LIFTED: Multimodal Mixture-of-Experts for Clinical Trial Outcome Prediction

Wenhao Zheng [1] **Dongsheng Peng** [1] **Hongxia Xu** [2] **Yun Li** [1] **Hongtu Zhu** [1] **Tianfan Fu** [3] **Huaxiu Yao** [1]

1. Introduction

The clinical trial is a crucial step in the development of new treatments to demonstrate the safety and efficacy of the drug. 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 (Blass, 2015; Huang et al., 2020a). 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 & Tang, 2019)). Recently, deep learning methods have been proposed for trial outcome predictions. However, those approachs rely 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 this, we aim to design a unified encoder to extract representations from various modalities, but it poses the following three 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 a graph, while disease names rely on relationships between different diseases.

- **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 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.

- **How to integrate extracted information from different modalities?** Extracted representations from various modalities need to be integrated for predictions. However, the contribution of extracted information from different modalities may vary significantly between samples. For instance, in one patient, a specific disease, such as type 2 diabetes mellitus, which is difficult to treat, may strongly influence the final outcome (Wu et al., 2022). In contrast, another patient's trial result may be primarily determined by the drugs they are prescribed, particularly if those medications have a high success rate in treating the disease.

To address those challenges, we propose an approach called mu**L**ti-modal m**I**x-of-experts **F**or ou**T**come pr**ED**iction (**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 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. 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 benchamark (Fu et al., 2022) to demonstrate the effectiveness of LIFTED and the effectiveness of our proposed components.

[1]UNC-Chapel Hill [2]Zhejiang University [3]Rensselaer Polytechnic Institute. Correspondence to: Huaxiu Yao <huaxiu@cs.unc.edu>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

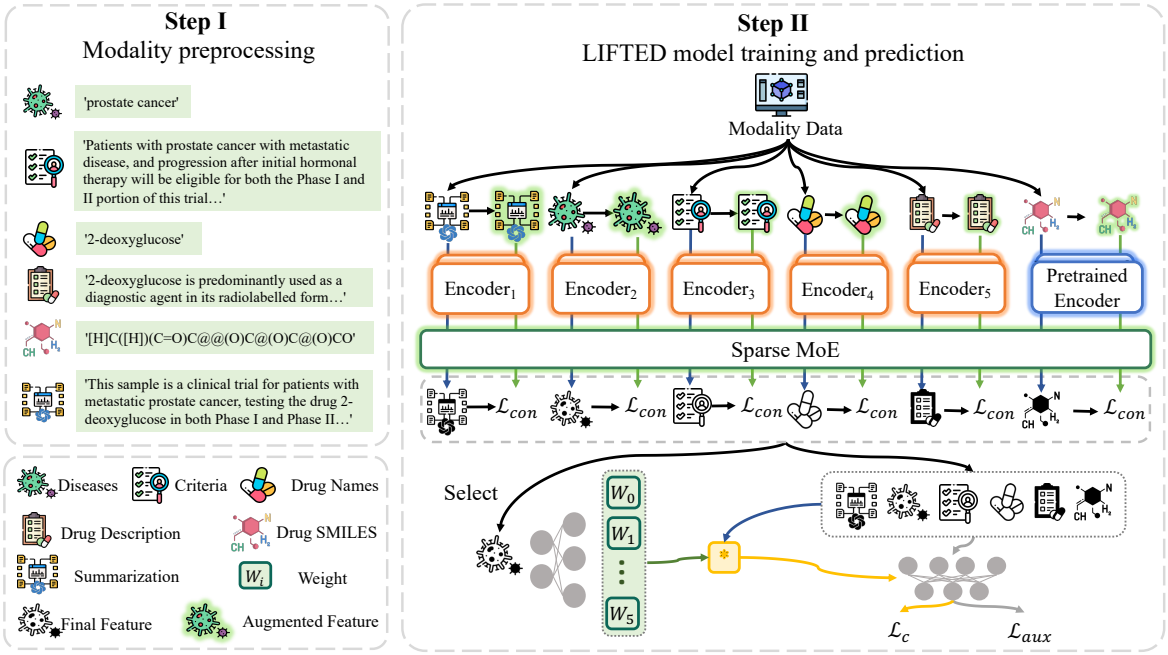

Figure 1. An overview of LIFTED. **Step 1**: Transforming multimodal data into natural language descriptions, where all modalities are converted into natural language descriptions to facilitate the representation extraction process of the transformer encoders. **Step 2**: Extract and combine representations 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.

2. Multimodal Mixture-of-Experts for Outcome Prediction

2.1. Overview

This section presents our proposed mu**L**ti-modal m**I**x-of-experts **F**or ou**T**come pr**ED**iction (**LIFTED**) method. The goal of LIFTED is to unify multimodal data using natural language descriptions and integrate this information within a Mixture-of-Experts (MoE) framework, as illustrated in Figure 1. To elaborate, we start by extracting specific modalities from the clinical trial dataset, subsequently transforming this multimodal data into natural language descriptions using a Large Language Model (LLM). Following this, we augment the embeddings of the language descriptions derived from these different modalities. We then feed both the original and augmented embeddings into transformer-based encoders for representation learning. Subsequently, an SMoE framework is utilized to route the embeddings from different modalities to different sets of experts, where similar information patterns in different modalities will be routed to the same experts while the different patterns will be routed to experts with more specialized knowledge. To enhance the robustness of encoders, we introduce a consistency loss that aligns the original representations with the augmented ones. Moving forward, we implement an MoE framework to integrate these representations for each trial, which originate from various modalities. Finally, these integrated representations are input into a classifier for prediction. Simultaneously, we introduce an auxiliary unimodal prediction loss to improve the quality of modal-specific representations. Below, we detail LIFTED.

2.2. Transforming Multimodal Data into Natural Language Descriptions

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 value is $x_{i,k}$.

Prompting. 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 $z_{i,k}$.

2.3. Representation Learning and Refinement

After transforming multimodal data into natural language descriptions, we build $K+1$ transformer-based encoders on the top on 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:

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

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:

$$\begin{aligned} \mathcal{G}(U_{i,k}) =& \mathrm{Softmax}(\mathrm{TopK}(\mathcal{P}(U_{i,k}), k)) \\ \mathcal{P}(U_{i,k}) =& U_{i,k} \cdot W_g + \mu \mathrm{Softplus}(U_{i,k} \cdot W_{\mathrm{noise}}) \end{aligned} \tag{2}$$

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.

2.4. Representation Augmentation and Consistency Loss

However, building informative modal-specific encoders and the SMoE framework solely from these modal-specific natural language descriptions remains challenging, primarily due to potential data noise introduced during the data collection process. To make the encoders and the SMoE framework more robust to the noise in the data, we augment the embeddings $u_{i,k}^t$ with a minor perturbation to $v_{i,k}^t$ and add a consistency loss to require the encoders and the SMoE framework insensitive to small perturbation, which is detailed as the following two steps:

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. Next, we proceed to sample a small value α_l from a uniform distribution $\mathrm{Uniform}(-\lambda, \lambda)$ for each selected element. Following this, each selected element is multiplied by $\exp(\alpha_l)$ to apply the perturbation, resulting in the perturbed vector $v_{i,k}^t$.

Consistency Loss. These perturbed embeddings $\{v_{i,k}^t\}_{t=1}^T$ 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 defined as the sum of MSE loss between $\tilde{U}_{i,k}$ and $\tilde{V}_{i,k}$.

2.5. Integrating Multimodal Information with Mixture-of-Experts

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. 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. The process can be formulated as follows:

$$\begin{aligned} W_{i,k} =& \mathrm{Softmax}(\mathcal{C}(\oplus_{j \in \mathcal{J}} U_{i,j}) * \gamma_k) \\ U_i =& \sum_{k=0}^{K} W_{i,k} * \tilde{U}_{i,k}, \end{aligned} \tag{3}$$

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 the cross entropy loss between the prediction $\hat{y}_i$ and ground truth label y_i. 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.

3. Experiments

In this section, we evaluate the performance of LIFTEDto compare with the existing methods. More experiments and analysis can be found in Appendix A.

3.1. Dataset Descriptions.

We evaluate our method and other baselines on the HINT dataset (Fu et al., 2022; Chen et al., 2024), which includes 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. 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 pre-trained 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 pre-trained.

3.2. Experimental Setup

Baselines. We compare LIFTED with both machine learning methods, including Logistic regression (LR) (Siah et al., 2021; Lo et al., 2019), Random Forest (RF) (Lo et al., 2019; Siah et al., 2021), XGBoost (Rajpurkar et al., 2020; Siah et al., 2021), Adaptive boosting (AdaBoost) (Fan et al., 2020), k Nearest Neighbor (kNN) + RF (Lo et al., 2019) and deep learning models, such as Feedforward Neural Network (FFNN) (Tranchevent et al., 2019), Multi Modal Fusion (MMF), DeepEnroll (Zhang et al., 2020), COMPOSE (Gao et al., 2020), HINT (Fu et al., 2022; Wang et al., 2024), SPOT (Wang et al., 2023b). Among these, HINT and SPOT are specifically designed for clinical trial outcome prediction. More details are presented in Appendix D.

Evaluation Metrics. Following Fu et al. (2022) and Wang et al. (2023b), we use F1 score, PR-AUC, and ROC-AUC to measure the performance of all methods. For all these three metrics, higher scores indicate better performance.

3.3. Overall Performance

Table 1. The clinical trial outcome performance (%) of LIFTED and baselines. 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 underlined, respectively.

Method	PR-AUC	F1	ROC-AUC
LR	50.0 ± 0.5	60.4 ± 0.5	52.0 ± 0.6
RF	51.8 ± 0.5	62.1 ± 0.5	52.5 ± 0.6
XGBoost	51.3 ± 6.0	62.1 ± 0.7	51.8 ± 0.6
AdaBoost	51.9 ± 0.5	62.2 ± 0.7	52.6 ± 0.6
kNN+RF	53.1 ± 0.6	62.5 ± 0.7	53.8 ± 0.5
FFNN	54.7 ± 1.0	63.4 ± 1.5	55.0 ± 1.0
MMF (early fusion)	60.6 ± 2.8	59.4 ± 2.3	54.4 ± 2.4
MMF (late fusion)	63.4 ± 3.0	67.5 ± 2.2	59.0 ± 2.8
DeepEnroll	56.8 ± 0.7	64.8 ± 1.1	57.5 ± 1.3
COMPOSE	56.4 ± 0.7	65.8 ± 0.9	57.1 ± 1.1
HINT†	58.4 ± 2.3	68.2 ± 1.7	62.1 ± 2.2
SPOT†	$\underline{69.8 \pm 1.7}$	$\underline{68.4 \pm 1.2}$	$\underline{64.6 \pm 2.1}$
LIFTED (ours)	$\mathbf{70.7 \pm 2.3}$	$\mathbf{71.6 \pm 1.4}$	$\mathbf{64.9 \pm 2.1}$

We conduct experiments to evaluate the performance of LIFTED on phase I compared to our baselines. The trial outcome prediction results of all models are reported in Table 1. We first observed that the deep learning-based methods, especially the methods designed for clinical trial outcome prediction including MMF, HINT and SPOT, outperforms the machine learning based methods with a significant performance gap, showcasing the powerful ability to extract critical information from different modalities in various formats of the deep learning encoders, especially 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, verifying its effectiveness in unifying different modalities and dynamically integrating them within the MoE.

Limitations Though LIFTED outperforms the existing models, there still are several limitations to our work. As we mentioned in Section 2.2, different modality data is transformed into natural language descriptions by LLM, specifically, the GPT-3.5. However, the output from LLM is unstable, which may affect the performance of LIFTED.

References

Ahmed, K., Baig, M. H., and Torresani, L. Network of experts for large-scale image categorization. In *Com-*

puter Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, pp. 516–532. Springer, 2016.

Aoki, R., Tung, F., and Oliveira, G. L. Heterogeneous multitask learning with expert diversity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19 (6):3093–3102, 2022.

Arik, S. Ö. and Pfister, T. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6679–6687, 2021.

Badirli, S., Liu, X., Xing, Z., Bhowmik, A., Doan, K., and Keerthi, S. S. Gradient boosting neural networks: Grownet. *arXiv preprint arXiv:2002.07971*, 2020.

Blass, B. E. *Basic principles of drug discovery and development*. Elsevier, 2015.

Chang, Y.-T., Hoffman, E. P., Yu, G., Herrington, D. M., Clarke, R., Wu, C.-T., Chen, L., and Wang, Y. Integrated identification of disease specific pathways using multiomics data. *bioRxiv*, pp. 666065, 2019.

Chen, J., Liao, K., Fang, Y., Chen, D., and Wu, J. Tabcaps: A capsule neural network for tabular data classification with bow routing. In *The Eleventh International Conference on Learning Representations*, 2022a.

Chen, J., Liao, K., Wan, Y., Chen, D. Z., and Wu, J. Danets: Deep abstract networks for tabular data classification and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3930–3938, 2022b.

Chen, J., Yan, J., Chen, D. Z., and Wu, J. Excelformer: A neural network surpassing gbdts on tabular data. *arXiv preprint arXiv:2301.02819*, 2023.

Chen, T., Huang, S., Xie, Y., Jiao, B., Jiang, D., Zhou, H., Li, J., and Wei, F. Task-specific expert pruning for sparse mixture-of-experts. *arXiv preprint arXiv:2206.00277*, 2022c.

Chen, T., Hao, N., Lu, Y., and Van Rechem, C. Uncertainty quantification on clinical trial outcome prediction. *arXiv preprint arXiv:2401.03482*, 2024.

Dai, Y., Tang, D., Liu, L., Tan, M., Zhou, C., Wang, J., Feng, Z., Zhang, F., Hu, X., and Shi, S. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *arXiv preprint arXiv:2205.06126*, 2022.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Eigen, D., Ranzato, M., and Sutskever, I. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

Fan, Z., Wang, L., Jiang, H., Lin, Y., and Wang, Z. Platelet dysfunction and its role in the pathogenesis of psoriasis. *Dermatology*, pp. 1 – 10, 2020. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102868717&partnerID=40&md5=6ddba49123974c7cd9a6b31c57bd0256. Cited by: 1.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

Fu, T., Xiao, C., Qian, C., Glass, L. M., and Sun, J. Probabilistic and dynamic molecule-disease interaction modeling for drug discovery. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 404–414, 2021.

Fu, T., Huang, K., Xiao, C., Glass, L. M., and Sun, J. Hint: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4):100445, April 2022. ISSN 26663899. doi: 10.1016/j.patter.2022.100445.

Fu, T., Huang, K., and Sun, J. Automated prediction of clinical trial outcome, February 2 2023. US Patent App. 17/749,065.

Gao, J., Xiao, C., Glass, L. M., and Sun, J. 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, pp. 803–812, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403123. URL https://doi.org/10.1145/3394486.3403123.

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

Gross, S., Ranzato, M., and Szlam, A. Hard mixtures of experts for large scale weakly supervised vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6865–6873, 2017.

Hazimeh, H., Ponomareva, N., Mol, P., Tan, Z., and Mazumder, R. The tree ensemble layer: Differentiability meets conditional computation. In *International Conference on Machine Learning*, pp. 4138–4148. PMLR, 2020.

Hazimeh, H., Zhao, Z., Chowdhery, A., Sathiamoorthy, M., Chen, Y., Mazumder, R., Hong, L., and Chi, E. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems*, 34:29335–29347, 2021.

Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., and Sun, J. Deeppurpose: A deep learning library for drug-target interaction prediction. *Bioinformatics*, 36(22-23):5545 – 5547, 2020a. doi: 10.1093/bioinformatics/btaa1005. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104047331&doi=10.1093%2fbioinformatics%2fbtaa1005&partnerID=40&md5=db3bedd7f81dd549a5d2220c9accdc77. Cited by: 128; All Open Access, Green Open Access, Hybrid Gold Open Access.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020b.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

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

Jordan, M. and Jacobs, R. Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pp. 1339–1344 vol.2, 1993. doi: 10.1109/IJCNN.1993.716791.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Lo, W.-S., Chiou, H.-W., Hsu, S.-C., Lee, Y.-M., and Cheng, L.-C. 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)*, pp. 7–12, 2019. doi: 10.1109/ITHERM.2019.8757347.

Lu, Y., Chen, T., Hao, N., Rechem, C. V., Chen, J., and Fu, T. Uncertainty quantification and interpretability for clinical trial approval prediction. *Health Data Science*, 2024.

Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., and Chi, E. H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1930–1939, 2018.

Martin, L., Hutchens, M., Hawkins, C., and Radnov, A. How much do clinical trials cost? *Nature Reviews Drug Discovery*, 16(6):381–382, June 2017. ISSN 1474-1784. doi: 10.1038/nrd.2017.70.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Popov, S., Morozov, S., and Babenko, A. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312*, 2019.

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

Rajpurkar, P., Yang, J., Dass, N., Vale, V., Keller, A. S., Irvin, J., Taylor, Z., Basu, S., Ng, A., and Williams, L. M. 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, 06 2020. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.6653.

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., and Houlsby, N. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595, 2021.

Shahbaba, B. and Neal, R. Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, 10(8), 2009.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Siah, K. W., Kelley, N. W., Ballerstedt, S., Holzhauer, B., Lyu, T., Mettler, D., Sun, S., Wandel, S., Zhong, Y., Zhou, B., Pan, S., Zhou, Y., and Lo, A. W. Predicting drug approvals: The novartis data science and artificial intelligence challenge. *Patterns*, 2(8):100312, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100312. URL https://www.sciencedirect.com/science/article/pii/S2666389921001550.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, Y., Lu, Y., Xu, Y., Ma, Z., Xu, H., Du, B., Gao, H., and Wu, J. Twin-gpt: Digital twins for clinical trials via large language model. *arXiv preprint arXiv:2404.01273*, 2024.

Wang, Z., Gao, C., Xiao, C., and Sun, J. Anypredict: Foundation model for tabular prediction, May 2023a.

Wang, Z., Xiao, C., and Sun, J. Spot: Sequential predictive modeling of clinical trial outcome with meta-learning, April 2023b.

Wu, C.-T., Parker, S. J., Cheng, Z., Saylor, G., Van Eyk, J. E., Yu, G., Clarke, R., Herrington, D. M., and Wang, Y. Cot: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinformatics Advances*, 2(1):vbac037, 2022.

Yan, J., Chen, J., Wu, Y., Chen, D. Z., and Wu, J. T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10720–10728, 2023.

Zhang, X., Xiao, C., Glass, L. M., and Sun, J. Deepenroll: patient-trial matching with deep embedding and entailment prediction. In *Proceedings of the web conference 2020*, pp. 1029–1037, 2020.

Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A. M., Le, Q. V., Laudon, J., et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

A. Experiments

A.1. Dataset Descriptions.

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). The detailed data statistics of the HINT dataset are shown in Table 2.

A.2. Overall Performance

We also conduct experiments to evaluate the performance of LIFTED on phase II and III compared to our baselines. The trial outcome prediction results of all models are reported in Table 3.

A.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. The ablation models are described as:

- 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 the 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 integrated by the MoE framework.
- 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 4, 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.

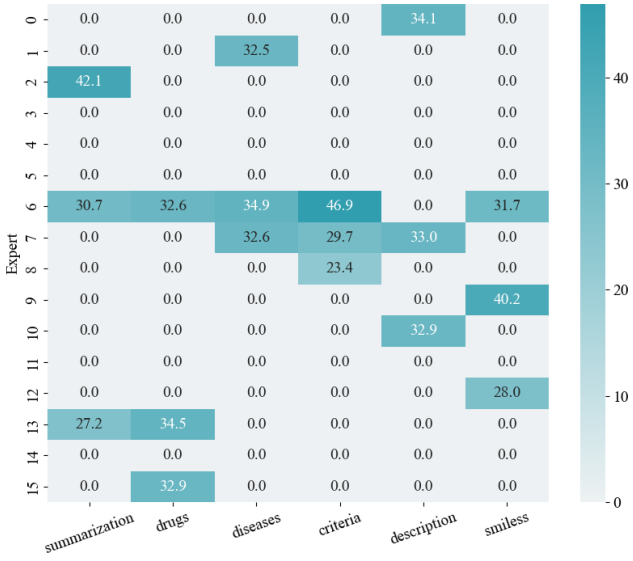

Figure 2. The SMoE experts' importance weights of our model predicting the knee osteoarthritis patient. Experts 6 and 7 play a crucial role in extracting common information patterns across modalities, while other experts specialize in a single specific modality.

A.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 5. The results indicate that LIFTED outperforms all unimodal models in terms of all metrics, demonstrating the effectiveness of multimodal integration. The reason why the F1 score of LIFTED is lower than that of the drug description unimodal models on phase I and III is that the drug description unimodal models, tend to produce all-positive or all-negative predictions, which results in unexpectedly high F1 score with near zero standard deviation due to the high success rate on the dataset. In addition, the results also demonstrate that the drug description modality is the least important modality, while the criteria modality is the most important modality. This is within expectation since the quality of recruited patients plays a crucial role in trial success (Jin et al., 2017).

A.5. Analysis of Sparse Mixture-of-Experts

In addition, we delve into an analysis of the Sparse MoE model to understand the performance enhancements ob-

Table 2. The statistics of the HINT Datasets (Fu et al., 2021). # is short for the number of. The number of Trials is shown by the split of train/validation/test sets.

	# Trials	# Drugs	# Diseases	# patients/trial	# Success	# Failure
Phase I	1,044/116/627	2,020	1,392	45	1,006	781
Phase II	4,004/445/1,653	5,610	2,824	183	3,039	3,063
Phase III	3,092/344/1,140	4,727	1,619	1418	3,104	1,472

Table 3. The clinical trial outcome performance (%) of LIFTED and baselines. 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 underlined, respectively. We observe that LIFTED consistently outperforms all other methods over all three phases.

	Phase II Trials			Phase III Trials		
Method	**PR-AUC**	**F1**	**ROC-AUC**	**PR-AUC**	**F1**	**ROC-AUC**
LR	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	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	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	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	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	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)	60.2 ± 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)	$\underline{62.9 \pm 2.0}$	63.0 ± 1.5	62.6 ± 1.6	$\underline{86.9 \pm 1.6}$	$\underline{83.1 \pm 1.1}$	$\underline{71.8 \pm 2.2}$
DeepEnroll	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	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†	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†	62.6 ± 0.7	$\underline{64.3 \pm 0.6}$	$\underline{63.0 \pm 0.6}$	81.7 ± 0.8	81.0 ± 0.4	71.0 ± 0.4
LIFTED (ours)	$\mathbf{69.8 \pm 1.8}$	$\mathbf{66.2 \pm 1.1}$	$\mathbf{65.1 \pm 1.4}$	$\mathbf{88.3 \pm 1.1}$	$\mathbf{83.8 \pm 0.8}$	$\mathbf{73.5 \pm 1.6}$

tained by the sparse MoE model. 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 2. 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 modalities and capturing specialized information patterns within a single modality.

A.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 3. 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.

B. Related Works

B.1. Clinical Trials 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 (Huang et al., 2020b; Arik & Pfister, 2021; Chen et al., 2022b; Klambauer et al., 2017; Badirli et al., 2020; Chen et al., 2023; Hazimeh et al., 2020; Chen et al., 2022a;

Table 4. 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.

	Phase I Trials			Phase II Trials			Phase III Trials		
Method	**PR-AUC**	**F1**	**ROC-AUC**	**PR-AUC**	**F1**	**ROC-AUC**	**PR-AUC**	**F1**	**ROC-AUC**
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)	$\mathbf{70.7 \pm 2.3}$	$\mathbf{71.6 \pm 1.4}$	$\mathbf{64.9 \pm 2.1}$	$\mathbf{69.8 \pm 1.8}$	$\mathbf{66.2 \pm 1.1}$	$\mathbf{65.1 \pm 1.4}$	$\mathbf{88.3 \pm 1.1}$	$\mathbf{83.8 \pm 0.8}$	$\mathbf{73.5 \pm 1.6}$

Table 5. Performance analysis of multimodal data integration. The best results and second best results are **bold** and underlined, respectively.

	Phase I Trials			Phase II Trials			Phase III Trials		
	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	$\underline{63.1 \pm 2.1}$	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)	$\mathbf{70.7 \pm 2.3}$	$\mathbf{71.6 \pm 1.4}$	$\mathbf{64.9 \pm 2.1}$	$\mathbf{69.8 \pm 1.8}$	$\mathbf{66.2 \pm 1.1}$	$\mathbf{65.1 \pm 1.4}$	$\mathbf{88.3 \pm 1.1}$	$\underline{83.8 \pm 0.8}$	$\mathbf{73.5 \pm 1.6}$

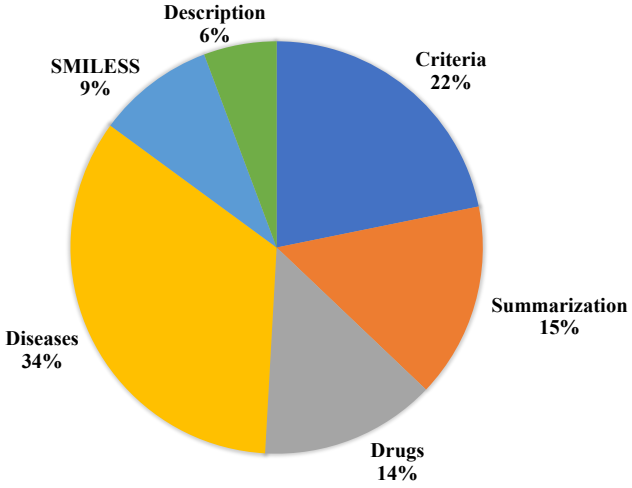

Figure 3. The modality importance weights of our model predicting the type 2 diabetes mellitus patient. LIFTED pay more attention to the disease modality as expected (Chang et al., 2019), since type 2 diabetes mellitus is hard to cure.

Popov et al., 2019; Yan et al., 2023). Specifically, Gayvert et al. (2016) employed random forests to integrate drug chemical structures and properties to predict drug toxicity; Qi & Tang (2019) utilized recurrent neural networks (RNNs) to predict pharmacokinetic outcomes at phase III, leveraging data acquired from phase II trials. Similarly, Lo et al. (2019) employed statistical machine learning models to predict drug approval. Recently, Fu et al. (2023); Lu et al. (2024) proposed a hierarchical interaction network employing different encoders to fuse multiple modal data and capture their correlations for trial outcome predictions; based on Fu et al. (2022), Chen et al. (2024) quantifies uncertainty in the prediction; 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.

B.2. Mixture-of-Experts

Mixture-of-Experts (MoE) is a special type of neural network whose parameters are partitioned into a series of submodules, called experts, functioning in a conditional computation fashion (Jacobs et al., 1991; Jordan & Jacobs, 1993).

Since traditional dense MoE models (Eigen et al., 2013) utilize all experts for each input, they are computationally expensive. Recently, Shazeer 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. Subsequently, Fedus et al. (2022) further reduced the routing computation cost of the MoE layer by routing one sample to only a single expert instead of K experts, enabling the scaling of language models to enormous sizes, such as trillions of parameters, without sacrificing performance. To encourage specialization and decrease redundancy among experts (Chen et al., 2022c), Dai et al. (2022) pre-defined 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; Hazimeh et al., 2021; Ma et al., 2018; Mittal et al., 2022). In addition, different neural network structures (Dauphin et al., 2017; Vaswani et al., 2017) have been proposed and achieved surprising successes in various NLP (Shahbaba & Neal, 2009; Lepikhin et al., 2020; Zhou et al., 2022; Dauphin et al., 2017) and vision (Riquelme et al., 2021; Eigen et al., 2013; Ahmed et al., 2016; Gross et al., 2017) tasks. To identify similar information patterns between different modalities and extract them with the same expert model, LIFTED follows the original design of Sparse Mixture-of-Experts (Shazeer et al., 2017), routing inputs to a subset of experts instead of just one expert, dynamically selecting the experts instead of using a pre-defined assignment.

C. Prompt

The whole prompt, including the system message, is demonstrated in Table 6, and some examples are demonstrated in Table 7.

Table 6. 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

D. 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).
- **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 imputate 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 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.

Table 7. 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 IIIBor 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 cmˆ2 transdermal patch', 'rivastigmine 10 cmˆ2 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.

- **DeepEnroll** (Zhang et al., 2020): initially intended for patient-trial matching, DeepEnroll employs three key components: (1) a pre-trained BERT model (Devlin et al., 2019) to encode eligibility criteria into sentence embeddings, (2) a hierarchical embedding model to handle disease information, and (3) an alignment model to capture interactions between eligibility criteria and diseases. In our experiments, to adapt DeepEnroll for predicting trial outcomes, its functionality is extended by concatenating the molecule embeddings (hm) computed by the MPNN algorithm (Huang et al., 2020a) over molecule graphs with the output of the alignment model.

- **COMPOSE** (Gao et al., 2020): similar to DeepEnroll, COMPOSE was originally proposed for patient-trial matching. A convolutional neural network and a memory network are employed to encode eligibility criteria and diseases, respectively. Similarly, a molecule embedding from MPNN is also concatenated with its embedding for trial outcome prediction.

- **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.

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 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.