MEDI-BERPT: A Novel Multitask Approach to Streamlining Chinese Healthcare

Stanford CS224N Custom Project

Bi Tian Yuan
Department of Computer Science
Stanford University
`jackyuan@stanford.edu`

Junyang Sun
Symbolic Systems
Stanford University
`sunnysun@stanford.edu`

Abstract

In the pursuit of revolutionizing the medical industry, Large Language Models (LLMs) have faced obstacles due to stringent quality requirements and numerous ethical concerns. Previous medical NLP approaches focused on building single-task models that were dedicated to simple classification or to generating doctor-like responses. Our study presents a multi-task transformer encoder-decoder model that takes a patient's symptom descriptions and generates predictions for both the appropriate medical department and an initial diagnosis or a follow-up question. The primary objective is to evaluate the extent to which the inclusion of a secondary task, specifically the doctor-diagnosis generation, enhances the performance of the medical department classification task.

We leverage a dataset comprising 3 million doctor-patient conversations scraped from Chinese online medical forums (MedDialog) to fine-tune BERT and GPT-2 models, establishing baseline performances. Our findings demonstrate that incorporating a secondary objective enables the model to capture more nuanced relationships between the doctor's response and the patient's symptom description, consequently enhancing the medical department classification process.

The proposed multi-objective transformer encoder-decoder model outperforms the original BERT encoder with an accuracy of 91% compared to 84% in predicting the top 10 most relevant medical departments. Nonetheless, our analysis highlights significant limitations in attempting to generate a doctor's response based solely on the patient's symptom description, underscoring the importance of developing AI-assisted tools that support patients within the medical system, rather than seeking to replace human expertise.

1 Key Information

- Mentor: Elaine Sui

2 Introduction

2.1 Motivation.

Recent advances in large language models (LLM) offer an opportunity to rethink AI systems, specifically in the medical domain. In the United States and China, patients often endure weeks or even months of waiting time for appointments with general practitioners, who are responsible for referring the patient's to the appropriate specialized doctor practitioners.

Developing a medical healthcare department recommendation system can substantially reduce this waste of time by directly referring the patients to the relevant specialized doctor practitioners, removing the necessity of having to meet with the general practitioner first, freeing up significant medical resources, and saving people time.

Researchers from Google DeepMind have sought to tackle this issue by training an LLM on U.S. medical exam data, encompassing multiple-choice and short-answer questions, and utilizing few-shot learning to answer general internet medical questions[1]. The ability of LLMs to medical question-answering might be an emergent ability combined with specific downstream finetuning of data. While contemporary LLMs can generate coherent sentences corresponding to patient symptoms, they often struggle to give a correct diagnosis.

By embracing the paradigm that LLM cannot replace human expertise in giving a patient diagnosis, we can concentrate our efforts on building models that alleviate the patient burden by guiding them toward the appropriate department, bypassing the need for initial consultations with general practitioners and subsequent referrals to specialized physicians.

2.2 Our contribution.

We introduce an AI agent MEDI-BERPT designed to refer patients to the appropriate medical department based on their symptoms. Our focus lies on a primary and secondary task: the primary task is to build a multi-class classifier directing the patient's inquiry to the appropriate medical department, and the secondary text generation providing initial diagnoses or follow-up questions from a general physician's perspective. The primary department classification task can remove the necessity of having to meet with a general practitioner first in order to be seen by the specialized doctor, and the secondary question-answering task can assist doctors in better understanding the patient's condition. By developing a scalable AI system deploy-able in clinics and hospitals in China, where a 1.4 billion population faces increased risks of system overflow, we aim to streamline the quality and efficiency of a general doctor checkup.

Considering the inherent risks of AI model-based misdiagnoses and the challenges in logically generating accurate medical advice based on symptoms, we prioritize leveraging the text-generation objective to improve the medical department classification objective, thus aiding patients in their hospital check-in process. We establish a baseline for department classification results using BERT models trained on Chinese medical corpora and anticipate a substantial improvement in classification results with the support of the question-answering objective.

We make use of an existing dataset collected from online medical forums between patients and doctors and a ground-truth label of the doctor's department to train and experiment with our models. Our MEDI-BERPT model incorporates a pre-trained BERT encoder with a pre-trained GPT2 decoder and simultaneously outputs a classification of the department and a predicted doctor's response to the question.

3 Related work.

3.1 Domain Specificity

Three dominant deep learning models in NLP are BERT[2], GPT, and T5. As per the rationale of the T5 paper ([3]), it is advantageous for a transformer encoder-decoder model to be pre-trained on a large text corpus and subsequently fine-tuned for specific downstream tasks. This approach provides more nuanced attention to domain-specific representations and improves the robustness of the language model.

3.2 Multi-Task Learning

The multi-task approach has been extensively examined in Kaiser et al.'s and Hu et al.'s papers [4][5]. Kaiser et al. introduced a single model trained concurrently on image classification, translation, and speech recognition tasks, utilizing convolutional neural networks and recurrent neural networks. They demonstrated the high performance achieved by a single model sharing the same encoder with unique decoders for different downstream tasks. Conversely, Hu et al. showed how a single model sharing the same encoder and decoder weights achieved high performance on various tasks through task-specific attention mechanisms. Jointly training multiple objectives enforces shared weights in different channels, leading to better input representations. Specifically, mixing computation and attention blocks may be an effective strategy to enhance performance across multiple tasks.

The Text-to-Text Transfer Transformer (T5) model also presents a novel approach to NLP and multitask language models [3]. It employs a robust encoder-decoder structure that generalizes to translation, sentiment analysis, word prediction, and summarization tasks, suggesting that the success of one task can directly inform other tasks.

Inspired by the multitasking approach and an encoder-decoder structure, we hypothesize that incorporating a seq2seq objective alongside our classification task could improve our model's performance in predicting the appropriate medical department for the patient.

3.3 Cross Attention & Encoder-Decoder Architecture

Human doctors must learn a holistic understanding of the patient's symptoms first before giving a medical diagnosis, and might constantly refer back to the list of the patient's symptoms while giving the diagnosis in real life. We implemented cross-attention so that the decoder, while giving a prediction of a doctor's diagnosis, can have access to the entire embedding representation of the patient's symptoms. This leads to better question-answering performance to produce more human-like responses. A better performance on the seq-to-seq task consequently helps boost the classification performance as these tasks reinforce each other.

We could not have used the existing hugging face encoder-decoder module because the hugging face encoder-decoder model is dedicated to a specific task. We built our own custom models of a BERT encoder, a classification layer appended to the output of the last hidden layer of the BERT encoder, and a GPT-2 decoder in order to backpropagate both the classification loss from the predictions on medical departments and the GPT-2 seq-to-seq loss.

3.4 Medical Datasets

Researchers from Sun Yatsen University used a Reinforcement Learning framework to predict 4 types of diseases based on a pre-defined 66 types of symptoms [6]. They have achieved quite high accuracy to build a dependency relationship between the symptoms and the diseases; however, they rely on the assumption that the patients would be able to describe their own symptoms using relevant medical terms that they have built their framework around. As a result, this wouldn't be able to scale and generalize to a large medical systems with a population who do not have clinical knowledge on describing their own symptoms using medically relevant terms.

Researchers from UCSD have attempted to carry out seq-to-seq generation on Chinese medical online forum conversations, yet have only achieved a meager 7% BLEU score for 2-gram words [7]. Even so, they still claimed that the generated responses were "clinically meaningful", without analyzing what further improvements they could be making to improve the generated responses. Without properly rethinking the whole model, it doesn't seem feasible to generate meaningful doctor responses, which is why we chose to focus on only using question-answering to improve our primary multi-class classification tasks.

4 Approach

4.1 Baselines

We started with BERT[2] (Bidirectional Encoder Representations from Transformers) and RoBERTa[8] (Robustly Optimized BERT Pretraining Approach) models in Chinese.

First Baseline: Our first baseline model for the encoder is a RoBERTa-Base classification model that was trained on CLUECorpusSmall[1] and finetuned on Chinese news articles[9][2]. The fine-tuned model is taken from HuggingFace and used directly with the provided tokenizer and multi-class model. BERT employs a ubiquitous transformer architecture, and the relevant architecture details include the 12 layers, 768 hidden dimensions, and the 10 different output classes.

Second Baseline: Our second baseline model for the encoder, MedBERT-wwm[10], is fine-tuned on Chinese medical corpus. Since this model is not a classifier, we modified the model structure and added an output layer with a dropout probability of 0.1 and a 10-class classification layer that generates 10-label classification accordingly.

Third Baseline: Our third baseline for the decoder comes from a GPT2 model trained on CLUECorpusSmall[3][11]. It is a general-domain text generation model that leverages the transformer model and generates the subsequent token based on the token that has come before. Developing these distinct baseline models, such as the standalone encoder and decoder, allows us to compare how combining the encoder-decoder in a multi-task setting improves the classification task or the question-answering task.

Our primary focus is on how seq2seq generation enhances the performance of the classification task. Consequently, we did not use the state-of-the-art T5 model because it is an encoder-decoder model that is dedicated to only the question-answering task. Our ultimate goal is to use multitask learning to improve our department classification task, making our BERPT (BERT + GPT2) multi-task model, which we will explore later, more relevant to our objective.

4.2 BERPT model architecture

We enhance the state-of-the-art BERT and GPT-2 models by integrating them into a multi-task encoder-decoder transformer architecture. We hypothesize that, by having a primary objective of department classification and a secondary objective of text-generation, the encoder can better learn the semantic reasoning behind the patient's symptoms and by developing a better rationale for its given medical department classification.

Encoder: The patient's symptoms are input into the BERT encoder with 12 layers and 768 hidden-size neurons. The last hidden layer of the BERT encoder, which has size $(batch\_size, sequence\_length, hidden\_size)$ is passed through an additional classification layer, which predicts department labels and back-propagates the associated losses throughout the network.

Decoder: Simultaneously, the decoder processes the input of the doctor's response to the patients' questions, predicting the subsequent token autoregressively while performing cross-attention with the BERT encoder's final hidden layer. The decoder then calculates the loss based on its autoregressively predicted response, compares it to the ground truth of the doctor's actual response, and backpropagates this loss throughout the entire network, including the encoder. By jointly backpropagating both the classification and seq2seq losses through the encoder, it learns to perform both tasks, leveraging knowledge gained from the text generation task to improve department classification performance.

We decided to use teacher-forcing in the training process of the decoder, where the ground-truth output from the previous time step is used as input. This is necessary because it guarantees a faster convergence to learn the target sequence more effectively and provides better training stability since an incorrect previous token will no longer be allowed to destabilize future token predictions.

Multi-task Back-Propagation: The back-propagation of the classifier's loss through the encoder assists in identifying which tokens to attend to for accurate medical department classification. Concurrently, backpropagating the decoder's seq-to-seq loss through the encoder

[1] `https://huggingface.co/uer/chinese_roberta_L-12_H-768`

[2] `https://huggingface.co/uer/roberta-base-finetuned-chinanews-chinese`

[3] `https://huggingface.co/uer/gpt2-chinese-cluecorpussmall`

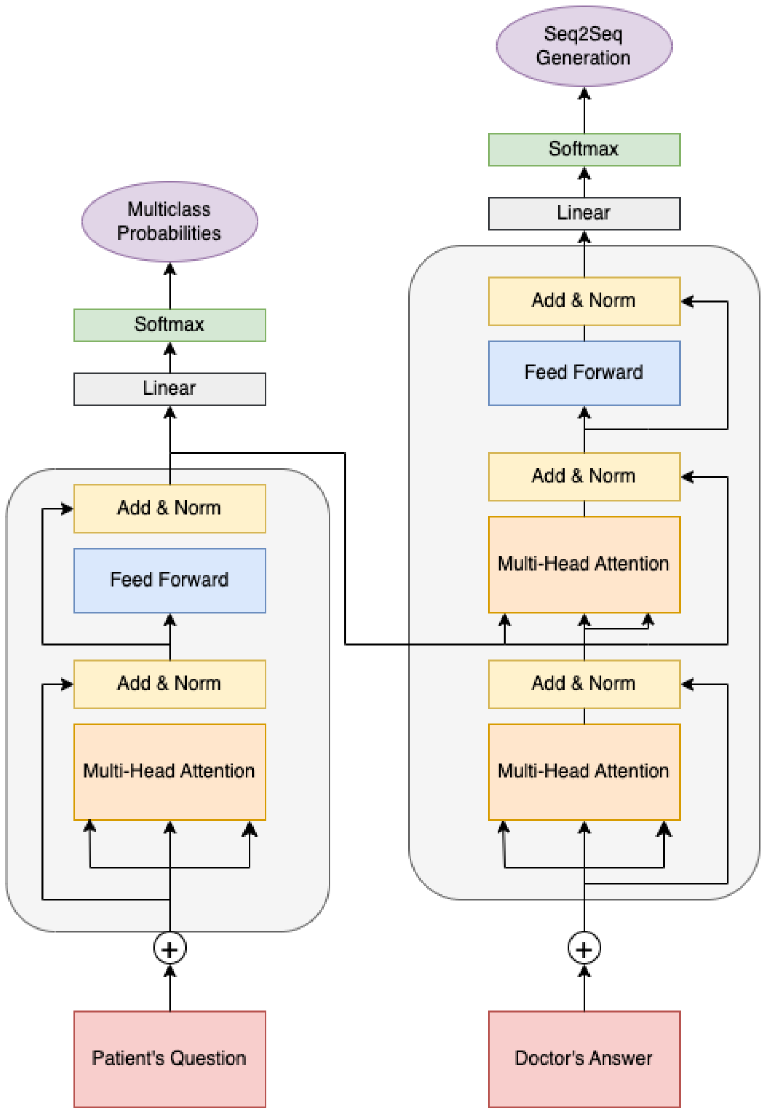

Figure 1: A diagram of our BERPT structure

helps the encoder comprehend the doctor's rationale in response to the patient's question, thereby enhancing the encoder's understanding of the patient's query and improving classification results.

5 Experiments

5.1 Data & Preprocessing

Our study utilizes MedDialog, an extensive Chinese medical corpus that was specifically designed for question-answering tasks [7]. This dataset contains over 3.4 million conversations between patients and accredited physicians on an online forum, making it the largest conversational-based Chinese dataset available for this purpose. The dataset and code are publicly accessible [4] and cover 29 different categories of specialties and 192 fine-grained specialties.

[4] https://github.com/UCSD-AI4H/Medical-Dialogue-System

The dataset was provided in ".txt" files for each year from 2010 to 2020, with each year adopting a different data format. We developed a custom preprocessing script to extract pertinent information from the physicians' raw data and dialogues on a file-by-file basis, eventually concatenating them. In particular, we extracted three components: the patient's description of symptoms, the department of the physician who responded to the patient's inquiry, and the physician's initial assessment of the patient's symptom description. We selected the 10 most prevalent medical departments and stored the relevant information in a CSV file for convenient access by our models. The resulting processed dataset comprises 597,906 training samples and 74,739 testing samples, each labeled with one of ten department categories, such as Gynecology, Ophthalmology, Obstetrics, and others.

5.2 Evaluation Methods

We evaluate our classification performance using accuracy, precision, recall, and F1-score. We also used BLEU and ROUGE scores to measure the equality of the output from seq2seq tasks. These are standard evaluation metrics to assess multi-class classification and text generation, which will not be explored extensively in the paper.

5.3 Experimental Details

Our experiments involve training 5 models, listed in the table. During training, we used a *batch_size* of 8, where input data has the shape of (*batch_size, sentence_length*), with a *max_length* of 200 tokens for the length of each sequence. We provided padding tokens in case the *sentence_length* is less than 200.

First Experiment: BERT + Grid Search In our first experiment, we utilized the BERT model to perform a multi-class classification task on the top 10 patient department categories in Chinese medical data. We set training epochs to 5 for all models. Using the original classification layers provided by BERT, we achieved an accuracy score of 0.84 and an F1 score of 0.81.

To enhance the model's performance, we removed the original classification layers and added a dropout layer with a 10% probability and a classification layer with random initialization. This approach resulted in a slightly lower accuracy score of 0.84 and an F1 score of 0.8. We also conducted a grid search across various *learning_rate, dropout_probability, temperature, batch_size* parameters, but none outperformed the original BERT model. We hypothesize that this is because the BERT model already identified the most suitable hyperparameters for most general tasks. In an attempt to further refine the model, we froze the BERT model parameters and trained only the parameters of the newly attached classification layer. However, this approach yielded a disappointing accuracy score of 0.39 and an F1 score of 0.24.

Second Experiment: medBERT + Parameter Freezing In our second experiment, we replicated the grid search and hyperparameter tuning process with the MedBERT-wwm model, a BERT model pre-trained on a Chinese medical corpus. Initially, we achieved a higher accuracy score of 0.86 and an F1 score of 0.85. However, upon freezing the MedBERT-wwm model parameters, we obtained a slightly worse accuracy score of 0.35 and an F1 score of 0.17.

Third Experiment: GPT2 decoder Our third experiment involved training a GPT2 decoder on the patient-doctor dialogue corpus. Although our primary goal was to utilize the text generation objective to assist in the classification task, we sought to determine if training the GPT2 decoder separately before incorporating it into the transformer encoder-decoder model could enhance its performance.

Fourth Experiment: medBERT + GPT2 Our fourth experiment implemented the BERPT model architecture, as specified earlier, using the MedBERT-wwm model trained on a Chinese medical corpus and a general Chinese GPT2 model. Due to the unavailability of a GPT2 model pre-trained on a medical corpus and limited AWS CUDA memory, we employed the smaller GPT2 Chinese model instead of the larger base GPT2 Chinese model. We trained each model from scratch, conducting validation tests at the end of every epoch and logging the results using Weights Biases.

Fifth Experiment: MEDI-BERPT Our fifth and final experiment involved loading the pre-trained weights from the standalone BERT and standalone GPT2 models into the multi-task BERPT model. We aimed to evaluate if pre-training individual components of the BERPT model could help boost performance on classification tasks.

5.4 Results

Discussion of Quantitative Results From Multi-class Classification: The multi-class medical department's classification results obtained by us can be seen in Table 1. While contrasting our results with CHMBERT[12], a department classification model while finetuning of Chinese Medical BERT, one can see that our MEDI-BERT + GPT2 model outperforms all the model listed in Wang et al. when looking at the overall department prediction accuracy. This result was initially unexpected, but after an analysis of the learning rate chosen by Want et al., we found that their learning rate and other hyperparameters diverged from our models, validating our findings. In Figure 3 we also display the cumulative classification loss of our MEDI-BERT + GPT2 model, and in Figure 4 we show the cumulative seq2seq loss from the GPT2 decoder of the MEDI-BERT + GPT2 model.

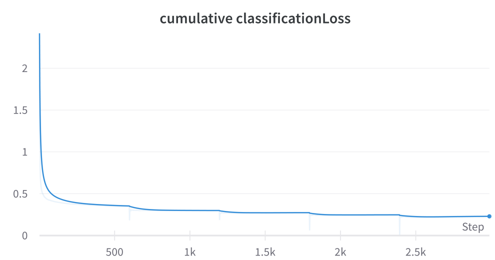

Figure 3: Classification Loss During Training

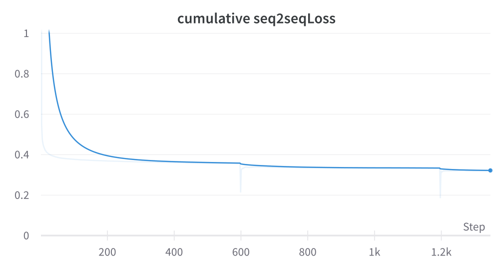

Figure 4: Seq-to-seq Loss During Training

Table 1: Performance comparison of experimental models on MedDialog

Model	Task	Accuracy	Precision	Recall	F1 Score
CHMBERT	0.86	N/A	N/A	N/A	0.74
BERT	Multiclass	0.84	0.81	0.8	0.81
medBERT	Multiclass	0.87	0.84	0.83	0.83
medBERT + GPT2	Multitask	**0.91**	0.85	**0.87**	**0.86**
MEDI-BERPT	Multitask	0.90	**0.86**	0.85	0.85

Discussion of Qualitative Results: Meanwhile, for the secondary seq-to-seq task in BERPT, we see an improvement in the higher medical reasoning tasks but a decrease in coherence when compared to the stand-alone GPT2 decoder. A stand-alone GPT2 decoder sentence is displayed in Table 2, where we can see that the GPT2 generated response was coherent and touched on general information on the symptoms, such as the hormone medications, yet its response was not tailored to the male patient's symptoms. Specifically, the stand-alone GPT2 decoder's sentences had higher sentence coherence compared to the transformer encoder-decoder. However, they had a lower BLEU score and were not discussing treatments that were specific to the patient.

The difficulty of understanding the rationale behind the doctor's responses to the patient's symptoms is a likely reason that the BLEU score is relatively low in both the stand-alone GPT2 decoder model and the transformer encoder-decoder model. We expect that by adding more high-quality doctor responses, the secondary doctor response generation tasks can be improved, and consequently improving the multi-class classification task.

6 Analysis

6.1 Multi-tasking objectives improves the performance of a single-task objective

Our most significant contribution lies in demonstrating the potential of incorporating supplementary tasks into an originally single-task encoder to enhance its performance.

The BERT Encoder's classification accuracy on medical departments reached 84%. Utilizing a BERT Encoder previously fine-tuned on Chinese medical data resulted in a 3% increase in classification accuracy. Most notably, we showed that by allowing the decoder to concurrently backpropagate its seq-to-seq loss and the multi-class classification task loss, we could outperform a single-task multi-classification BERT encoder by 7%. This substantial improvement highlights the feasibility of training any model optimized for a primary objective to be augmented by a secondary objective in the medical domain.

We hypothesize that this is because the encoder, aided by the decoder, learns to focus on specific aspects of the patient's symptoms and reasons through the high-level rationale of the disease symptoms. Through qualitative analyses, we have observed that the generated text provides justifications for the classification of certain departments. For example, some doctor's responses include keywords that directly connect to a department, such as "skin," "endocrine," "bone," etc. These would correspond to department labels of dermatology, endocrinology, and orthopedics. Even when the output is not coherent, it produces more medically-informative terms that show an understanding of the patient's input, helping the classifier distinguish between classes. The original BERT encoder's performance is suboptimal due to the absence of a secondary objective during training. Ultimately, we have shown that fine-grained tuning can facilitate core tuning.

6.2 BERT Baseline Takeaways

Based on our experiments, we conclude that using a pre-trained model that is fine-tuned on a similar corpus to our target task produces superior results than just using a general model without fine-tuning. Carrying out a grid search with different hyperparameters can at most achieve a marginal improvement compared to just using the original BERT parameters because the original hyperparameters were already state of the art. Furthermore, freezing specific parts of the models to finetune only parts of the model would not be desirable in our case since the layers are ideally trained in conjunction with all the other layers. Going forward, we plan on using the encoder of our baseline MedBERT-wwm model to finetune the transformer encoder-decoder we will use for our multi-task objective.

6.3 Cross-Attention

Moreover, we implement cross-attention between the decoder's multi-attention heads and the encoder's final hidden state. As these hidden states represent high-level abstractions of the patients' questions, enabling the decoder to perform cross-attention in relation to these hidden states can bolster the decoder's performance. By transferring the hidden states from the encoder to the decoder, the latter carries out cross-attention between the hidden states and the forward pass attention of the doctor's answer during training. This comparison with the ground truth allows the decoder to backpropagate its loss back to the encoder, further refining the model.

6.4 Limitations

Our work faced limitations primarily due to AWS virtual machine constraints and the low quality of some doctor's responses. The virtual machine's training time restricted the number of experiments we could conduct, while memory resource limitations confined us to a GPT-2 model trained on only 1/10 of the Chinese text compared to the base GPT-2 model.

Furthermore, the initial responses from doctors were not always medically relevant to the patient's symptoms, complicating the task for our decoder to generate appropriate responses. For instance, many doctor's responses began with "Thank you for your trust... please send a photo of your injured area" (translated). We cannot assume that a large dataset would compensate for the low quality of these responses.

Despite utilizing an in-domain MedBERT model with greater exposure to medical vocabulary, we observed instances where certain characters in the patient's question were encoded with '<UNK>', the unknown token. This disrupts the flow of the patient's question and hinders the model's ability to effectively learn representations due to limited vocabulary. It would be beneficial to develop a larger vocabulary that encompasses uncommon but critical terms in the doctor-patient space.

Evaluating our LLM's ability to encode clinical knowledge is challenging. Providing high-quality answers to medical questions necessitates both a deep understanding of disease mechanisms and the ability to interpret the patient's self-description of symptoms. Patient descriptions can vary greatly, making it difficult to generate accurate responses. Standard metrics like BLEU or ROUGE scores only measure coherence without the detailed analysis required for real-world clinical applications. Due to the high risks associated with providing diagnoses and suggestions to patients, we prioritize improving the multi-class classification objective, without focusing heavily on the outcomes of the sequence-to-sequence task itself.

7 Conclusion

In our experiment, we have successfully demonstrated that seq-to-seq tasks can improve the performance of classification tasks. However, it is also possible that the reverse is true.

While comparing the outputs generated by the stand-alone GPT2 and MEDBERT's GPT2 decoder, we found that the stand-alone GPT2 generated coherent sentences that were medically relevant yet not pertinent to the patient's specific conditions. In contrast, MEDBERT's GPT2 decoder generated sentences with less coherence but demonstrated significantly more understanding of the patient's specific condition. Due to the low BLEU score of both predictions, we decided not to include the relative results in this paper. However, this observation suggests the possibility of the classification task assisting the seq-to-seq task.

8 Future Work

Having demonstrated that a secondary seq-to-seq task can enhance the performance of the primary multi-class classification task, we aim to investigate whether the reverse relationship holds true as well.

Moreover, we plan to improve the low BLEU score of the decoder in our BERPT model by employing a more powerful version of the Chinese GPT2. Our current work was constrained by the AWS memory limitations, and we hypothesize that utilizing a better fine-tuned and more powerful GPT2 could improve the decoder's BLEU score. As a higher BLEU score indicates a better understanding of the doctor's response to the patient's symptoms, enhancing this metric might further increase the performance of our multi-class classification task.

Additionally, the quality of our classification tasks could be significantly improved by employing a team of human doctors to filter through and clean our data by removing suboptimal doctor responses. Although the online doctors whose conversations we used for our dataset were all clinically certified, we still observed instances of low-quality diagnostic language during our analysis. By refining the dataset, we aim to further enhance the performance of our model.

9 References

[1] Karan Singhal et al. *Large Language Models Encode Clinical Knowledge*. 2022. arXiv: `2212.13138 [cs.CL]`.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[3] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *CoRR* abs/1910.10683 (2019). arXiv: `1910.10683`. URL: `http://arxiv.org/abs/1910.10683`.

[4] Lukasz Kaiser et al. "One Model To Learn Them All". In: *CoRR* abs/1706.05137 (2017). arXiv: `1706.05137`. URL: `http://arxiv.org/abs/1706.05137`.

[5] Ronghang Hu and Amanpreet Singh. "Transformer is All You Need: Multimodal Multitask Learning with a Unified Transformer". In: *CoRR* abs/2102.10772 (2021). arXiv: `2102.10772`. URL: `https://arxiv.org/abs/2102.10772`.

[6] Lin Xu et al. "End-to-End Knowledge-Routed Relational Dialogue System for Automatic Diagnosis". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 7346–7353. DOI: `10.1609/aaai.v33i01.33017346`. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/4722`.

[7] Guangtao Zeng et al. "MedDialog: Large-scale Medical Dialogue Datasets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9241–9250. DOI: `10.18653/v1/2020.emnlp-main.743`. URL: `https://aclanthology.org/2020.emnlp-main.743`.

[8] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[9] Iulia Turc et al. "Well-Read Students Learn Better: On the Importance of Pre-training Compact Models". In: *arXiv preprint arXiv:1908.08962v2* (2019).

[10] Jiao Li Feihong Yang Xuwen Wang. "Using BERT in Chinese Medical Natural Language Processing to Explore Applications and Research". In: *EB/OL* (2021).

[11] Zeyao Du. *GPT2-Chinese: Tools for training GPT2 model in Chinese language.* `https://github.com/Morizeyao/GPT2-Chinese`. 2019.

[12] Junshu Wang et al. "Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical Bert - Journal of Cloud Computing". In: *SpringerLink* (Jan. 2021). URL: `https://link.springer.com/article/10.1186/s13677-020-00218-2`.

10 Appendix

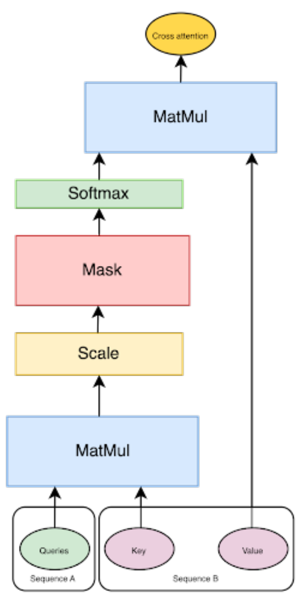

Figure 2: Mechanisms for Cross Attention

Table 2: Qualitative analysis of the medical question answering system

Patient's Question	
脂溢性脱发脱发，皮肤科医生说脱发是内分泌失调引起的，雄激素分泌过多，和油脂分泌过多使其毛囊堵塞请问脱发是不是内分泌失调引起的啊，我看内分泌科是否能解决脱发问题。	”Seborrheic hair loss, the dermatologist said that hair loss is caused by endocrine disorders, excessive secretion of androgens, and excessive secretion of sebum that clogs the hair follicles. Is hair loss caused by endocrine disorders? Can the problem of hair loss be solved by seeing an endocrinologist?”
Doctor's Response	
Ground Truth	
你好，男性脂溢性脱发一般与内分泌无关，非那雄胺治疗脂溢性脱发最有效的药物了。建议同时外用米诺地尔溶液	”Hello, male seborrheic hair loss is generally unrelated to endocrine disorders. Finasteride is the most effective medication for treating seborrheic hair loss. It is recommended to use minoxidil solution topically at the same time.”
Prediction, stand-alone GPT2 decoder	
西南皮肤病医院科室：中国人民解放军第 302 医院用药情况：雄性激素异常脱发，雌性激素，是激素分泌过量引起的，与皮肤有机物排泄障碍是直接相关。一旦雌雄同源，有可能会诱发激素失调的产生，应当考虑使用某种激素，如雌激素类药物或激素类药物治疗，同时可以进行另一种激素的减轻症状，如皮炎，血小板减少等，有一定疗效。建议：患病的女性服用雄激素药，但必须经过消炎	The department at Southwest Skin Disease Hospital is: Department of the Chinese People's Liberation Army No. 302 Hospital. The medication situation is as follows: male hormone abnormal hair loss, estrogen is related to excessive hormone secretion and directly related to organic matter excretion disorder of the skin. Once estrogen and androgen are of the same origin, it may induce hormone imbalance. Therefore, consideration should be given to using a certain hormone medication, such as estrogen or hormone medication for treatment. At the same time, another hormone can be used to alleviate symptoms, such as dermatitis, thrombocytopenia, etc., with certain therapeutic effects. It is recommended that female patients with the disease take hormone drugs, but they must also be treated with anti-inflammatory drugs.
Department Classification	
皮肤科	Dermatology