Multimodal Clinical Prediction with Unified Prompts and Pretrained Large-Language Models

Caleb Winston*, Chloe Winston†, Cailin Winston‡, Claris Winston‡, Cleah Winston‡
*Department of Computer Science, Stanford University, Stanford, USA
calebwin@cs.stanford.edu
†Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
winstonc@pennmedicine.upenn.edu
‡Department of Computer Science, University of Washington, Seattle, USA
{cailinw, clarisw, cleahw}@cs.washington.edu

***Abstract*—Clinical prediction models (CPMs) increasingly rely on multiple modalities to predict clinical outcomes. The use of free-text data sources (e.g., chief complaint, medical notes) presents new challenges in the heterogeneity of the text across providers and patients and the need to convert the free text to numerical features before combining with other modalities. Prior work has employed multi-head architectures to learn separate heads for different modalities. In this work, we propose a new approach of constructing a unified text prompt that captures information from multiple modalities. We then use existing large-language models (LLMs), optimized with diagnosis-contrastive learning (DCL) to encode the unified prompt and make clinical predictions. We test our approach on the prediction of a variety of outcomes from emergency department visits using a free-text chief complaint and structured numerical data, including demographic information and vital signs. We find that optimized LLMs with unified prompts outperform LLMs that only use the chief complaint by 0.02 weighted F1 score ($p < 0.0001$) and models that only use the structured data modality as numerical inputs by 0.15 ($p < 0.0001$) in predicting acuity. We also observe improvements in the prediction of clinical outcomes, including hospital admission, length of stay, and time to revisit.**

***Index Terms*—multimodal, clinical prediction, large-language models, LLM, unified, prompting, self-supervised learning, contrastive learning, EHR**

I. Introduction

Clinical prediction models (CPMs) integrate an increasing number of data sources - including the patient interview, bedside monitoring, and the electronic health record (EHR) - to assist in tasks ranging from determining patient acuity in the emergency department to predicting patient outcomes, such as length of stay, probability of hospital readmission, and risk of specific diseases like asthma exacerbations [1]–[4]. With an increasing variety of data sources, CPMs must simultaneously handle multiple modalities of input features. Table I surveys and compares cost of acquisition of different data modalities and features. Free-text chief complaints requires no processing and are often provided directly by a patient early in an encounter. Structured data (i.e., categorical variables like demographics or numerical data like vital signs or lab results) may be collected and processed later. Finally, imaging data and lab tests are most expensive time-consuming to acquire and process [5].

Different data modalities present unique challenges. In particular, the text modality presents challenges in heterogeneity across providers and patients [6], frequent misspellings and uncommon abbreviations [7], and the need to convert free text into numerical or categorical features in order to combine with other modalities [8].

Prior work has proposed training unified models to understand multiple data modalities using separate model heads and make clinical predictions [9]. This approach requires a multi-head architecture trained from scratch with a separate encoding for each modality. For example, text is tokenized while structured data is passed through a linear projection layer and the results are passed through a multi-head attention layer.

In this work, we propose a method for clinical prediction using unified prompts and existing LLMs optimized with diagnosis-contrastive learning. We hypothesize that since numerical data modalities can be reduced to text - for example, demographics and vital signs can simply be described in text - a single encoder model can then be supplied a single unified prompt that captures information from multiple modalities. We construct this unified prompt by simply concatenating string representations of both unstructured data (i.e., chief complaint) and structured data (i.e., demographic information and vital signs) and then finetune an existing LLM to predict clinical outcomes given the unified prompt.

Our contributions are as follows:

1) Demonstrate that existing LLMs can be used for multimodal clinical prediction.
2) Present a method of constructing unified prompts for existing LLMs, optimized with diagnosis-contrastive learning, to predict varied clinical outcomes.
3) Evaluate the ability of pretrained LLMs to generalize to diverse multimodal clinical prediction tasks.

II. Methods

In this work, we focus on clinical prediction using multiple modalities: text and structured data. Our proposed method comprises of (a) constructing a unified prompt and (b) using existing LLMs (c) pretrained with self-supervised diagnosis-contrastive learning (DCL) to encode the unified prompt for clinical prediction.

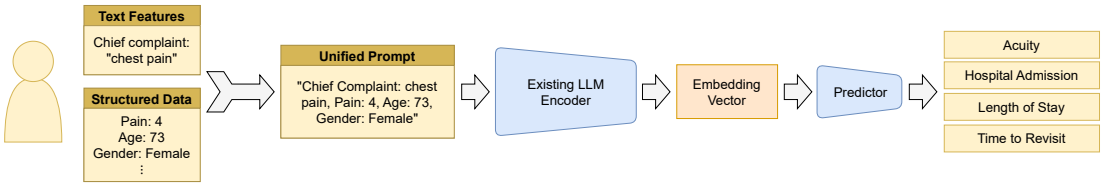

Fig. 1. **End-to-end pipeline.** Our proposed method of combining multiple modalities into a unified prompt and using existing LLMs for clinical prediction.

TABLE I
DIFFERENT MODALITIES WITH VARYING ACQUISITION COST

Modality	Feature	Acquisition Cost
Text	Chief Complaint	Reported directly by the patient either at admission time or through a tele-health application [3].
Structured Data	Demographics	Collected by dispatch in prehospital assessment [10] or $<$ 1 minute to collect in clinic [5].
	Vital Signs	Monitored in real-time with wearables [11] or collected within an hour of arrival [12].
Image	Lab Tests	Varying door-to-X time for imaging [13], ECG [14], cath lab [15]

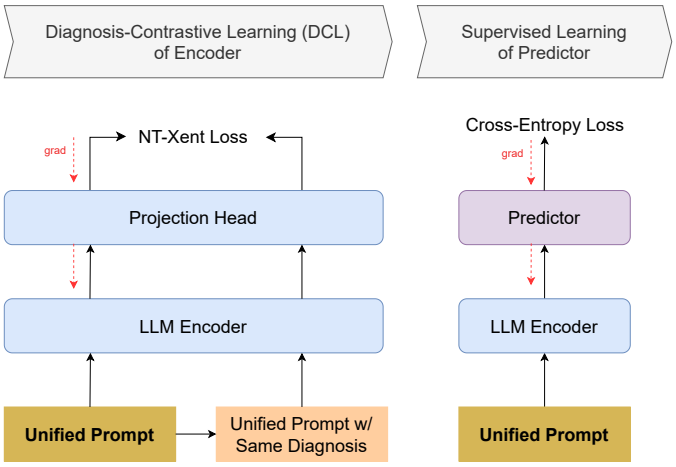

Fig. 2. **Training workflow.** Proposed method for learning to encode multimodal input with (1) a DCL-optimized LLM encoder and (2) a finetuned predictor model.

A. Unified Prompt Construction

Given a set of input features, we generate a single string prompt that serves as the input to the LLM. Each input feature is converted to a string. Chief complaints are assumed to be free text. Demographics and vitals are a mix of categorical and numerical variables. These are again converted to strings without any binning or one-hot encoding. For each feature, we prefix with a short string label - e.g., "Gender: " for gender. The resulting strings are concatenated to form the prompt that is then supplied to the LLM. The output embedding of the last layer of the LLM is then classified by the predictor head.

B. Existing LLMs

We propose using existing LLMs to encode unified prompts. Although the approach taken in prior work has been to train a transformer model from scratch to predict clinical outcomes [9], we observe that many LLMs have been pretrained to understand information sourced from different data modalities [16]–[18]. For example, ClinicalBERT [16] has been pretrained on clinical reports which captures information from chief complaints, demographics, vitals, and lab tests. Indeed, in our experiments we find that ClinicalBERT used as an encoding model for unified prompts outperforms other LLMs including BERT and BiomedBERT, which is pretrained on scientific papers instead of clinical reports.

C. Predictor Head

Finally, we construct a classification head that takes LLM embeddings as input and generates the output label depending on the task. The classification head is a small feedforward neural network (FNN) trained on a smaller amount of supervised training data. We found that a nonlinear FNN yields significantly higher accuracy than a simple linear probe.

D. Diagnosis-Contrastive Learning

We optimize the LLM with contrastive loss to align embeddings with diagnostic information. We adapt the NT-Xent loss function [19] and the SimCLR method [19], defining positive pairs as pairs of unified prompts corresponding to the same diagnosis (ICD-10 code). By minimizing this diagnosis-contrastive loss function, the LLM learns similar embeddings for different inputs with same diagnosis. This method is entirely task-agnostic and since many clinical outcomes correlate with the diagnosis, the resulting optimized LLM can generalize to varied clinical prediction tasks (evaluation results shown in Figure 4).

III. RESULTS

A. Setup

1) Dataset: We evaluate our approach for multimodal clinical prediction with the MIMIC-IV-ED dataset of emergency department (ED) visit records from the Beth Israel Deaconess Medical Center from 2011 to 2019 [20]. Our data processing workflow for sampling, filtering, and splitting the dataset is shown in Figure 3.

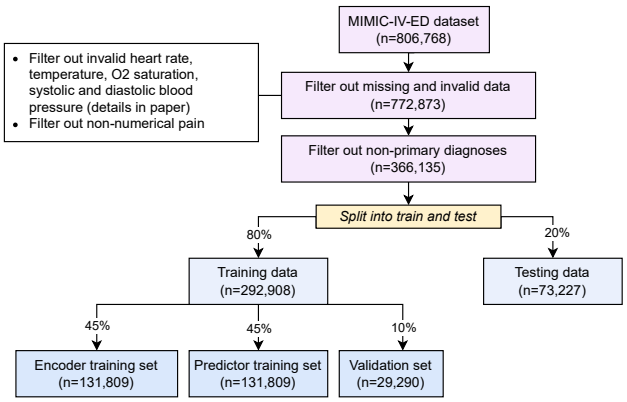

Fig. 3. **Dataset processing steps.**

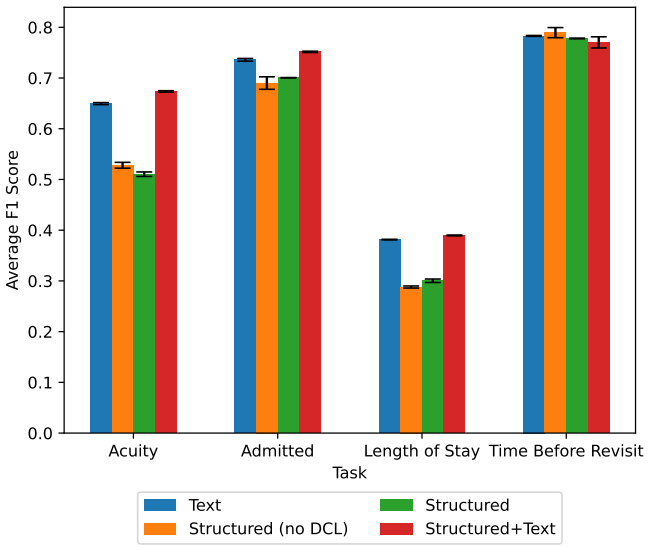

Fig. 4. **Performance improvement from combining modalities and DCL.** Mean and standard error shown for the prediction of various tasks using DCL-optimized ClinicalBERT on unified text prompts consisting of different sets of features. Structured refers to structured data modality features and Text refers to text modality features.

Given significant class imbalance, we (1) split the data using stratified sampling across classes and (2) oversample the data using the class weighting approach introduced in [21]. In oversampling, each class with n samples is weighted by $1/E_n$ where the effective number of samples is defined by $E_n = (1-\beta^n)/(1-\beta)$ with $\beta = 0.99999$.

In our work we focus on both multiclass and binary classification with 3 output classes for the tasks of predicting acuity, length of stay, and time before revisit and the 2 classes for predicting hospital admission.

2) Models: In our evaluation, we use ClinicalBERT [22], a large-language model (LLM) trained on clinical reports, as an encoder model. We also evaluated BERT and BiomedBERT, which is trained on PubMed articles, but found that ClinicalBERT generally outperforms [17], [18]. Although ClinicalBERT is trained for text generation, we repurpose it as an embedding model by taking the hidden state of the last layer as the output embedding. For the predictor head, we use a small feedforward neural network with one size-128 hidden layer.

Although we do not include a detailed comparison in this paper of our method against baseline methods from prior work such as using TF-IDF [23] or Word2Vec [24], we have found that TF-IDF and Word2Vec both tend to underperform BERT encoding models for chief complaint classification.

3) Hyperparameters: We tuned the following hyperparameters for the LLM encoder and predictor models: batch size $\in$ [32, 64, 128, 256, 512], learning rate $\in$ [1e-5, 2e-5, 3e-5, 5e-5], and NT-Xent loss temperature $\in$ [0.05, 0.1]. For each configuration, we ran five trials each with a random split of the dataset, and then chose the hyperparameters yielding the highest overall or per-class accuracy.

B. Benefit of multiple modalities

We first evaluate the benefit of our proposed approach across a variety of clinical prediction tasks. As points of comparison, we test using only the text modality or only the structured data modality. For a baseline, we also test using only structured data with a typical multi-head architecture. Our results are shown in Figure 4.

We find that across 3 out of 4 tasks, combining modalities yields a higher performance. In predicting acuity, using both text and structured data achieves a weighted F1 score 0.02 higher ($P < 0.0001$) than just text, 0.16 higher ($P < 0.0001$) than just structured data, and 0.15 higher ($P < 0.0001$) than just structured data and no LLM. The exception is the task of predicting time to revisit where we observed higher variance in performance.

Figure 4 shows that Structured+Text has a smaller performance gain over Text-only than over Structured-only. This can be attributed to different modalities contributing differently to performance. We evaluate this in Section III-C1.

C. Ablation study

We conduct an ablation study to quantify and understand (1) the importance of different features, (2) the effect of pretraining with diagnosis-contrastive learning (DCL), and (3) using a unified prompt instead of a multi-head architecture.

1) Importance of features: To understand feature importance, we remove different sets of features and compare the change in performance measured by weighted F1 score. The 4 sets of features we consider are: chief complaint, pain level, demographics (gender and age), vitals (temperature, heart rate, respiratory rate, blood pressure O_2 saturation). Table II shows the change in performance of predicting acuity with different feature sets removed. We find that removing chief complaint has the largest impact on performance with a decrease of 0.16 ($p < 0.0001$) in weighted F1 score. Demographics and vitals each have the next largest impact on performance with 0.01 each ($p = 0.0015$). The highest performance is achieved when all features are used, again demonstrating the benefit of combining the text modality (chief complaints) with the structured

TABLE II

ABLATION OF SETS OF FEATURES OF DIFFERENT MODALITIES. MEAN AND STANDARD ERROR SHOWN FOR THE PREDICTION OF VARIOUS TASKS USING DCL-OPTIMIZED CLINICALBERT ON UNIFIED TEXT PROMPTS CONSISTING OF DIFFERENT SETS OF FEATURES. ENTRIES WITHOUT STANDARD ERROR ARE RESULTS FROM TWO TRIALS.

Features	Acuity	Admitted	Length of Stay	Time to Revisit
All Features	$0.67 \pm 1.45 \times 10^{-3}$	0.75	0.39	0.77
− Pain	$0.67 \pm 2.76 \times 10^{-3}$	$0.75 \pm 8.48 \times 10^{-4}$	$0.39 \pm 2.37 \times 10^{-3}$	$0.77 \pm 3.69 \times 10^{-3}$
− Demographics	$0.66 \pm 1.13 \times 10^{-3}$	$0.74 \pm 1.17 \times 10^{-3}$	$0.38 \pm 2.88 \times 10^{-3}$	$0.76 \pm 3.48 \times 10^{-3}$
− Vitals	$0.66 \pm 1.51 \times 10^{-3}$	$0.75 \pm 1.04 \times 10^{-3}$	$0.39 \pm 6.27 \times 10^{-4}$	$0.76 \pm 4.79 \times 10^{-3}$

data modality (demographics and vitals). Prior studies support that chief complaints contribute more to predictive accuracy than structured data like laboratory results [9], and our results demonstrate further performance gain with DCL.

2) Effect of pretraining: In Table III, we evaluate the impact on performance by pretraining with self-supervised diagnosis-contrastive learning (DCL). In the task of predicting acuity, we find that DCL yields a significant improvement of 0.06 ($p < 0.0001$) in weighted F1 score. We also observe a performance improvement with DCL in the tasks of predicting hospital admission (+0.08) and length of stay (+0.25). The effect of DCL is less pronounced when using a multi-head architecture.

3) Comparison with multi-head architecture: Finally, we compare our proposed approach of using a unified prompt against a more common approach of creating a multi-head architecture. In this architecture there are two one-layer heads: one for the text (chief complaint) embedding vector and the other for the structured data (demographics and vitals) vector. The outputs of each head are combined by summing before passing through the rest of the network. We find there is no significant improvement in performance from using the multi-head architecture, demonstrating that a single LLM encoder and unified prompt can perform on par (Table III).

IV. CONCLUSION

In this paper, we demonstrated that existing LLMs can be prompted to encode multiple modalities of input to complete clinical prediction tasks by constructing unified cross-modality text prompts. We further showed the benefit of self-supervised diagnosis-contrastive learning to improve predictive accuracy with multiple input modalities. We found that our proposed approach yielded a significant improvement in performance (weighted F1 score) across the tasks of predicting hospital admission, length of stay, and acuity (improved by 0.15, $p < 0.0001$) using unified prompts instead of structured data only. Our approach demonstrates that contrastive LLM representation learning with simple unified prompts can effectively combine the modalities of text chief complaints and structured data to achieve improved accuracy on clinical prediction tasks.

REFERENCES

[1] S.-C. Lu, C. Xu, C. H. Nguyen, Y. Geng, A. Pfob, and C. Sidey-Gibbons, "Machine learning–based short-term mortality prediction models for patients with cancer using electronic health record data: systematic review and critical appraisal," *JMIR medical informatics*, vol. 10, no. 3, p. e33182, 2022.

[2] A. Martin, V. Bauer, A. Datta, C. Masi, G. Mosnaim, A. Solomonides, and G. Rao, "Development and validation of an asthma exacerbation prediction model using electronic health record (ehr) data," *Journal of Asthma*, vol. 57, no. 12, pp. 1339–1346, 2020.

[3] J. Shi, M. Ye, H. Chen, Y. Lu, Z. Tan, Z. Fan, and J. Zhao, "Enhancing efficiency and capacity of telehealth services with intelligent triage: a bidirectional lstm neural network model employing character embedding," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 269, 2023.

[4] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang, "Combining structured and unstructured data for predictive models: a deep learning approach," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–11, 2020.

[5] G. Erion, J. D. Janizek, C. Hudelson, R. B. Utarnachitt, A. M. McCoy, M. R. Sayre, N. J. White, and S.-I. Lee, "Coai: Cost-aware artificial intelligence for health care," *Nat Biomed Eng*, 2022.

[6] R. L. Richesson, J. P. Turley, K. A. Johnson-Throop, C. Eick, M. S. Tuttle, and R. Richesson, "Obtaining comparable presenting complaint data from heterogeneous emergency department databases," *A Process For Achieving Comparable Data From Heterogeneous Databases*, p. 93, 2006.

[7] J. Dara, J. N. Dowling, D. Travers, G. F. Cooper, and W. W. Chapman, "Evaluation of preprocessing techniques for chief complaint classification," *Journal of biomedical informatics*, vol. 41, no. 4, pp. 613–623, 2008.

[8] D. A. Thompson, D. Eitel, C. M. Fernandes, J. M. Pines, J. Amsterdam, and S. J. Davidson, "Coded chief complaints—automated analysis of free-text complaints," *Academic emergency medicine*, vol. 13, no. 7, pp. 774–782, 2006.

[9] H.-Y. Zhou, Y. Yu, C. Wang, S. Zhang, Y. Gao, J. Pan, J. Shao, G. Lu, K. Zhang, and W. Li, "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 743–755, 2023.

[10] A. R. Wheeler, C. Cuenca, A. D. Fisher, M. D. April, S. A. Shackelford, and S. G. Schauer, "Development of prehospital assessment findings associated with massive transfusion," *Transfusion*, vol. 60, pp. S70–S76, 2020.

[11] Z. Wang, Z. Yang, and T. Dong, "A review of wearable technologies for elderly care that can accurately track indoor position, recognize physical activities and monitor vital signs in real time," *Sensors*, vol. 17, no. 2, p. 341, 2017.

[12] S. S. Falconer, C. M. Karuppan, E. Kiehne, and S. Rama, "Ed triage process improvement: timely vital signs for less acute patients," *Journal of Emergency Nursing*, vol. 44, no. 6, pp. 589–597, 2018.

[13] C. B. Lin, E. D. Peterson, E. E. Smith, J. L. Saver, L. Liang, Y. Xian, D. M. Olson, B. R. Shah, A. F. Hernandez, L. H. Schwamm *et al.*, "Emergency medical service hospital prenotification is associated with improved evaluation and treatment of acute ischemic stroke," *Circulation: Cardiovascular quality and outcomes*, vol. 5, no. 4, pp. 514–522, 2012.

[14] D. B. Diercks, J. D. Kirk, C. J. Lindsell, C. V. Pollack Jr, J. W. Hoekstra, W. B. Gibler, and J. E. Hollander, "Door-to-ecg time in patients with chest pain presenting to the ed," *The American journal of emergency medicine*, vol. 24, no. 1, pp. 1–7, 2006.

[15] E. H. Bradley, J. Herrin, Y. Wang, B. A. Barton, T. R. Webster, J. A. Mattera, S. A. Roumanis, J. P. Curtis, B. K. Nallamothu, D. J. Magid *et al.*, "Strategies for reducing the door-to-balloon time in acute

TABLE III

ABLATION OF UNIFIED PROMPTS AND DCL. MEAN AND STANDARD ERROR SHOWN FOR THE PREDICTION OF VARIOUS TASKS USING CLINICALBERT WITH AND WITHOUT DCL PRETRAINING USING UNIFIED PROMPTS OR THE MULT-HEAD APPROACH. ENTRIES WITHOUT STANDARD ERROR ARE RESULTS FROM TWO TRIALS.

Input Encoding	LLM Optimization	Acuity	Admitted	Length of Stay	Time to Revisit
Unified Prompt	DCL	$0.67 \pm 1.45 \times 10^{-3}$	0.75	0.39	0.77
	− DCL	$0.62 \pm 6.65 \times 10^{-3}$	0.68 ± 0.04	$0.14 \pm 5.62 \times 10^{-3}$	$0.78 \pm 8.20 \times 10^{-3}$
Multi-Head	DCL	$0.68 \pm 3.46 \times 10^{-3}$	$0.76 \pm 3.81 \times 10^{-3}$	$0.39 \pm 2.06 \times 10^{-3}$	$0.77 \pm 7.63 \times 10^{-3}$
	− DCL	$0.66 \pm 5.88 \times 10^{-3}$	0.74 ± 0.01	0.36 ± 0.02	0.78 ± 0.01

myocardial infarction," *New England Journal of Medicine*, vol. 355, no. 22, pp. 2308–2320, 2006.

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

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," 2020.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[20] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[21] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.

[22] G. Wang, X. Liu, Z. Ying, G. Yang, Z. Chen, Z. Liu, M. Zhang, H. Yan, Y. Lu, Y. Gao *et al.*, "Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial," *Nature Medicine*, vol. 29, no. 10, pp. 2633–2642, 2023.

[23] I. Valmianski, C. Goodwin, I. M. Finn, N. Khan, and D. S. Zisook, "Evaluating robustness of language models for chief complaint extraction from patient-generated text," *arXiv preprint arXiv:1911.06915*, 2019.

[24] T.-L. Chen, J. C. Chen, W.-H. Chang, W. Tsai, M.-C. Shih, and A. W. Nabila, "Imbalanced prediction of emergency department admission using natural language processing and deep neural network," *Journal of Biomedical Informatics*, vol. 133, p. 104171, 2022.