Building a Medical Q&A System Based on Deep Learning and Knowledge Graphs

Zhi Wang

1840782209@qq.com

College of Computer Science, Hubei University of Technology, Wuhan, China

Abstract. The aim of this paper is to build a localized question and answer system for medical big language models. Deep learning is used to equip the big language model with the ability to intelligently recognize and generate answers. In this paper, the completeness of knowledge storage of knowledge graph is used to provide the model with a more comprehensive knowledge of the medical specialized domain, which in turn is transformed into a dataset that can be used to train the model. In order to make the model more accurate, langchain is introduced as a new model prompt word design method, so that the new model inputs have identification and collect contextual information according to the established input index to expand the knowledge of the inputs, and then merge with the prompt to generate the inputs needed by the new model format, and the model makes use of the text2vec to deal with the Chinese participle in advance in the process of processing, so as to achieve a slight decoupling and through the comparison, the model is able to generate the answers in a more comprehensive way. The model uses text2vec to process Chinese word separation in advance during processing, achieving a slight decoupling and, by comparison, text2vec-english has a better effect on Chinese word separation. In order to fit the model localization deployment and Chinese Q&A, this paper chooses the Chinese Q&A model chatGLM-6B model learning to achieve Q&A, and uses qlora to fine-tune the model data, which will achieve the localization deployment with small memory.

Keywords: knowledge graph; langchain; text2vec-chinese; chatGLM-6B; qlora; localized medical Q&A

1. Introduction

Medical problems is a major livelihood issues in any country. Doctor-patient disputes, large-scale dispersed knowledge data to make medical universal access is very inconvenient to process the growing expansion of large-scale data, and the correlation between the knowledge of the various pieces of information can not be demonstrated with the traditional data processing storage. Moreover, the problem of difficult to see a doctor exists in underdeveloped areas and hospitals, excellent medical Q&A online system can alleviate this problem to a certain extent.And after reducing hallucinations, the model's answers will not be artificially false information transmission caused by low internet barriers like search engines.

Since Berners-Lee, the father of the World Wide Web, proposed the Semantic Web in 1998, which expresses the understanding of the objective world on electronic carriers to form a conceptual standardization, textual semantics has become an increasing topic of concern. 2012,

PMIS 2024, March 15-17, Changsha, People's Republic of China

DOI 10.4108/eai.15-3-2024.2346400

after Google launched the Knowledge Graph, there are more practical and convenient applications for dealing with semantics. In agriculture, medical, biological, financial and other aspects of the knowledge graph show emerging. Knowledge graphs can, to a certain extent, deal with knowledge extraction, representation, reasoning, Q&A, storage and other problems, describe entity relationships, associate domain knowledge, solve the problem of dispersion, and are friendly to text processing, so they have become a new research hot spot. It is because of the extensive and closely associated characteristics of knowledge graphs, large-scale knowledge graphs contain a huge amount of information, how to make good use of knowledge graph data is a major problem[1].

With the development of the Internet, the era of big data coming, traditional data processing methods can not meet the needs of machine learning applications, even for deep learning, greatly simplifying the traditional machine learning algorithms and processes. More importantly, Deep learning improves the precision and accuracy which the traditional machine learning algorithms can not reach. Deep learning is optimized for data analysis and the modelling process is shortened, with neural networks unifying the wide variety of algorithms used in the original machine learning. Before deep learning is officially used on a large scale, data collection and processing, trying a variety of different feature extraction machine learning algorithms, or combining a variety of different features on the data classification and regression is a very complex part of the traditional machine learning, while deep learning is to optimize the data processing and model building process. Therefore deep learning has become the current artificial intelligence hot spot.

Since the transformer framework was proposed, the attention mechanism has been widely used in natural language processing, and its secondary frameworks BERT and GPT have won awards in related fields in successive years[2], especially the research of BERT. However, there are few products that can really be widely used commercially. Since the emergence of chatGPT in 2022, the natural language Q&A application has reached its peak, and its advantages of universality and high availability have attracted the public's attention, and the breakthroughs in natural language processing have made the research on GPT a hot topic, and a variety of model fine-tuning methods have emerged[3], among which the qlora[4] fine-tuning method proposed by the University of Washington has become a hot topic in recent years[5][6].

The SELF-INSTRUCT[7] method was proposed in a paper in 2022 as a method for constructing instruction datasets, which is efficient and usable, and the instruction data generated with fine-tuning input data into the GPT dramatically improves the effect of the original GPT-generated data[8][9]. The use of professional instructions in the professional field also improves the effect over open instructions, which largely solves the problem of data quality, diversity, and creativity, but it requires fine-tuning the data on closed-source GPT, and the high cost of the openAI api prevents the construction of small Q&A data. The open-source alternative is recently proposed by the natural science processing direction model DARWIN[10], which uses open-source models such as LLaMA[11] and RWKV[12], introduced the Scientific Instruction Generation (SIG) model to automatically generate instructions from scientific text. This removes the need for manual extraction or domain-specific knowledge graphs and effectively injects scientific knowledge into the model, using open source models also removes the costly expense of data construction, while DARWIN researchers have found their model data to be more effective[13].At the same time, improving data diversity and enhancing generalization, introducing Langchain to construct document knowledge and generate instruction data, selecting reasonable prompt[14] templates to enhance self validation of generated responses and cots when generating datasets, and optimizing self consistency to select relatively better Q&A[15][16].

Simultaneously introducing an optimizer and contrasting loss to adjust the model learning rate in an attempt to optimize production speed and move towards fast modeling. Although the actual speed may not be immediate, there is indeed a feeling of acceleration this time.

Therefore, in this paper, we choose to collect the knowledge graph of complex and diverse medical knowledge obtained online, based on which we use knowledge extraction to get part of the quality data,then feed the data to a large language model to get a domain-specific Q&A model, and transform the knowledge graph into a dataset of knowledge Q&A instructions to be processed. The dataset is used to construct a private Q&A system based on Langchain and ChatGLM-6B[17][18], and then QLora is used to tune the chatGLM-6B model to get a small private Q&A repository that can handle large-scale data, and the efficiency is close to the large-scale and latest GPT Q&A model, and the data in this paper is based on medical data, which is more standardized and comprehensive and practical, and it is a small-scale based of professional knowledge base that can be deployed on local links and is more professional than GPT3 Q&A.

2. Background and Significance of the Study

2.1 Research Background

2.1.1 Natural Language Processing

The concept of the semantic web and knowledge base mentioned above has long been mentioned, and the early ones are mainly small-scale specialized knowledge base construction, whose using methods are mainly semantic analysis. However, the earliest manual annotation used in processing is time-consuming and laborious. Later, under the development of machine learning, semantic parsing based on weakly supervised learning methods began to explore, but machine learning is more widely developed in image processing and other natural language has been lagging behind other research. The previous mainstream research methods are mainly divided into two categories: methods based on semantic analysis and methods based on information retrieval.

Methods based on semantic analysis use logical expressions instead of natural language to obtain answers from a knowledge base. While the information retrieval based methods focus on extracting relevant features to rank the candidate answers.

With the continuous progress of neural network technology, Bordes et al. introduced feed-forward neural network to process the question and the answer into two vectors separately, then the dot product of the two vectors is finally used as the score of the candidate answer.Li et al. used CNN to process the type of sentence, relationship, and context of the query and answer respectively, finally the dot product of the vectors composed of the semantics of the three types of information is weighted to get the final score. The final score is obtained by weighting the dot product of the semantic vectors of these three types of information. In recent years, attention mechanisms have been introduced to train different sentence representations for different contents of candidate answers.

In the past two years, ChatGPT developed by OpenAI has become a big hit, which has pushed the research of natural language and Q&A system to a new climax[19].

In the Chinese domain, the KBQA task in 2016 provided a large-scale knowledge base, which is the prototype of Chinese Q&A processing in recent years. Langchain generates a knowledge base based on Q&A data such as individual documents, which has also become a popular issue in recent years. Convolutional neural networks and GUP models[20] were first used for semantic representation of interrogative sentences CNN networks were developed for named entity recognition, BiLSTM and CNN were used to specifically implement attribute mapping. With the deepening of research, the chatGLM-6B model researched by Tsinghua University becomes the leader in the Chinese language field. And the methods based on model tuning and cue word engineering emerged in language model processing in recent years can improve the model capability efficiently.

2.1.2 Knowledge graph research

In recent years, AI has been widely used in various fields of life. Among them, it is mainly divided into simulating the physical structure of human nerves (i.e. neural networks) and simulating human brain memory (i.e. knowledge graph) simulating human minds.

In 2012,Google introduced the Knowledge Graph, which is a structured semantic knowledge base, which uses triples to represent knowledge (entity 1-relationship-entity 2 or entity-attribute-attribute value), structure knowledge, be good at describing relationships and extract and associate complex and scattered data, and the research of knowledge graph has gradually become popular in recent years.

With its powerful semantic processing and open interconnection capabilities, the knowledge graph can lay a solid foundation for knowledge interconnection on the World Wide Web, making the vision of the "web of knowledge" proposed by Web 3.0 possible.

For complex human cognitive knowledge, the computer is convenient to display, but the computer is difficult to actively identify, and the relevant information is scattered, such as natural language, AI voice communication, recommendation system, question answering system, etc., here we can add a knowledge graph on the basis of neural network reinforcement learning, the human cognition is recognized by the computer, and it is connected in series. For example, the book "Those Things in the Ming Dynasty" contains the author, outline, type, evaluation, etc., and its data is represented in the form of a knowledge graph and recognized by the computer, and the value of clear, concise and comprehensive knowledge is huge, and this article analyzes data and trains it to build a question and answer system.

The knowledge graph based on neural networks can help us deal with this kind of problem.First of all, obtaining the knowledge graph from Internet. Knowledge graph processing is turned into a dataset that the computer can recognize the processed data structure. Finally, a dataset of learning training questions and answers is provided for model. In this way, we can extract the key information needed from the complex text and process it into the Q&A system.

This paper introduces deep learning and related algorithms to form a question answering system. Through dataset training and qlora parameter tuning, the optimized model can be obtained close to the ChatGPT effect. And the comprehensiveness of knowledge graph lays a foundation for data provision. This paper uses the open source scientific instruction generation (SIG) model method to build a question answering instruction dataset like 'Task: -Input: -Output:' mode, so that the model can be specialize, and the question answering system can be deployed on the local link.

2.2 Research significance

ChatGPT came out last year with its effective text-generated Q&A and its high scalability which makes it comprehensive in terms of the domains involved, deal with more complex in terms of processing data, and breakthrough in natural language processing. However, the shortcomings of these models are equally obvious:

- Chinese language limitations:

ChatGPT as a model based on the development of English datasets, in dealing with the Chinese language has just begun, the model of Chinese learning should be further. This paper uses the Chinese medical more comprehensive knowledge graph data coupled with the Scientific Instruction Generation (SIG) model approach to build a Chinese instruction dataset which computer can be better identified by the prompt type.

- Optimization of deployment environment:

ChatGPT-3 is a model with 175 billion parameters, and its deployment is extremely demanding in terms of graphics cards, models, datasets, etc. Optimizing deployment and computation on common platforms can both save on setup and maintenance costs, be more research friendly, and improve efficiency and wider application of Q&A-like systems.

- Small private domain applications:

ChatGPT deployment is based on the cloud, can not be deployed offline, and Samsung not long ago because of the GPT information leakage incident, private security is a problem. So we want to build a more efficient, lighter and safer application in small private local domains, and ChatGPT3 to ChatGPT4 and other iterative expansion is very inconvenient, our private dataset at any time to feed the ability to get new learning results.

- Professional domain optimization:

Although ChatGPT is highly extensible and its applications basically cover all major domains, it performs generally in specific domains, such as the medical domain in this paper, for example, it is not explicitly trained in the medical domain, which leads to unsatisfactory accuracy in diagnosis, drug recommendation and other medical suggestions.

Therefore, this paper will aim to optimize chatGPT for these large model inconveniences and explore methods for large-scale use of knowledge quizzes.

3. Research Content

3.1 Research content

In this paper, the knowledge base query and retrieval of the medical knowledge graph will be carried out to establish the question answering system. In order to establish a robust knowledge base and realize the significance of the research, this study is based on the following aspects:

- Research uses medical knowledge graph and command data construction method, DARWEN model type, to build question and answer data to achieve professional question answering, and sets up a variety of prompt forms to make full use of knowledge. Optimizing the prompt template through cots to obtain optimal Q&A results, and in order to filter the optimal answers, self instrument generates data by using multiple cots to filter for the best quality data. This is achieved by introducing self consistency decoding.The public and self-built Chinese medical knowledge base is adopted, mainly referring to the cMeKG knowledge graph. The medical knowledge base builds instruction datasets around the maps of diseases, drugs, examination indicators, etc.. And the fields include complications, risk factors, histological examination, clinical symptoms, drug treatment, adjuvant therapy, etc. Improving data diversity is also an aspect of improving model performance. Utilizing Langchain to generate high-quality datasets of professional documents enhances usability[21][22].
- Several techniques were employed to train LLM that was easy to deploy. It is worth noting that we were able to train ChatGLM-6B on ordinary personal GPUs in a short period of time, which means that LLM with private healthcare use can be very affordable, while efficient processing of text such as documents can also improve the analysis efficiency of private knowledge text[23].
- GPT, the secondary architecture of the transformer with the attention mechanism of the most popular natural language processing architecture, is adopted, and then the pre-training of the model is improved into vector, which is a computer recognition language, and the natural language is encoded.
- The use of langchain+chatGLM large language model with qlora fine-tuning, not exactly the same as the original finetune of the transformer[24][25], using input, using representations to find similar related paragraphs, cracking the problem and related paragraphs to achieve the effect of question and answer, while finetune uses specific fields, can not be effectively expanded, high fine-tuning cost, poor real-time data, experience-based, slow update iteration. The model is a security model that is fine-tuned to obtain privatized data based on domain data, and the model Lora model and base model are combined to obtain 4-bit quantification to facilitate qlora tuning to obtain low graphics memory operation.Improve the optimization of prompt on question answering datasets and outputs to generate optimal solutions, and optimize the prompt template for cot[26].
- Introducing optimizers and contrast-loss to accelerate convergence and achieve faster and more accurate results.

3.2 Significant point

Based on the medical knowledge graph, this paper conducts deep learning training, so that the model has the ability to question and answer the medical knowledge. The total points in this article are as follows:

- Based on the new encoder processing context generation vector, the langchain+chatGLM large language model is fine-tuned with qlora, which makes the dialogue more efficient and builds a miniature model suitable for private professions, prevents data leakage, applies specific fields, improves training effects, and obtains lower graphics card memory.
- This article not only handles private questions and answers, but also has a good effect in efficiently reading document files for question and answering, making the ability of the language processing question answering system more comprehensive.
- Using the langchain+chatGLM large language model, qlora is used to fine-tune chatGLM-6B to process large data sets, and the prompt prompt word is directly trained to the model to obtain new responds, so as to build a private question answering library.Improve the optimization of prompt on question answering datasets and outputs to generate optimal solutions, and optimize the prompt template for cot.
- Using the DARWEN model to construct a new type of Chinese medical Q&A instruction dataset (i.e. JSON format) based on the prompt template construction method, the prompt template is cot-tuned to obtain optimal Q&A. At the same time, in order to filter the optimal answers, multiple cots are used to filter the best quality data when self instrument generates data, which introduces self consistency decoding.
- Attempt to improve convergence and optimize answer retrieval to increase accuracy and speed.
- Improving data diversity is also an aspect of improving model performance. Utilizing Langchain to generate high-quality datasets of professional documents enhances usability.

4. Method

4.1 Acquisition of training sets

The quality of fine-tuning model training depends on the quality of instruction data, and fine-tuning instructions can improve the quality of the model without relying on other external methods, so the construction of a good instruction dataset can directly improve the training quality. At present, the medical instruction data set is mainly in English, Chinese Q&A first manually written data is time-consuming and laborious, secondly, Chinese instruction data set is less, currently open source many datasets, but either use question answers, or translate English data, these massive field data effects are unpredictable, or the use of InstructGPT fine-tuning data, the cost is extremely high, the cost of scientific instruction generation (SIG) model is low and the way to generate data by yourself is more creative.

Chinese knowledge graph has developed relatively maturely, there are many Chinese medical knowledge graphs, structured knowledge can enhance the knowledge expression of LLMs. In this paper, the Chinese medical knowledge graph will be fine-tuned with the scientific instruction generation (SIG) model method to fine-tune the open source model Vicuna-7B[27] that performs well in the Chinese field, and directly convert the knowledge graph into an instruction dataset to solve the shortcomings of the Chinese instruction dataset. This paper not only enhances the data availability of the training set, but also the robustness of the model through instruction tuning[28], and the data robustness of the scientific instruction generation (SIG) model is better than the commonly used chatGPT translation dataset.

The experiment is also to choose the stanford_alpaca[29] template prompt word 'Task: -Input: -Output' form to set the prompt_dict[30][31][32][33][34][35]:

{

"description": "Template used by Med Instruction Tuning",

"prompt_input": "Here is a question that uses medical knowledge to answer the question correctly. \n#### Question:\n{instruction}\n### Answer:\n", "prompt_input".

"prompt_no_input": "Here is a question that uses medical knowledge to answer the question correctly. \n### Question:\n{instruction}\n### Answer:\n", "response_split".

"response_split": "### Answer:"

}

To modify the template of DARWEN-SIG, and then use the langchain file interface form to get part of the Q&A instances as seed QA through the relational extraction of the knowledge graph, finally input them into the Vicuna-7B model in csv file format with prompts to generate the DARWEN-SIG model of medicine.

The DARWEN-SIG model generates instructions with the given prompts and Q&A pair inputs to get a new DARWEN-SIG (switching to instructions instead of prompts), so that a rich set of instructions is obtained based on the knowledge graph and the model. Finally training to get QA: Using the trained DARWEN-SIG, Q&A pairs are automatically generated based on instructions and inputs at low cost.Such as Table 1. (These QA pairs can be directly converted to instruction data).

Table 1 Example of command tuning mode

Introduction:
Translate the following sentences into Chinese.
Input:
What are the causes of headaches?
Output:
(in Chinese)What are the causes of headaches?

The generated question-answer pair used translation instructions are namely in the form of the data in Table 2 (used to train the test set).

Table 2 Input Instructions

	Instance(in Chinese)	Instance(translate into English)
Question	(in Chinese) Xiao Zhang has been feeling unwell recently with palpitations and shortness of breath. Physical examination revealed an enlarged and weakened heart beat.	Xiao Zhang has been feeling unwell recently with palpitations and shortness of breath. Physical examination revealed an enlarged and weakened heart beat.
Answer	(in Chinese) Zhang may be suffering from myocarditis, and tests such as ECG and cardiac ultrasound are recommended to confirm the diagnosis. The treatment plan includes the use of medications such as prednisone, Shengqin Drink and adenosine triphosphate, while proper temperature control and good nutritional status are recommended.	Zhang may be suffering from myocarditis, and tests such as ECG and cardiac ultrasound are recommended to confirm the diagnosis. The treatment plan includes the use of medications such as prednisone, Shengqin Drink and adenosine triphosphate, while proper temperature control and good nutritional status are recommended.

The Q&A pairs from the training test set are then inputted to really generate the model to get the target instruction data. And the DARWEN model is also generated in the form of input and instruction merged data in the same form as GPT, such as the json format mentioned above:

{"instruction": (in Chinese)Xiao Zhang has been feeling unwell recently with palpitations and shortness of breath. Physical examination revealed an enlarged and weakened heart beat., "input": "", "output": (in Chinese)Zhang may be suffering from myocarditis, and tests such as ECG and cardiac ultrasound are recommended to confirm the diagnosis. The treatment plan includes the use of medications such as prednisone, Shengqin Drink and adenosine triphosphate, while proper temperature control and good nutritional status are recommended.)}

The generated dataset is based on question answering examples, and in order to improve data diversity and conform to clinical question answering forms, this article also constructs instruction datasets based on medical texts and generates self instruction datasets. Due to the localization of deployment, the dataset is still only targeted at text forms.

When selecting Langchain for input processing in this article, the Structured Output Parser is used to convert professional texts into JSON format. Combined with the chatglm in this article, clinical instruction datasets related to professional texts are generated based on the above prompt. The high-quality dataset generated from the text and the knowledge graph are used to fine tune and train the large model chatglm based on self instruction new data[36].

4.2 Model Architecture Rationale and Selection

The large language model used for natural language processing is a large deep neural network. It deals with textual data to recognize human language data, and transforms data into vector format to train recognition, on the basis of which it understands and generates human language associated with the processed data. Its implementation is based on the following three steps:

- Semantic representation of text based on word embedding;

- Transformer based on attention mechanism (e.g. Transformer);
- Self-supervised learning based on predicting the next word.

In the first step of these, the data to be processed is vectorized (a form of data that can be recognized by the computer). Before the earliest word-document matrices used were too complex and labour costly, the advent of neural networks introduced word embedding methods. Since then, the Word2Vec model has been further developed to deal with word vectors specifically.Word2Vec is a simple network structure and efficient training method to deal with large-scale text data. However, the shortcomings of Word2Vec are also obvious, for example, there is no concept of pre-training, it is not combined with the later steps, and the processed data can not be adapted to the complex big data. What's more, the processing is not contextualized and the results are not good enough.

In recent years, word embedding technology has entered the era of pre-trained language models. models such as BERT and GPT can not only generate high-quality word embeddings, but also process utterance-level and even longer texts. This provides a powerful tool for solving NLP tasks. text2vec-chinese based on Chinese has also been effective in recent years.

Prior to ChatGPT, BERT, another secondary architecture of Transformer, was the focus of NLP research. due to the excellent computational power and unique attention mechanism of Transformer, it is able to train large models with billions of parameters and has outstanding features in translation, summarization, and text generation, etc. BERT has been a research hot spot before, and various derivative models have been studied and BERT has been a hot research topic before, and various derivative models have been researched and used.

The ability of chatGPT to process B-scale data is outstanding, MedPaLM has outstanding results in biomedicine, and chatCAD's ability to process images provides interactive assisted diagnosis potential. ChatGLM researched by Tsinghua is very close to ChatGPT in Chinese domain ability, and chatGLM-6b is a Chinese trained model, this paper is going to perform Chinese fine-tuning, Chinese training data with the same language Chinese fine-tuning dataset may work better. In this paper, we choose this model for qlora fine-tuning to make a private medical chat model, and select the text2vec split word with better effect theory on the preprocessing of split word.

4.3 Qlora fine-tuning training

As mentioned before, GPT is a secondary architecture based on the transformer model, which of course has the mechanism structure of transformer.The ChatGLM-6B model is based on the Generalized Language Model (GLM), which has 6.2 billion parameters and 7 million trainable parameters, and the fine-tuning of the transformer has always been an important method to improve the quality of the model.

The transformer is a coding model that encodes utterance words and positional inputs into vectors, and then calculates the relationship between the target word and the context based on the coding matrix and introduces an attention mechanism to calculate the attention coefficient of each target word to other words for word training, which is more robust. After that, Fine-tuning was used to optimize the model, but the model was costly and difficult, so various optimization schemes emerged.

Adapter is almost the earliest Fine-tuning-based model. With keeping the parameters of the transformer-based model unchanged, it add some parameters for pre-training, and then fine-tune the data downstream to make the model better. the structure of the adapter layer is very simple, and projects down to a smaller dimension. After a layer of nonlinear activation function, the model is projected upward to the original dimension. In addition, there is a residual connection between the inputs and outputs of the entire adapter layer. This type of adapter is also figuratively called bottleneck adapter, i.e., the new parameters to be added in the lower layer are trained at the connection, as shown in Fig. 1.

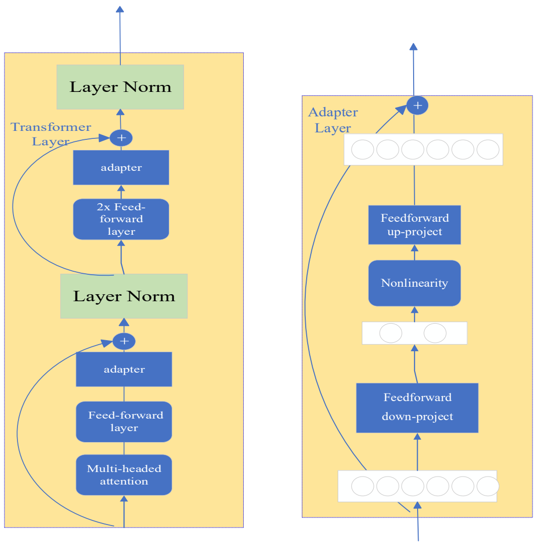

Figure 1 adapter model

LN (Layer Norm Normalization) in Transformer is to do normalization (calculate the variance and mean of each sample) for the dimension of each sample, that is, to do the operation for different features of a single sample. Therefore LN can be independent of the number of samples. For each tensor (multidimensional array) to be computed, higher dimensional matrices, vectors are created. Multiple tensor matrices are computed in a sequence, the time_step dimension of the batch_size tensor matrices are processed into vectors at the same time. batch_size is based on the specific number of data batch inside the statement sample data, divided into x batches of discretionary size processing, batch_size is the average batch of the number of matrix data(batch/x).

The essence of LoRA is to insert a number of new parameters, called adapter, into the original Transformer model.During training, the parameters of the original model are frozen and only the parameters of the adapter are updated. For different base models, the number of parameters of the adapter is generally several millions to tens of millions, adding different sets of adapter parameters A,B decomposition matrix to change the weights, while A and B contain trainable parameters. Random Gaussian initialization is used for A, and zero initialization is used for B.

The advantage of LoRA is the ability to use less GPU resources to fine-tune large models in downstream tasks. In the open source community, developers have used LoRA to fine-tune

Stable Diffusion with very good results. However, the parameter fine-tuning utilization of LoRA is not high, and the cost of fine-tuning 7 million trainable parameters for large models is high[37].

In this year, the University of Washington proposed the qlora (4-bit base model + LoRA) fine-tuning method, quantizing data on the basis of lora can reduce the memory occupation and efficiency enhancement in a large number of ways, which gives the necessary conditions for the civilianization of large models. Quantize quantization can improve the accuracy problem of lora, the common accuracy of full-precision float32(FP32), half-precision float16(FP16) and bfloat16(BF16), which can put the medical data into the relevant precision of the instruction set, and the computer can identify the accuracy of the storage match, while the instruction can support parallel computing, but also greatly reduces the model occupancy, and at the same time, people found that the half-precision float16(FP16) and bfloat16(BF16) i.e. (full precision float32(FP32)) after quantization the model size is reduced, the performance is instead improved, the 16-bit float with 4-bit to express, equivalent to make the data more closely match the computer calculation, and secondly in the model of all the fully connected layer at all the adapter inserted to increase the training parameters, to make up for the loss of performance caused by the precision, equivalent to increase the fine-tuning to maintain the precision performance This is the best performance fine-tuning (adapter upgraded version) method under the most resource-saving (model reduced) situation. It is due to the characteristics of qlora that the model can be fine-tuned on a normal gpu. The chatGLM and qlora files are merged to form a new model, and the model of the chatGLM-6B is quantized in 4bit, the model file is about 3.7G, which is loaded into the GPU. G, loading into the GPU takes up less than 5G of video card memory.

4.4 Langchain

Langchain is not only a solution for prompt design, but also for Q&A and document reading. Langchain is an application framework that includes components, interfaces and other tools to help create applications related to language models such as the ChatGLM in this paper, easy interact with ChatGPT, and integrate a large number of related components, which simplifies the construction of applications. Langchain is currently a popular framework tool for language modelling.

The function of cue design is to optimize the previous output unreliability problem of ChatGPT, so we use a cue designer module to preprocess user input.

The user's knowledge base (i.e., dataset) is encoded into a vector knowledge base. And when questions are asked, the keywords about the disease entered by the user are extracted and encoded, and a vector correlation algorithm (e.g., cosine algorithm) is used to find the few most matching pieces of the knowledge base, and the knowledge found from the knowledge base with the user's inputs is constructed and used as a Q&A in its classical template way:

{

"known information":"{context}".

Answer the user's question concisely and professionally based on the above known information. If you can't get an answer from it, please say "Can't answer this question based on known information" or "Not enough information provided", no making up in the answer is allowed, and the answer should be in Chinese.

The question is: {question}

}

4.5 Vector model text2vec specific principles

Human natural language utterances want to be read by the computer must be transformed into vectors, this paper proposes text2vec method, similar to word2vec, bert, gpt, bilstm, etc. to construct a model of split word vectors. This model also uses word2vec (based on Tencent's 8 million Chinese words training), SBERT (Sentence-BERT), CoSENT (Cosine Sentence) three representations of training, can be understood as the use of bert vectorisation on top of Bert. Comparison carried out through online information, this split word effect is more ideal.

4.6 Prompt enhances interpretability

In LLM, prompt words are the key to generating professional answers. Given specific prompts, the model should know the approximate detailed direction of the answer. Here, a specific description of prompt will be provided. After learning Prompt, it quickly evolved into Demonstration Learning (DL) and Context Learning (ICL), adding some examples of zero shot or few shot[38][39] to generate answers in large languages. CoT: Compared to the Stanford template Prompt for summarizing answers, the thought chain has added some intermediate processes in the generation process.

CoT also has a few shot form. With the feed shot CoT, the model provides an intermediate reasoning stage, allowing the model to learn the reasoning logic and thinking methods in the intermediate process. According to the COT prompts, the answer will be added to the analysis, and the answer data will be more accurate and transparent, following the template and rich in content. This article provides a dataset, selecting some as feed shot examples to fill in the following template:

Add let's think step by step in Langchian's prompt to obtain rich answers.

During training, based on the single question and answer format dataset generated by self instructions and langchain methods in this article, these question and answer formats are designated as long question and answer formats. When training and answering question and answer instruction data:

{

Instruction: The following are effective and scientifically rich explanations for medical issues, answer them step by step in Chinese.

Question:

Explanation:

Answer

}

As mentioned earlier, Langchain not only generates datasets, but also provides a document question and answer format, which is very useful for real-time document text knowledge that can be provided at any time. When calling this interface for document reading, in this case, the model makes an enhanced prompt for the following cot:

{

"Known information": "{context}",

Based on the known information above, provide concise and professional answers to user questions. If you are unable to obtain an answer from it, please say "unable to answer the question based on known information" or "insufficient information provided", and it is not allowed to add fabricated elements to the answer. Please answer them step by step in Chinese.

The problem is: {question}

Process:

Answer:

}

4.7 Self consistency enhances model inference

The above CoT inference chain is a model that selects a self perceived optimal inference result from the inference process, which is the direct source of the model's potential hallucinations. To solve this problem, Google proposes self consistency to improve it. However, let me mention in advance that this method will increase model training time, which is not in line with the requirements of this article. Therefore, this method will be applied to the construction stage of the dataset to optimize it and reduce hallucinations:

- Using multiple thought chains (CoT) to suggest language models;
- Sample from the decoder of the language model, replace the decoding in the CoT prompt, and generate a set of different inference paths[40];
- Obtain the aggregated result by selecting the most consistent answer in the final answer set.
- When constructing training data, it will consume a lot of resources. Although it does not affect deployment, in order to facilitate construction and because the model performance will be saturated, the number of thought chain reasoning is set to 5, which slightly enhances the data quality for the original CoT's greedy selection of only one decoding strategy.

4.8 Adjusting the learning rate and attempting to increase training efficiency

The essence of training a model is to converge the loss function as much as possible during the encoding process, as shown in Figure 2.

Due to the random initialization of model parameters, in the initial stage of training, they may be far from the lowest point or relatively close to the lowest point. In order to enable the model to converge, smaller learning rates are more likely to achieve the target than larger learning rates, at least not causing the model to diverge.

In an ideal state, the learning rate is dynamic: there is a higher learning rate when moving away from the lowest point, and a lower learning rate when approaching the lowest point.

Properly adjusting the learning rate will enable the model to quickly converge to the optimal parameters at the beginning and reduce redundant parameters at the end, thus attempting to improve efficiency and save time.

The calculation of text similarity after encoding in this article is:

Due to the involvement of classification problems, the loss function adopts the cross entropy algorithm and uses the LAMB optimizer to accelerate convergence to the optimal parameter values considered by the model. Therefore, the set parameter learning rate is also extremely low: 0.01.

4.9 Contrast loss increases ROUGE-L similarity

The segmentation in this article is open-source text2vec. In order to increase retrieval strength and enhance robustness, the contrast-loss method is introduced to enhance similarity:

When we select a context, we can select several related similar problems and several unrelated problems, then calculate the similarity between these problems and the context, wait for a similarity vector, and then use the classification encoding of one hot to set the related to 1 and the unrelated to 0, in order to calculate the loss and retain the correlation.

5. Experimental Process And Results

5.1 Experimental process

5.1.1 Acquire dataset

In this paper, the use of easy access to the Chinese medical knowledge graph plus darwen model method is transformed into a command dataset that can train the Q&A, and this paper only requires the model Chinese inputs, the output is generated by itself, so this paper only needs to retain the model inputs for the Chinese command dataset can be stored in the json file. Generation process is described in detail in the previous section.

Here is an introduction to the Langchain interface for processing text and generating professional clinical Q&A steps. Select high-quality medical text on PubMed, and introduce the Langchain interface to define the JSON Q&A template format that needs to be converted_ Instructions, using the StructuredOutputParser function to receive return instruction data and fill it into the prompt template. This class can be used to build prompt word templates based on professional documents, thereby increasing the professional dataset through external knowledge.

5.1.2 Training

Before the training data in this paper, we first perform model merging to get the model in qlora form, in order to further save the memory of the graphics card, that is, in this paper, we merge the ChatGLM-6B model tuning Lora model and base model and quantize them into 4bits to get the qlora tuned model, which can be run on less than 8G of the graphics card memory, and when merging, we choose the text2vec-english mode of split word decompression into the model file.

The data from the processed dataset is fed into the qlora-tuned ChatGLM-6B via Langchain in training and test sets. With the original model parameters unchanged, in this paper, qlora is fine-tuned to add new parameters to the connection layer with an accuracy that is consistent with the form of instruction data storage, which affects the downstream module through a nonlinear activation function, while using parallel multi-distributed adapters and increasing the parameters to ensure the model performance. The accuracy of such model processing will make the instruction data more suitable for computers to read and compute, thus increasing the speed, saving resources, and can get high accuracy and reliable Q&A system at a very small cost[41].

Subsequently, this paper uses Langchain to process the input again, and even provides a document text reading interface to process a wider range of data with greater practicality.Langchain extracts the keywords of the input to improve the accuracy and reliability of the model to recognize the utterance. It also simplifies the complexity of the data input to the model in this paper, so that the computation is reduced significantly, and finally get the lightweight, efficient and pervasive Q&A system in this paper.

5.1.3 Principle of using langchain+ChatGLM

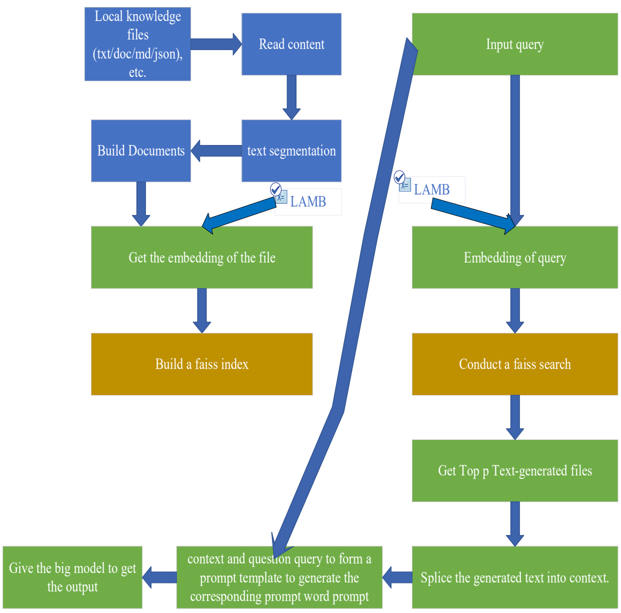

Fig. 2 Langchain and chatGLM operation principle

As shown in Figure 2, this is the principle of operation of this system. In this paper, we get the trained system and then test it for use, read the test set or local knowledge document from the local area, carry out the operations such as slicing to construct the vector of text, and use the vector to get the faiss index; when the user inputs the relevant questions, the system internally will use the question vector to get the text answers with different similarities using the faiss index, and splice the K texts with high indexed similarity to get the answers. At this point, the loss value must be considered. The learning rate is fixed at the beginning, and high learning rates are used for model learning to improve model efficiency. However, in order to converge and obtain an efficient model for optimization, the learning rate will be adjusted lower, reducing unnecessary participation in later data training. Therefore, the traditional cosine learning rate scheduler is introduced to adapt to various stages of learning, adjust learning speed and training set utilization. At the same time, in parameter settings such as batch 4, the optimizer is used to maximize data utilization and reduce loss rate. The most commonly used and most effective optimizers are AdamW and LAMB, while LAMB is the choice because it can adaptively correct model gradient updates, improve model efficiency, reduce time overhead.And use the questions and answers to The questions and answers are merged according to the set prompt template to get the LLM prompt, and finally get the Q&A results. langchain provides great convenience for small-scale application of local knowledge, install langchain package when using it, and then import langchain classes into the big model directly. class into the big model can be fused to use.

Experiment with localisation:

Directly change the parameter path in the model loading surface to

"local_address/chatglm-6b-int4",

"init_embedding_model=local_address/text2vec-chinese" .

Quantitative model for experiments:

The model startup run is to execute web_demo.py, web_demo2.py, api.py, cli_demo.py one of them, because langchain is to provide the agent standard interface, it is the api form of running, so this paper is also run api.py. as for the model, this paper with qlora fine-tuning and 4int quantization, has been quantized chatglm-6b-int4+qlora to save resources, although the model is compressed in advance of the loss of precision, but allegedly the performance impact is not large, and lora parameters previously mentioned the utilization rate is not high, so run by step directly select chatglm-6b-int4 can be followed in the qlora parameter tuning, the model introduces merge_lora_and_quantize module can be quantize module[42]. As shown in Figure 3:

```
embedding_model_dict = {
    "ernie-tiny": "nghuyong/ernie-3.0-nano-zh",
    "ernie-base": "nghuyong/ernie-3.0-base-zh",
    "ernie-medium": "nghuyong/ernie-3.0-medium-zh",
    "ernie-xbase": "nghuyong/ernie-3.0-xbase-zh",
    "text2vec-base": "GanymedeNil/text2vec-base-chinese",
    'simbert-base-chinese': 'WangZeJun/simbert-base-chinese',
    'paraphrase-multilingual-MiniLM-L12-v2': "sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2"
}


llm_model_dict = {
    chatglm": {
        "ChatGLM-6B": "THUDM/chatglm-6b",
        "ChatGLM-6B-int4": "THUDM/chatglm-6b-int4",
        "ChatGLM-6B-int8": "THUDM/chatglm-6b-int8",
        "ChatGLM-6b-int4-qe": "THUDM/chatglm-6b-int4-qe"
    ,
    belle": {
        "BELLE-LLaMA-Local": "/pretrainmodel/belle",
    ,
    vicuna": {
        "Vicuna-Local": "/pretrainmodel/vicuna",
```

Fig. 3 Selection of quantitative model

Experimental setup data:

- Top P parameters: text generation requires context sampling and then decoding, commonly used is top k, GPT using top p more emphasis on probability, in order to prevent the attention mechanism of the most recent word generation bias on the model calculation, the likelihood value is greater than P is the first few token probability and the largest but p, get the sum does not exceed the p inside the collection of tokens may be generated inside the candidate word, each time the answer According to the collection of words inside the proportion of the size of randomly generated text, the closer the value of 1 diversity, the more enough, this paper is a professional field of question and answer, the sample knowledge is less, you can sacrifice part of the diversity. Settings are generally 0.7, in this paper, to increase the accuracy to 0.5.
- Temperature: temperature sampling control text distribution model prediction: the output is too high may be biased, the value of 0-2, the default 0.1, are text control is generally not recommended and top-p change at the same time, it will lead to uncontrollable, and the top-p output quality has a higher impact, so only to change the value of small top-p.

qLora merge model after some of the parameters according to the example: batch_size (train and dev to maintain consistency, set a low value to ensure that the graphics card memory occupation is small): 4.

steps:500.

INITIAL_ LEARNING_ RATE: 1e-3.

To ensure speed, the epoch is not large: 3.

Optim: LAMB.

To make sure the model is quantitative, the parameter qbits :4 after merging.

Pip the qlora library again and merge it into the model store after tuning the parameters:

"Torchrun--nproc_per_node=2train_qlora.py--qlora/chatGLM_6B_QLoRA.json"

The inputs are trained to get the loss values, and then merge the merge_lora_and_quantize setup paths into chatglm to get the final training-tuned chatglm model.

5.2 Experimental results

5.2.1 Example data set

We obtain the knowledge base from Chinese Medical Knowledge Graph CMeKG, generate instructions to be classified by Darwin method, set the prompt word template prompt, fine-tune the input template to construct the Task: -Input: -Output: mode of Q&A instruction data, make it more robust, feed the input data to Darwin, generate Q&A and collect the data so as to get the The instruction dataset shown in the following example:

{"instruction": (in Chinese)A woman aged 70 has symptoms such as fainting, involuntary tremors, and emotional instability. Please provide a detailed explanation of her surgical treatment and preoperative preparation."input": "", "output":The disease requires the placement of electrode wires, pulse generators, and permanent pacemakers, as well as the use of sedatives and local anesthesia for surgical treatment. Preoperative preparation includes 1-3 days for preoperative examinations and lifestyle adjustments}；

{"instruction":(in Chinese)Is it necessary to undergo surgical treatment when a patient's ovarian small cell carcinoma metastasizes to other areas? "input": "", "output":When ovarian small cell carcinoma metastasizes to other areas, the effectiveness of surgical treatment may not be ideal, so surgical treatment is generally not recommended. For metastatic lesions, comprehensive treatment methods such as chemotherapy and radiotherapy can be used.}

5.2.2 Post-training results

Performance test:inference_test.py calculates the number of tokens per unit time.

python3 inference_test.py --model_path THUDM/chatglm-6b

The result is 19.62.

Gpu graphics memory only takes up 6-7GB.

Parameter results are shown in Figure 4:

For comparison, relevant Q&A on the web is borrowed to compare the Q&A of chatGPT and normal ChatGLM and fine-tuned DoctorGLM[43], after running on the platform with local address connection and public url, the local link is also to avoid the risk of information leakage. The data input is the json format dataset file fed directly to the Q&A system model training through the langchain interface, as shown in Figure 4. With DoctorGLM paper provides Q&A reference[44][45], in similar problems, this paper builds the model than gpt3 in more comprehensive and accurate, unlike gpt3 general "other reasons" and precautions,, and not like chatGLM condition is too perfunctory, the answer is specific and professional guidance, just in the The answer is too absolute and not as professional and clear as GLM. As a comparison, the relevant statements are labelled by human beings, and the green part is the possible wrong answer, while the blue part is the inadequacy of the text, which should be revised and evaluated, as shown in Table 3.

```
"epoch": 2.99,
"current_steps": 12358,
"loss": 0.5578,
"learning_rate": 4.0431266846361185e-06,
```

Fig. 4 Parameter results

Table 3 Response Comparison

Q :(in Chinese)In the past month or so, I have experienced discomfort and pain in my upper body, including my ribs, chest, sides, and back, as well as the upper part of my spine, including the fanbone. My upper body is very conflicted, as if I have a hand inside, feeling uncomfortable and uncomfortable. Sometimes my chest and ribs are uncomfortable, sometimes my back is uncomfortable and uncomfortable, and at the same time, I experience pain. Twist your body, do stretching activities, or use your fists to rub and massage areas that are uncomfortable or painful, and your body will stretch and feel better. In general, discomfort and dull pain occur in areas with ribs in the upper body and within the range of the upper vertebrae, while there is no pain in

areas without ribs, such as in the abdomen. I underwent cholecystectomy and common bile duct stone surgery in November 2011. It was not a laparoscopic procedure, but a large 18cm incision and two small drainage holes were made. This kind of discomfort and dull pain occurred once in the spring of 2012, and it automatically improved after a short period of time. But this time it was severe for a long time, and the discomfort and dull pain sometimes appeared in blocks and sometimes in strips. Generally, there were two or three areas of discomfort and dull pain at the same time, which was very uncomfortable, and there was a growing feeling.

(paper)(in Chinese)

(The pain you describe may be related to postoperative recovery, as well as poor posture, bone problems, or other muscle or ligament injuries. I suggest you consult an internist or thoracic surgeon for evaluation and accurate diagnosis.

During the waiting period for medical treatment, you can try the following measures to alleviate pain:

1. Maintain a good posture and avoid staying in the same position for extended periods of time.

Moderate exercise and avoid prolonged inactivity.

3. Slowly engage in gentle stretching exercises, such as rotating shoulders, bending and extending necks, etc.

Use hot or cold compress to relieve pain.

Try massaging the painful area or seek the help of a professional massage therapist to relieve muscle tension or ligament pain.

Please remember that these are only temporary relief methods. If the pain persists or worsens, it is still recommended to seek medical attention as soon as possible. Only doctors can provide a detailed assessment of your symptoms and provide accurate treatment recommendations.)

(DoctorGLM)(in Chinese)

(Disease analysis: Hello, this situation is likely caused by pain from intercostal neuritis.

Guidance: It is recommended to take oral anti-inflammatory drugs, painkillers, and vitamin B supplements according to the course of treatment. After the pain is relieved, practicing chest lifting exercises can alleviate the pain. In addition to regular treatment for gallstones, patients need to consult with experts and communicate with doctors. Patients also need to pay attention to nursing aspects, such as reasonable diet and maintaining a happy mood. At the same time, patients should also choose a professional hospital for diagnosis and treatment, Only in this way can good therapeutic effects be achieved.)

(ChatGPT) (in Chinese)

(I'm sorry to hear that you have been experiencing these discomforts and dull pain recently. According to your description, your symptoms seem to involve skeletal structures such as ribs and vertebrae in multiple parts of the upper body, as well as the accompanying discomfort and dull pain. Due to your previous experience with gallbladder removal and common bile duct stone surgery, these symptoms may be related to the surgery itself, but they may also be caused by other reasons. Therefore, it is recommended that you seek medical attention in a timely manner and undergo detailed physical and related examinations to determine the specific causes and treatment methods of the symptoms. You can consult with your family doctor or internist, who can conduct preliminary examinations and evaluations, and perform corresponding examinations as needed, such as X-rays, CT, MRI, or blood tests. While waiting for medical treatment, you can pay attention to rest, maintain good posture, avoid excessive exercise or lifting heavy objects, and use methods such as hot or cold compress to relieve pain and discomfort. Finally, it is recommended that you maintain an optimistic attitude and actively cooperate with doctors for treatment and rehabilitation. Wishing you a speedy recovery!)

(ChatGLM) (in Chinese)

(The symptoms described may be caused by various factors, including physical, psychological, and emotional aspects. Based on the symptoms described, it is recommended to perform the following examinations and tests:

1. Doctor's physical examination: Conduct a comprehensive physical examination, including heart and lung auscultation, blood pressure testing, vision, hearing, skin testing, etc., to determine

any potential health issues.
2. Electrocardiogram (EKG): Test heart health to assess potential heart problems.
3. Magnetic Resonance Imaging (MRI): tests the health of the skeletal and muscular systems, and evaluates for potential neurological issues.
4. Blood test: Test for anemia, infection, tumors, and other issues.
5. Psychological health testing: testing emotional and mental health, such as stress testing, anxiety testing, etc.
If symptoms persist or become more severe, please consult a doctor as soon as possible for professional advice and treatment.)

Enhance the dataset with CoT and self consistency, and enrich the text data with langchain.

The CoT fine-tuning of the Prompt template feed shot adapts to the input

In addition, the similarity ROUGE-L (0.6342) between the long text and the reference data of the validation set was calculated based on the recall rate, and the effect of this article was validated using the unmodified chatglm as shown in Table 4:

Table 4 Comparison of similarity with chatglm

	ROUGE-L
module	0.7942
chatglm	0.5469

5.2.3Project operating environment

The project is based on the open-source webui project, the quantitative model is introduced, in which it is merged with chatGLM to form a new quantitative language model, to get a lower configuration can run the Q&A system, and the instruction dataset is processed individually locally, and the data is added to the model to get the results shown above after the project is run. However, the minimum requirement of 10GB of graphics memory for the instruction dataset construction is higher than the local demand, based on the low-cost requirements, so this paper obtains the dataset through the A100 cloud server in the QI platform, and runs the project locally with 8GB of graphics memory on a 4070 graphics card with low graphics memory.

Running environment: This paper only runs in 4070 graphics card with low graphics memory of 8GB graphics memory due to the quantitative model of qlora merge4int as shown above.

6. Conclusion

This paper aim to help solve the difficult about Medical problems and to achieve a slight decoupling and through the comparison, the model is able to generate the answers in a more comprehensive way with slight deployment.

It aims to study a medical LLM model that assists patients in consulting medical information through question answering, providing reference opinions, and attempting to use QLORA fine-tuning to achieve localized deployment in low graphics memory, using Langchain to process input to achieve platform interaction question answering interface, using self instruction method on the dataset. Due to the high cost of generating data from the ChatGPT closed source model, we refer to the Dawen model and use open-source LLM to leverage the knowledge graph's comprehensiveness, Construct and train instruction datasets using professional prompt and scientific instruction generation (SIG) model method for knowledge generation.Optimizing the prompt template through cots to obtain optimal Q&A results, and in order to filter the optimal answers, self instrument generates data by using multiple cots to filter for the best quality data.Utilizing Langchain to generate high-quality datasets of professional documents enhances usability.Select chatglm with good performance and lightweight data for data fine-tuning training on the training model. The experiment requires the use of high-quality models to generate data, with at least 10GB of graphics memory higher than the local configuration. The data generation module is implemented at a low cost on the Qizhi Cloud platform, while the chatglm+langchain deployment can be deployed locally for quantization friendly operation. Although the quality of the generated data in the experiment is only close to gpt3.5, considering the ability to localize the deployment model, we attempt to lay a foundation for the widespread research on nlp with low threshold, At the same time, explore a medical Q&A path with certain reference value.

Last,Attempt to improve convergence and optimize answer retrieval to increase accuracy and speed for deployment.

Shortcomings: langchain after processing the input data although improved fault tolerance, outside the dataset of professional Q&A error is less, but there are still wrong answers or a certain probability of outputting abnormal answers, after all, it is in the low-platform Q&A system, chatglm2-6b second generation version of the release of the perhaps can be resolved, but the graphics memory consumption may be still chatglm-6b more friendly, and the Q&A system can only be equivalent to chatGPT3.5 with a certain degree of accuracy, but generative text still can not be highly accurate, after all, qlora use graphics memory although small, and therefore become less. And the Q&A system can only be comparable to chatGPT3.5, there is a certain degree of accuracy,the cot actually improve the ability of module, but medical generative text requires highly accurate,it can be referenced but not completely dependent. After all, qlora use graphics memory although small, the computing power has thus become lower, and chatglm-6b model itself there is a gap in the answer is slow and the graphics memory is almost full after the run, a few rounds of dialogue will be a few rounds of graphics memory explosion, paper try to reduce the difficulty, maybe use docker can help our private in a short time.

References

[1] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, “K-BERT: enabling language representation with knowledge graph,” in The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 2020, pp. 2901–2908. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5681.

[2] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, “Harnessing the power of llms in practice: A survey on chatgpt and beyond,” arXiv preprint arXiv:2304.13712, 2023.

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and Alex Ray. Training language models to follow

instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744,2022.
[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer.QLORA: Efficient Finetuning of Quantized LLMs.
[5] Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns, 3(4), 2022.
[6] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fifine-tuned on llama model using medical domain knowledge. arXiv preprint arXiv:2303.14070, 2023.
[7] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Model with Self Generated Instructions. arXiv preprint arXiv:2212.10560, 2022.
[8] S. An, Y. Li, Z. Lin, Q. Liu, B. Chen, Q. Fu, W. Chen, N. Zheng, and J.-G. Lou. Input-tuning: Adapting unfamiliar inputs to frozen pretrained models. arXiv preprint arXiv:2203.03131, 2022.
[9] LAION. Open-instruction-generalist dataset.https://github.com/LAION-AI/Open-Instruction-Generalist, 2023.
[10] Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, Imran Razzak, and Bram Hoex.DARWIN Series: Domain Specific Large Language Models for Natural Science. arXiv preprint arXiv.2308.13565, 2023.
[11] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.
[12] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, and Kranthi Kiran GV. RWKV: Reinventing RNNs for the Transformer Era. arXiv preprint arXiv:2305.13048, 2023.
[13] Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns, 3(4):100488, 2022.
[14] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," in NeurIPS 2022 Foundation Models for Decision Making Workshop.
[15] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691, 2021.
[16] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.
[17] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infifilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, 2022. 4.
[18] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang

Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pretrained model. In The Eleventh International Conference on Learning Representations (ICLR).
[19] Florin Eggmann, Roland Weiger, Nicola U Zitzmann, and Markus B Blatz. Implications of large language models such as chatgpt for dental medicine. Journal of Esthetic and Restorative Dentistry, 2023.
[20] T. Chen, B. Xu, C. Zhang, and C. Guestrin. Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174, 2016.
[21] S. H. Bach, V. Sanh, Z.-X. Yong, A. Webson, C. Raffel, N. V. Nayak, A. Sharma, T. Kim, M. S. Bari, T. Fevry, et al. Promptsource: An integrated development environment and repository for natural language prompts. arXiv preprint arXiv:2202.01279, 2022.
[22] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975, 2023.
[23] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. Chataug: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007, 2023.
[24] Hongying Feng, Samir H Patel, William W Wong, James E Younkin, Gregory P Penoncello, Danairis Hernandez Morales, Joshua B Stoker, Daniel G Robertson, Mirek Fatyga, Martin Bues, et al. Gpu-accelerated monte carlo-based online adaptive proton therapy: A feasibility study. Medical physics, 49(6):3550–3563, 2022.
[25] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416, 2022.
[26] T. Dettmers and L. Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. arXiv preprint arXiv:2212.09720, 2022.
[27] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.
[28] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. arXiv preprint arXiv:2301.13688, 2023.
[29] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
[30] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, et al. Differentiate chatgpt-generated and human-written medical texts. arXiv preprint arXiv:2304.11567, 2023.
[31] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454, 2023.
[32] Stanford CRFM – crfm.stanford.edu. https://crfm.stanford.edu/2023/03/13/alpaca.html. [Accessed 09-Jun-2023].
[33] Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM— databricks.com.https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm. [Accessed 09-Jun- 2023].

[34] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

[35] Zachary Huemann, Changhee Lee,Junjie Hu,Tyler Bradshaw,Steve Y. Cho.Domain-adapted large language models for classifying nuclear medicine reports: arXiv preprint arXiv2303.01258,2023.

[36] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specifific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1):1–23, 2021. 3.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[38] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Advances in Neural Information Processing Systems, 35:1950–1965, 2022.

[39] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032, 2023.

[40] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020.

[41] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323, 2022.

[42] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer. 8-bit optimizers via block-wise quantization. 9th International Conference on Learning Representations, ICLR, 2022.

[43] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, Dinggang Shen.DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task. arXiv preprint arXiv:2304.01097, 2023.

[44] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110, 2022.

[45] T. Liao, R. Taori, I. D. Raji, and L. Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.