSE score between 10 and 24. The inclusion criteria also require patients to have a caregiver who can attend all study visits and have received continuous treatment with donepezil for at least 6 months prior to screening. The exclusion criteria include various medical conditions and disabilities that may interfere with the study.

fusion strategies. In the early fusion approach, various modal inputs are first concatenated before being fed into the prediction model. Conversely, in the late fusion variant, multiple prediction models are employed on each modal input, and the ultimate prediction is derived through fusion techniques, such as voting, which integrates predictions from each modality.

- **HINT** (Fu et al., 2022): several key components are integrated with HINT, including a drug molecule encoder utilizing MPNN algorithm, a disease ontology encoder based on GRAM, a trial eligibility criteria encoder leveraging BERT, and also, a drug molecule pharmacokinetic encoder, surplus a graph neural network to capture feature interactions. After the interacted features are encoded, they are fed into a prediction model for accurate outcome predictions.

- **SPOT** (Wang et al., 2023b): SPOT contains several steps. Firstly, trial topics are identified to group the diverse trial data from multiple sources into relevant trial topics. Subsequently, trial embeddings are produced and organized according to topic and timestamp to construct organized clinical trial sequences. Finally, each trial sequence is treated as a separate task, and a meta-learning approach is employed to adapt to new tasks with minimal modifications.

D Dataset Descriptions

Table 6: The statistics of the HINT Dataset and the CTOD Dataset. The number of Trials is shown by the split of train/validation/test sets.

	# Trials	# Drugs	# Diseases	# Success	# Failure
HINT					
Phase I	1,044/116/627	2,020	1,392	1,006	781
Phase II	4,004/445/1,653	5,610	2,824	3,039	3,063
Phase III	3,092/344/1,140	4,727	1,619	3,104	1,472
CTOD					
Phase I	1,012/149/627	1,638	913	1,129	659
Phase II	3,950/501/1,653	5,003	2,254	3,949	2,156
Phase III	3,075/363/1,140	3,863	1,533	3,643	941

The HINT dataset (Fu et al., 2022; Chen et al., 2024b) and CTOD dataset (Gao et al., 2024) include the information on diseases, the name, description, and SMILES string of drugs, eligibility criteria for each clinical trial record, the phase, and also, the trial outcome labels as success or failure covering Phases I, II and III trials. The HINT dataset contains 17,538 clinical trial records, with

1,787 trials in Phase I, 6,102 trials in Phase II, and 4,576 trials in Phase III (Fu et al., 2021). We utilize parts of the CTOD dataset to test LIFTED's performance, with 1,788 trials in Phase I, 6,105 trials in Phase II, and 4,584 trials in Phase III, for a total of 12,477 trials. (Gao et al., 2024). The detailed data statistics of the HINT dataset and the CTOD dataset are shown in Table 6.

In our implementation, we incorporate all modalities, including disease, the name, description and SMILES string of drugs and criteria, totaling five modalities. Additionally, we include phase information when generating the natural language summarization for samples. The transform-based encoder for the SMILES string modality is pretrained and the corresponding tokenizer is specifically designed for SMILES string data. However, all the other modalities are tokenized by a unified tokenizer and none of the other encoders are pretrained.

E Hyperparameter Settings

We follow the settings of most hyperparameters in HINT (Fu et al., 2022). The models are trained for a total of 5 epochs using a mini-batch size of 32 on one NVIDIA 4090 GPUs, which will take up to 2 hours. We employ the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 3×10^{-4}, β values of $(0.9, 0.99)$, and a weight decay of 1×10^{-2} with a CosineAnnealing learning rate scheduler. Besides this, we leverage the GPT-4o as the LLM model to generate natural language description for our data during our prompting stage.

F Detailed Results

The detailed experiments results with variants are presented in Table 7, 8, 9

Table 7: The clinical trial outcome performance (%) of LIFTED and baselines on HINT dataset and CTOD dataset. The mean and standard deviations are calculated from 30 independent runs with different random seeds. †: The results of the HINT and SPOT methods were obtained by running their released codes. The best results and second best results are **bold** and <u>underlined</u>, respectively. We observe that LIFTED consistently outperforms all other methods over all three phases.

HINT

Method	Phase I Trials			Phase II Trials			Phase III Trials		
	PR-AUC	F1	ROC-AUC	PR-AUC	F1	ROC-AUC	PR-AUC	F1	ROC-AUC
LR	50.0 ± 0.5	60.4 ± 0.5	52.0 ± 0.6	56.5 ± 0.5	55.5 ± 0.6	58.7 ± 0.9	68.7 ± 0.5	69.8 ± 0.5	65.0 ± 0.7
RF	51.8 ± 0.5	62.1 ± 0.5	52.5 ± 0.6	57.8 ± 0.8	56.3 ± 0.9	58.8 ± 0.9	69.2 ± 0.4	68.6 ± 1.0	66.3 ± 0.7
XGBoost	51.3 ± 6.0	62.1 ± 0.7	51.8 ± 0.6	58.6 ± 0.6	57.0 ± 0.9	60.0 ± 0.7	69.7 ± 0.7	69.6 ± 0.5	66.7 ± 0.5
AdaBoost	51.9 ± 0.5	62.2 ± 0.7	52.6 ± 0.6	58.6 ± 0.9	58.3 ± 0.8	60.3 ± 0.7	70.1 ± 0.5	69.5 ± 0.5	67.0 ± 0.4
kNN+RF	53.1 ± 0.6	62.5 ± 0.7	53.8 ± 0.5	59.4 ± 0.8	59.0 ± 0.6	59.7 ± 0.8	70.7 ± 0.7	69.8 ± 0.8	67.8 ± 1.0
FFNN	54.7 ± 1.0	63.4 ± 1.5	55.0 ± 1.0	60.4 ± 1.0	59.9 ± 1.2	61.1 ± 1.1	74.7 ± 1.1	74.8 ± 0.9	68.1 ± 0.8
MMF (early fusion)	60.6 ± 2.8	59.4 ± 2.3	54.4 ± 2.4	60.2 ± 1.9	62.6 ± 1.4	60.7 ± 1.3	85.5 ± 1.4	81.5 ± 0.9	70.6 ± 1.7
MMF (late fusion)	63.4 ± 3.0	67.5 ± 2.2	59.0 ± 2.8	<u>62.9 ± 2.0</u>	63.0 ± 1.5	62.6 ± 1.6	<u>86.9 ± 1.6</u>	<u>83.1 ± 1.1</u>	<u>71.8 ± 2.2</u>
DeepEnroll	56.8 ± 0.7	64.8 ± 1.1	57.5 ± 1.3	60.0 ± 1.0	59.8 ± 0.7	62.5 ± 0.8	77.7 ± 0.8	78.6 ± 0.7	69.9 ± 0.8
COMPOSE	56.4 ± 0.7	65.8 ± 0.9	57.1 ± 1.1	60.4 ± 0.7	59.7 ± 0.6	62.8 ± 0.9	78.2 ± 0.8	79.2 ± 0.7	70.0 ± 0.7
HINT†	58.4 ± 2.3	68.2 ± 1.7	62.1 ± 2.2	59.1 ± 1.2	63.9 ± 1.2	62.8 ± 1.4	85.9 ± 1.1	80.9 ± 0.8	70.8 ± 1.3
SPOT†	<u>69.8 ± 1.7</u>	<u>68.4 ± 1.2</u>	<u>64.6 ± 2.1</u>	62.6 ± 0.7	<u>64.3 ± 0.6</u>	<u>63.0 ± 0.6</u>	81.7 ± 0.8	81.0 ± 0.4	71.0 ± 0.4
LIFTED (ours)	**70.7 ± 2.3**	**71.6 ± 1.4**	**64.9 ± 2.1**	**69.8 ± 1.8**	**66.2 ± 1.1**	**65.1 ± 1.4**	**88.3 ± 1.1**	**83.8 ± 0.8**	**73.5 ± 1.6**

CTOD

Method	Phase I Trials			Phase II Trials			Phase III Trials		
	PR-AUC	F1	ROC-AUC	PR-AUC	F1	ROC-AUC	PR-AUC	F1	ROC-AUC
LR	55.6 ± 2.3	58.5 ± 2.1	51.5 ± 2.2	56.0 ± 1.4	63.6 ± 1.1	56.7 ± 1.0	75.0 ± 2.2	73.8 ± 1.6	51.9 ± 2.0
RF	58.3 ± 2.9	70.0 ± 1.6	54.0 ± 2.0	60.8 ± 1.3	70.9 ± 0.8	56.0 ± 1.3	75.1 ± 1.6	85.1 ± 0.7	50.7 ± 2.2
XGBoost	57.4 ± 3.5	68.0 ± 1.9	50.5 ± 2.3	61.2 ± 1.5	70.6 ± 1.2	61.3 ± 1.3	78.7 ± 1.8	84.9 ± 0.7	57.1 ± 2.0
AdaBoost	58.8 ± 3.2	66.7 ± 2.4	53.5 ± 2.3	60.2 ± 1.8	68.9 ± 0.8	55.2 ± 1.3	79.5 ± 1.3	84.6 ± 0.8	56.9 ± 1.9
kNN+RF	58.4 ± 3.9	70.3 ± 2.0	58.7 ± 2.1	61.2 ± 1.7	70.9 ± 0.9	56.6 ± 1.5	76.8 ± 2.2	85.2 ± 1.1	52.5 ± 1.9
FFNN	55.2 ± 2.2	58.4 ± 1.4	48.7 ± 1.7	57.2 ± 1.5	66.0 ± 1.3	53.0 ± 1.2	76.1 ± 1.4	79.6 ± 1.1	52.6 ± 1.5
MMF (early fusion)	61.2 ± 2.8	64.2 ± 2.1	56.2 ± 2.8	64.0 ± 1.2	70.1 ± 0.9	59.0 ± 1.1	83.3 ± 1.4	84.4 ± 1.0	64.6 ± 1.9
MMF (late fusion)	63.0 ± 3.1	70.5 ± 1.8	57.4 ± 2.4	<u>65.7 ± 2.1</u>	71.4 ± 0.8	60.3 ± 2.6	<u>83.5 ± 1.7</u>	85.2 ± 0.8	65.8 ± 1.6
HINT†	63.4 ± 2.9	<u>71.0 ± 1.2</u>	57.6 ± 2.4	64.6 ± 2.4	71.2 ± 1.3	<u>60.6 ± 2.0</u>	81.8 ± 1.9	<u>85.7 ± 0.7</u>	60.8 ± 2.0
SPOT†	<u>66.8 ± 1.9</u>	70.0 ± 0.9	<u>62.3 ± 1.0</u>	64.5 ± 1.7	<u>71.8 ± 0.7</u>	58.8 ± 1.4	83.2 ± 2.3	77.6 ± 1.0	<u>67.5 ± 1.7</u>
LIFTED (ours)	**69.7 ± 3.0**	**71.8 ± 1.4**	**63.4 ± 2.3**	**67.7 ± 2.0**	**72.0 ± 1.4**	**63.0 ± 1.2**	**86.7 ± 1.2**	**85.9 ± 1.0**	**69.5 ± 1.8**

Table 8: The clinical trial outcome prediction performance (%) of LIFTED and variants without certain key component. The best results are **bold**. LIFTED outperforms all variants, showcasing the effectiveness of our proposed components.

HINT

Method	Phase I Trials			Phase II Trials			Phase III Trials		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
LIFTED-aug	68.4 ± 2.0	69.8 ± 2.1	64.8 ± 1.5	69.5 ± 1.4	66.0 ± 1.1	64.3 ± 0.8	86.9 ± 1.7	82.4 ± 0.9	72.1 ± 1.6
LIFTED-aux	69.0 ± 2.9	71.2 ± 1.7	63.7 ± 1.6	69.6 ± 1.6	64.5 ± 1.5	64.6 ± 1.3	87.4 ± 1.4	82.8 ± 1.0	71.1 ± 2.2
LIFTED-LLM	68.5 ± 2.7	70.8 ± 1.3	64.0 ± 2.3	69.7 ± 2.0	64.9 ± 1.4	65.0 ± 1.5	86.7 ± 1.0	82.7 ± 1.0	70.8 ± 1.3
LIFTED-gating	69.9 ± 2.3	71.3 ± 1.8	64.9 ± 1.9	69.7 ± 1.7	65.5 ± 1.4	65.0 ± 1.6	87.0 ± 0.8	82.7 ± 0.8	72.4 ± 1.1
LIFTED (ours)	**70.7 ± 2.3**	**71.6 ± 1.4**	**64.9 ± 2.1**	**69.8 ± 1.8**	**66.2 ± 1.1**	**65.1 ± 1.4**	**88.3 ± 1.1**	**83.8 ± 0.8**	**73.5 ± 1.6**

CTOD

Method	Phase I Trials			Phase II Trials			Phase III Trials		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
LIFTED-aug	65.1 ± 3.7	70.6 ± 1.7	61.9 ± 2.4	67.1 ± 1.8	70.5 ± 1.1	62.2 ± 1.4	85.2 ± 1.4	82.4 ± 1.1	67.8 ± 1.6
LIFTED-aux	64.4 ± 3.3	71.1 ± 1.8	60.0 ± 2.8	60.0 ± 1.4	71.0 ± 0.9	54.9 ± 1.0	83.8 ± 1.8	85.5 ± 0.9	64.6 ± 2.4
LIFTED-LLM	65.1 ± 2.7	71.4 ± 1.6	58.5 ± 3.2	66.3 ± 1.9	71.1 ± 1.1	61.2 ± 1.5	83.9 ± 1.3	85.4 ± 1.0	65.3 ± 2.0
LIFTED-gating	67.1 ± 2.4	71.4 ± 1.8	60.8 ± 2.5	65.1 ± 1.4	70.4 ± 1.3	61.9 ± 1.5	85.0 ± 1.5	85.8 ± 0.7	68.8 ± 1.8
LIFTED (ours)	**69.7 ± 3.0**	**71.8 ± 1.4**	**63.4 ± 2.3**	**67.7 ± 2.0**	**72.0 ± 1.4**	**63.0 ± 1.2**	**86.7 ± 1.2**	**85.9 ± 1.0**	**69.5 ± 1.8**

Table 9: Performance analysis of multimodal data integration. The best results and second best results are **bold** and <u>underlined</u>, respectively.

	Phase I Trials			Phase II Trials			Phase III Trials		
HINT									
	PR-AUC	**F1**	**ROC-AUC**	**PR-AUC**	**F1**	**ROC-AUC**	**PR-AUC**	**F1**	**ROC-AUC**
Summarization	63.2 ± 2.4	69.9 ± 2.2	57.8 ± 2.2	66.1 ± 1.3	61.2 ± 1.5	61.2 ± 1.0	85.1 ± 1.0	80.7 ± 1.0	66.6 ± 1.6
Drugs	62.1 ± 1.2	67.2 ± 1.5	57.9 ± 1.1	60.5 ± 1.1	62.3 ± 1.5	55.8 ± 1.1	83.8 ± 0.7	81.7 ± 1.0	63.8 ± 1.3
Disease	65.3 ± 1.1	67.3 ± 1.8	59.7 ± 1.3	$\underline{68.0 \pm 0.5}$	59.9 ± 1.3	62.4 ± 0.6	$\underline{86.0 \pm 0.8}$	80.5 ± 1.1	$\underline{69.1 \pm 0.9}$
Description	55.5 ± 0.5	$\underline{71.3 \pm 0.1}$	50.2 ± 1.0	55.5 ± 0.7	0.0 ± 0.0	50.0 ± 1.3	74.9 ± 0.6	$\mathbf{85.7 \pm 0.0}$	49.7 ± 1.3
SMILES	62.8 ± 0.7	69.6 ± 1.7	58.5 ± 0.9	59.3 ± 0.6	58.3 ± 2.2	54.9 ± 0.7	76.1 ± 1.5	83.6 ± 0.5	51.1 ± 2.5
Criteria	$\underline{68.0 \pm 3.0}$	70.5 ± 1.9	63.1 ± 2.1	67.6 ± 1.1	$\underline{64.4 \pm 1.0}$	$\underline{63.0 \pm 1.3}$	83.7 ± 1.2	82.7 ± 0.7	65.0 ± 2.1
All (LIFTED)	$\mathbf{70.7 \pm 2.3}$	$\mathbf{71.6 \pm 1.4}$	$\mathbf{64.9 \pm 2.1}$	$\mathbf{69.8 \pm 1.8}$	$\mathbf{66.2 \pm 1.1}$	$\mathbf{65.1 \pm 1.4}$	$\mathbf{88.3 \pm 1.1}$	$\underline{83.8 \pm 0.8}$	$\mathbf{73.5 \pm 1.6}$
CTOD									
	PR-AUC	**F1**	**ROC-AUC**	**PR-AUC**	**F1**	**ROC-AUC**	**PR-AUC**	**F1**	**ROC-AUC**
Summarization	61.6 ± 3.2	70.9 ± 2.2	56.9 ± 2.6	65.4 ± 1.3	$\underline{71.5 \pm 0.9}$	61.1 ± 1.3	84.1 ± 1.3	84.8 ± 0.9	64.9 ± 1.9
Drugs	60.8 ± 2.8	70.6 ± 1.4	57.6 ± 2.4	58.0 ± 1.8	71.3 ± 1.1	53.8 ± 1.2	77.3 ± 1.7	85.2 ± 1.0	54.5 ± 1.9
Disease	65.2 ± 2.0	70.2 ± 1.9	61.3 ± 1.7	68.3 ± 1.5	71.1 ± 1.2	62.6 ± 1.0	85.3 ± 1.4	85.5 ± 0.7	68.4 ± 2.0
Description	56.6 ± 2.6	70.7 ± 1.5	52.3 ± 2.7	56.0 ± 1.8	71.3 ± 0.8	52.4 ± 1.7	76.5 ± 1.2	$\underline{85.7 \pm 0.7}$	52.9 ± 1.9
SMILES	61.6 ± 2.8	70.6 ± 1.7	56.0 ± 2.4	58.0 ± 1.8	$\underline{71.5 \pm 1.2}$	53.1 ± 1.2	75.1 ± 1.6	85.6 ± 0.7	50.8 ± 2.4
Criteria	60.5 ± 3.0	$\underline{71.2 \pm 1.7}$	55.8 ± 3.1	59.9 ± 2.5	71.2 ± 1.2	53.9 ± 2.0	75.7 ± 2.0	84.6 ± 0.7	52.5 ± 2.6
LIFTED (ours)	$\mathbf{69.7 \pm 3.0}$	$\mathbf{71.8 \pm 1.4}$	$\mathbf{63.4 \pm 2.3}$	$\mathbf{67.7 \pm 2.0}$	$\mathbf{72.0 \pm 1.4}$	$\mathbf{63.0 \pm 1.2}$	$\mathbf{86.7 \pm 1.2}$	$\mathbf{85.9 \pm 1.0}$	$\mathbf{69.5 \pm 1.8}$