NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

From Symptoms to Services: An LLM Chatbot for Effective Departmental Referral

Qi Shi

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

From Symptoms to Services: An LLM Chatbot for Effective Departmental Referral

by

Qi Shi

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science.

Supervised by

Miguel de Castro Neto, PhD, NOVA IMS

Qiwei Han, PhD, NOVA SBE

November, 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 01/11/2023]

ABSTRACT

This study explores integrating large language models (LLM) into the medical domain, focusing on developing and using the LLM tool Chat-SymChecker. Although LLM technology, such as ChatGPT and Copilot, is evolving rapidly, their use in medical consultations is still limited due to their complexity. To address this issue, we propose using Chat-Symptom Checker to supplement primary care consultation and specialist referral. Chat-Symptom Checker is based on the LLaMA model and is trained on extensive medical Question-answer datasets and patient-specific data from electronic health records, allowing it to provide rapid initial assessment and efficiently direct patients to the proper medical department. This article describes Chat-Symptom Checker's development process, functionality, and potential impact in increasing hospital efficiency, accelerating diagnostic procedures, and enhancing patient care. Chat-Symptom Checker shows the capability of processing complex natural language input, allowing users to describe symptoms and receive clear, individualized feedback. By integrating comprehensive patient data, such as past medical history and family history, the system will guide users to the proper medical department and provide initial recommendations and potential diagnoses, which can significantly decrease wait times and labor costs while simultaneously improving service efficiency. However, there are several challenges with our model. Issues such as redundant or nonsensical queries still need to be refined. In addition, an evaluation of the text quality of LLMs reveals that data volume is not necessarily correlated with enhanced performance. Studies reveal that smaller datasets with better text quality can perform better than more enormous datasets that lack context coherence. This shows the significance of data quality and contextual relevance in LLM medical model training. Despite some remaining limitations, Chat-Symptom Checker can serve as a beneficial healthcare support tool.

KEYWORDS

Large Language Models (LLMs); Symptom Checker; Medical Diagnosis Support; Artificial Intelligence

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

LLM	Large Language Model
EHR	Electronic Health Records
EMR	Electronic Medical Records
CMeKG	Chinese Medical Knowledge Graph
LLaMA	Large Language Model Meta AI
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformers
API	Application Programming Interface
ICL	In Context Learning
PEFT	Parameter-Efficient Finetuning
LoRA	Low-Rank Adaptation
RLHF	reinforcement learning from human feedback

1. INTRODUCTION

Because of the rapid growth of Large Language Model (LLM) technology, chatbot-based apps have become popular in the app business. LLMs such as ChatGPT (OpenAI, 2022) are highly successful in understanding instructions and generating human-like dialogues. A tremendous amount of training data gives them verbal comprehension and flexibility across several languages and domains. ChatGPT-based programs, such as AutoGPT(Significant-Gravitas, 2023), Copilot(GitHub, 2023), etc., are well-known. However, applying LLM models in the medical field remains relatively limited, especially in medical consultation scenarios. The complexity of medical consultations frequently exceeds the capabilities of general-purpose LLMs because general-purpose LLMs regularly provide diagnosis and treatment advice based on a single-round conversation. In the context of a medical consultation, a doctor must participate in multiple rounds of talks with patients to get enough information, respond appropriately at each round, ensure correct information is obtained, and finally give the proper medical advice. Therefore, one-round-based medical advice may result in potential inaccuracies and misleading information, which limits the application of LLMs in medical contexts.

To improve accuracy and practicality, we designed LLM as a tool to assist doctors rather than completely replace them in tasks such as providing medical advice. The whole Medical procedure can be systematically divided, and LLMs are trained specifically with different medical domain knowledge for each segment. The approach enables LLMs to focus on specific data in each model segment and decreases the resources and time needed for training. With reduced complexity, LLMs could also present the possibility of better performance. In our project, we've divided the medical procedure into stages: primary consultation, referral, specialist consultation, diagnosis, and treatment. We propose an LLM solution for primary consultation and specialist referral, namely the Chat-Symptom Checker.

Typically, the standard procedure for patients is to visit the primary doctor in a hospital and wait until they are referred to a proper specialist. This is not only time-consuming but also exhausts medical resources. Thus, the Chat-Symptom Checker is provided to solve this problem, which can speed up the primary consultation process for patients while simultaneously offering recommendations for the proper specialist. There are several features of our Symptom Checker LLM.

Swift Preliminary Assessments: Symptom Checker LLM provides a quick preliminary assessment. Patients answer a series of questions about their symptoms, following which the checker will provide initial recommendations. It helps gather a patient's condition more quickly, speeding up the entire diagnostic process.

Enhancing Hospital Efficiency: One of its main goals is to lead patients to the correct medical department based on their symptoms. Patients are often unclear about where to go for medical care. In such circumstances, the symptom checker provides ideas to guide them in

visiting a specific department. This decreases unnecessary trips to primary care and ensures they receive treatment rapidly.

Data Logging and Record Keeping: Aside from direct patient interactions, the symptom checker is excellent at recording dialogues. It generates log files on its own, which is useful to help medical specialists review the patient's condition. Besides this, these records can also be used to create a summary for Electronic Health Records (EHR) or Electronic Medical Records (EMR), improving overall patient care efficiency.

Our model is developed using LLaMA as a base model. The training process integrates a large amount of medical QA datasets and patient-specific data from electronic health records (EHR). This integration is essential because it allows LLMs to have adequate medical knowledge and improves their suggestions' potential accuracy and relevancy depending on the patient's health history. However, it is crucial to note that symptom checks are not a substitute for medical decisions. An experienced physician must make the final diagnostic and treatment decisions, for medical diagnosis decisions demand extreme precision. Due to the limited logic capabilities of LLMs, Chat-SymCheker serves as a supplement for initial assessment and referral. Nonetheless, its primary purpose is to guide patients to the appropriate treatment department, which can enhance hospital diagnosis, treatment, and the patient's medical experience. Consequently, LLMS still has broad application prospects in these fields.

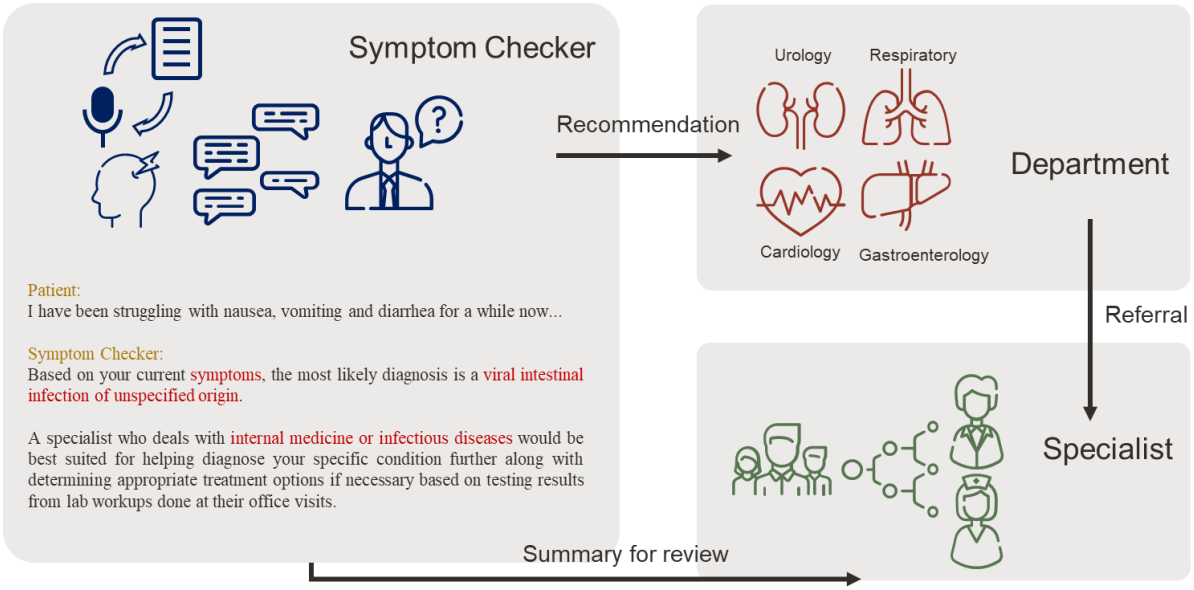


Figure 1.1 Symptom checker and referral recommendation system

2. LITERATURE REVIEW

Symptom checkers aim to improve the precision and effectiveness of medical condition diagnosis, and their development and use have drawn more attention in recent years. A symptom checker can monitor symptoms in various medical circumstances and provide rapid reports, especially during the pandemic of COVID-19 these years, where they can play a crucial role in involving public safety and COVID-19 monitoring.

Research (Zobel, Knapp, Nateqi, & Martin, 2023) shows that online symptom checkers, such as Symptoma, can effectively track national trends in coronavirus infections, providing a valuable, non-traditional source of information for epidemic policymaking. In addition, these symptom detectors frequently detect trends before official infection numbers are released, which could provide early warning signals.

Regardless of user demographics, another study (A. W. Liu et al., 2022) reveals that COVID-19 symptom monitors with self-triage and self-scheduling capabilities tend to have high user satisfaction. By empowering users to self-triage and schedule tests or consultations, such tools can reduce the number of unnecessary contacts and messages to physicians.

In a study by Simpson et al. (Simpson et al., 2022), contact tracing techniques were deployed within 24 hours for suspected cases of Monkeypox. Within four days of a confirmed diagnosis, the system is operational, using clinical follow-up with healthcare providers and permitting real-time updates for enhanced monitoring and tracking. The study highlighted the rapid development and implementation of a mobile response investigation solution for potential exposure situations, exposure risk assessment, and symptom monitoring for Monkeypox, demonstrating the efficacy of symptom checkers in tracking and addressing contagious diseases.

Because Symptom Checker is a tool that uses computer algorithms and machine learning to aid self-diagnosis, higher accuracy rates indicate more benefits for patients. As computer technology develops, increasing accuracy is also of concern. A study by Peven et al. (Peven et al., 2023) evaluated the accuracy of three symptom checkers developed by Flo Health for diagnosing endometriosis, uterine fibroids, and polycystic ovary syndrome. The research discovered that symptom checkers have high accuracy for these illnesses, implying that they have the potential to reduce the time it takes to detect reproductive health diseases, which proves that digital tools can enhance healthcare outcomes in the reproductive field.

Maturana et al. (Maturana et al., 2022) conducted a comprehensive review of the use of artificial intelligence tools in malaria diagnosis, particularly in contexts with limited resources. The authors demonstrate the potential for artificial intelligence-based systems to provide cost-effective malaria diagnosis, indicating that incorporating artificial intelligence into microscopy imaging could facilitate precise diagnosis, allowing for faster therapeutic intervention and reducing the prevalence of this epidemic disease.

A study (Hennemann, Kuhn, Witthöft, & Jungmann, 2022) tested the diagnostic capabilities of a widely available symptom checker against the formal diagnosis of mental illness. Compared to interviews with therapists, the symptom checker ADA showed excellent performance, with an accuracy rate of 69%. Markus Gräf et al. (Gräf et al., 2022) compared the diagnostic accuracy between the symptom checker Ada and medical professionals in inflammatory rheumatic disorders. Ada showed higher accuracy than doctors. These studies demonstrate the benefits of symptom checkers as a supplemental tool for healthcare and show the potential to enhance diagnostic results.

Interestingly, despite sometimes beating doctors in accuracy, the symptom checker completely fails at emergency diagnosis, which is more crucial to patients. In the study by Laure Abensur Vuillaume et al. (Abensur Vuillaume et al., 2023), Objective Structured Clinical Examinations were used to compare the performance of the symptom checker and emergency physicians. It was discovered that emergency physicians performed better than the symptom checker on both primary diagnosis and correlation of primary and secondary diagnoses. At this point, the symptom checker's accuracy still limits its usage in hospitals.

Since disease diagnosis and treatment cannot be performed, symptom checkers should be used more on the service side. Mohammed1 et al. (Mohammed, Mohammed, Mohammed, & Applications, 2022) researched how automation in digital medical systems is affected by artificial intelligence. The results show that AI technology can improve healthcare by automating processes and enhancing patient outcomes. However, in terms of service, the symptom checker still leaves space to improve. According to a study by Yue You et al. (You & Gui, 2020), some of the services provided by current symptom checkers include establishing a user's health history, inputting symptoms, presenting probing questions, and supporting a variety of diseases and user groups but still do not perform completely. The study's findings also show that consumers are worried about the conversational design of CSC applications' input restrictions and language understandability.

However, with the rapid development of LLM technology, large language models have raised the possibility of even understanding users' natural language. Currently, several LLM models based on medical information have been released. Besides the non-open source OpenAI ChatGPT(OpenAI, 2022) and Google's PaLM2(Google, 2023), there are also some open-source models. HuaTuo(H. Wang et al., 2023), an open-source Chinese biomedical LLM finetuned based on Large Language Model Meta-artificial Intelligence (LLaMA) with knowledge-guided data. The data includes structured and unstructured medical knowledge from the Chinese Medical Knowledge Graph (CMeKG), providing the model domain-specific knowledge with accuracy. Another open-source model, ChatDoctor(Li et al., 2023), uses medical domain knowledge and an online doctor-patient conversation database to finetune a medical chat model on LLaMA.

However, even with LLMs, effective deployment still presents several unresolved problems and difficulties, including acquiring adequate medical domain knowledge, considering medical

history information, conversation recording, user interface design, etc. Given the challenges above, the primary objective of this project is to investigate how to apply conversational artificial intelligence in the context of hospital networks. We are trying to offer virtual consultations, address patients' questions, and lead them to the proper department for needed medical care. The model focuses not on diagnosis and treatment but on enhancing patient-centered medical experience and efficiency.

3. METHODOLOGY

3.1. MODEL

3.1.1. Base Model

LLaMA: LLaMA is a collection of pre-trained generating text models with different parameters ranging from 7 billion to 70 billion (Touvron et al., 2023), and their source codes are available for research purposes. There are already numerous fine-tuned models on the market that perform excellent outcomes. For example, Vicuna-13B is a model fine-tuned based on LLaMA-13B, which can achieve equivalent performance to GPT-4. It is regarded as a free alternative to GPT-4(Thissen, 2023). According to LMFlow's research, the fully fine-tuned LLaMA-7B model performs even better than ChatGPT in the pubmedQA evaluation (Diao et al., 2023). This demonstrates that even a model with fewer parameters may perform better after fine-tuning. When fine-tuned in specific fields such as medicine, the outstanding results even outperform those of a general broad language model. Because of this, We decided to use the LLaMA-7B model in our project. It has the potential to achieve excellent performance in professional domains and requires less training time and resources. Additionally, the smaller size makes the implementation in potential future deployment simpler.

3.1.2. Evaluation model

BERT: BERT (Bidirectional Encoder Representations from Transformers)(Devlin, Chang, Lee, & Toutanova, 2018) is an NLP model based on the Transformer architecture, similar to LLM. Using a self-attention mechanism, this architecture captures long-term dependencies in text. It captures pre- and post-contextual information. Using large-scale text data for training, they can learn and comprehend text data's complex patterns and structures.

ClinicalBERT: Since our project is about the healthcare system, we used the ClinicalBERT model (G. Wang et al., 2023). ClinicalBERT is a medically-specific language model. Its training is based on a large-scale multi-center data set, containing a multi-disease corpus of 1.2 billion words and then fine-tuned with a large corpus of EHRs (electronic health records) containing over 3 million patient records.

3.2. DATA

Due to the special characteristics of the medical area, LLM models must have a wide range of knowledge to perform at their best. For instance, ChatGPT is trained on medical professional data, which the open-source, pre-trained LLaMA model lacks. Therefore, to realize the scenario where patients seek health and medical advice, we intend to train medical knowledge in LLAMA in two steps:

3.2.1. Medical Databases

PubmedQA: A question-and-answer dataset for the biomedical area based on the PubMed database. It includes 1,000 expert annotations QA instances, 612,000 unlabeled QA instances, and 211.3 thousand artificially generated QA instances. (Q. Jin, Dhingra, Liu, Cohen, & Lu, 2019)

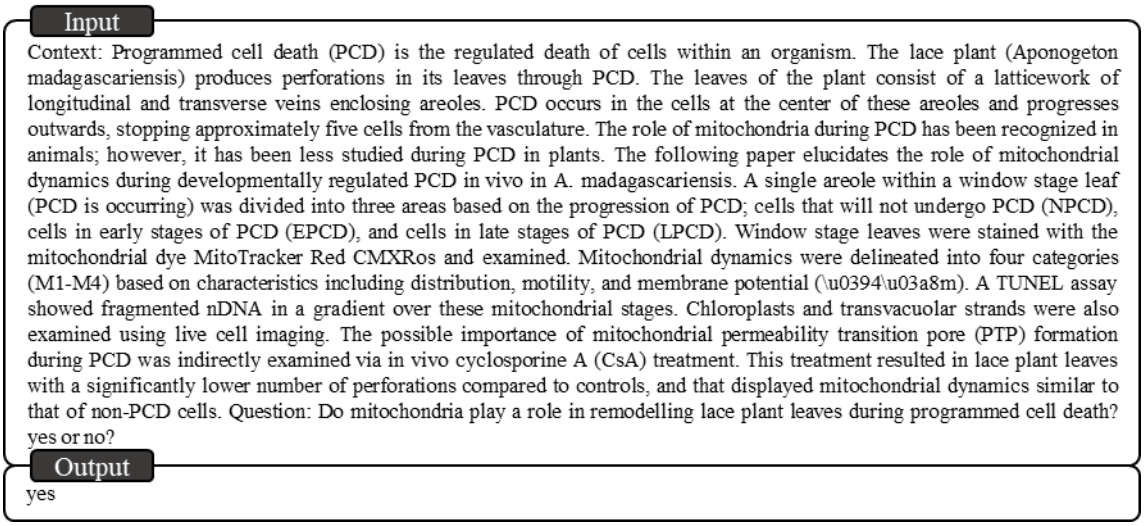


Figure 3.1 An example of the PubMedQA dataset

MedQA-USMLE: A question-and-answer dataset aimed exclusively at US medical students and physicians in preparation for the US Medical Licensing Examination, which contains 12,723 questions. (D. Jin et al., 2021)

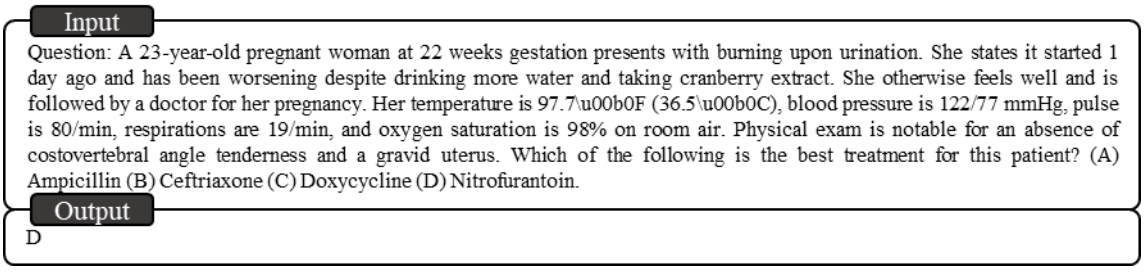


Figure 3.2 An example of the MedQA-USMLE dataset

MedMCQA: A vast multi-choice question-answer (MCQA) dataset containing medical entrance exam questions. It collects approximately 194,000 records, covering more than 2,400 healthcare topics and 21 medical subjects. (Pal, Umapathi, & Sankarasubbu, 2022)

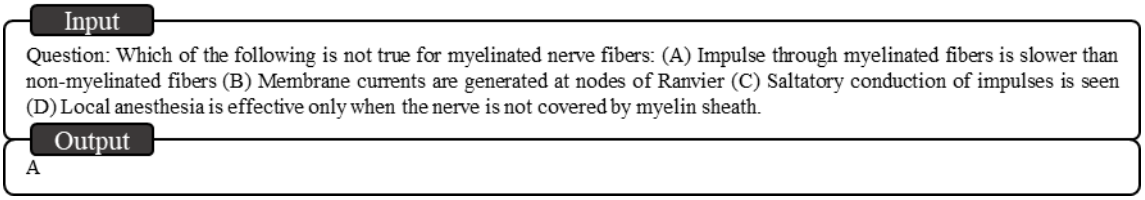
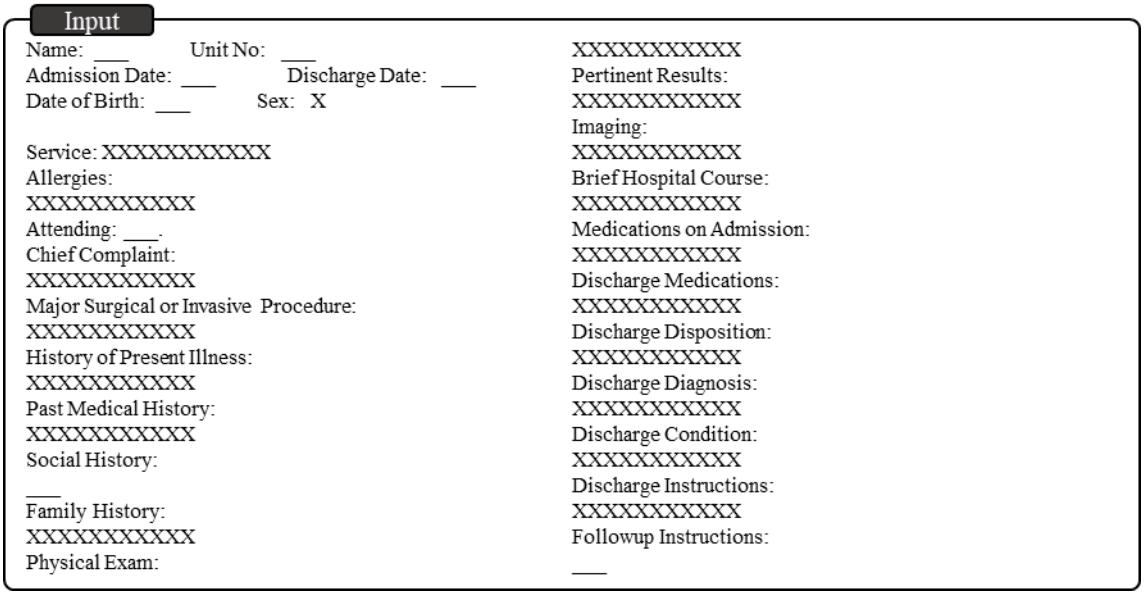


Figure 3.3 An example of the MedMCQA dataset

MIMIC-IV note: This clinical database contains de-identified free-text clinical notes from patients. Specifically, MIMIC-IV-Note has unidentifiable discharge summaries from 145,915

patients treated in Boston's Beth Israel Deaconess Medical Center's hospital and emergency departments. (Goldberger et al., 2000; Johnson, Pollard, Horng, Celi, & Mark, 2023)


Input

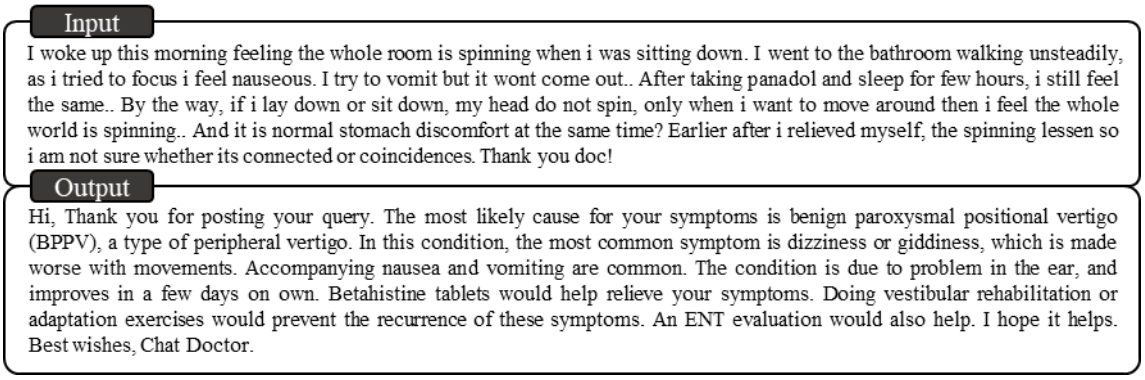
Name: ___ Unit No: ___
Admission Date: ___ Discharge Date: ___
Date of Birth: ___ Sex: X

Service: XXXXXXXXXXXX
Allergies:
XXXXXXXXXXXX
Attending: ___.
Chief Complaint:
XXXXXXXXXXXX
Major Surgical or Invasive Procedure:
XXXXXXXXXXXX
History of Present Illness:
XXXXXXXXXXXX
Past Medical History:
XXXXXXXXXXXX
Social History:
___
Family History:
XXXXXXXXXXXX
Physical Exam:
XXXXXXXXXXXX
Pertinent Results:
XXXXXXXXXXXX
Imaging:
XXXXXXXXXXXX
Brief Hospital Course:
XXXXXXXXXXXX
Medications on Admission:
XXXXXXXXXXXX
Discharge Medications:
XXXXXXXXXXXX
Discharge Disposition:
XXXXXXXXXXXX
Discharge Diagnosis:
XXXXXXXXXXXX
Discharge Condition:
XXXXXXXXXXXX
Discharge Instructions:
XXXXXXXXXXXX
Followup Instructions:
___

Figure 3.4 An example of the MIMIC-IV note text

3.2.2. Medical Dialogue Datasets

MedDialog: MedDialog (Zeng et al., 2020) is a large set of medical conversation datasets with two languages. The Chinese dataset covers 172 professional diseases with 3.4 million patient-doctor conversations, including 11.3 million utterances and 660.2 million tokens. The English dataset covers 96 specialties of disease with 260 thousand dialogues, including 510 thousand utterances and 44.53 million tokens. It is reported that MedDialog is the world's largest medical conversation dataset. The authors demonstrate that a clinical, doctor-like medical dialogue model could be generated by training on the Chinese dialogue dataset. Furthermore, their studies also show that the performance of medical dialogue generation tasks can be significantly improved by fine-tuning with MedDialog.


Input

I woke up this morning feeling the whole room is spinning when i was sitting down. I went to the bathroom walking unsteadily, as i tried to focus i feel nauseous. I try to vomit but it wont come out.. After taking panadol and sleep for few hours, i still feel the same.. By the way, if i lay down or sit down, my head do not spin, only when i want to move around then i feel the whole world is spinning.. And it is normal stomach discomfort at the same time? Earlier after i relieved myself, the spinning lessen so i am not sure whether its connected or coincidences. Thank you doc!

Output

Hi, Thank you for posting your query. The most likely cause for your symptoms is benign paroxysmal positional vertigo (BPPV), a type of peripheral vertigo. In this condition, the most common symptom is dizziness or giddiness, which is made worse with movements. Accompanying nausea and vomiting are common. The condition is due to problem in the ear, and improves in a few days on own. Betahistine tablets would help relieve your symptoms. Doing vestibular rehabilitation or adaptation exercises would prevent the recurrence of these symptoms. An ENT evaluation would also help. I hope it helps. Best wishes, Chat Doctor.

Figure 3.5 An example of the MedDialog

DDXPlus


```
{
  "name": "E_130",
  "code_question": "E_129",
  "question_fr": "De quelle couleur sont les lésions?",
  "question_en": "What color is the rash?",
  "is_antecedent": false,
  "default_value": "V_11",
  "value_meaning": {
    "V_11": {"fr": "NA", "en": "NA"},
    "V_86": {"fr": "foncée", "en": "dark"},
    "V_107": {"fr": "jaune", "en": "yellow"},
    "V_138": {"fr": "pâle", "en": "pale"},
    "V_156": {"fr": "rose", "en": "pink"},
    "V_157": {"fr": "rouge", "en": "red"}
  },
  "possible-values": [
    "V_11",
    "V_86",
    "V_107",
    "V_138",
    "V_156",
    "V_157"
  ],
  "data_type": "C"
}
```

```
{
  "condition_name": "Myasthenia gravis",
  "cond-name-fr": "Myasthénie grave",
  "cond-name-eng": "Myasthenia gravis",
  "icd10-id": "G70.0",
  "symptoms": {
    "E_65": {},
    "E_63": {},
    "E_52": {},
    "E_172": {},
    "E_84": {},
    "E_66": {},
    "E_90": {},
    "E_38": {},
    "E_176": {}
  },
  "antecedents": {
    "E_28": {},
    "E_204": {}
  },
  "severity": 3
}
```


Figure 3.6 An example of evidence and pathology description in the DDXplus

DDXPlus: DDXPlus(Fansi Tchango, Goel, Wen, Martel, & Ghosn, 2022) provides a large-scale, exhaustive dataset of approximately 1,300,000 patients, which includes differential diagnoses as well as the true pathology, symptoms, and history information of each patient. In contrast to existing datasets that only contain binary symptoms and antecedents, DDXPlus also includes categorical and multi-select symptoms and antecedents. In addition, the hierarchical organization of symptoms makes it possible to design systems that can interact with patients in a logical manner.

E_183
Before:
Do you live in a rural area?

After:
1. Is your residency situated in a rural area?
2. Is your place of residence located in a rural area?
3. Is your home in a countryside setting?
4. Do you live in a rural area?

Before:
'iliac wing(R)':

After:
- "Right side of my lower back"
- "Near my right hip"
- "Right side above the hip"
- "Right side of my waist"
- "On my right, where my waistline is"

E_59
Before:
How fast did the pain appear?

After:
How fast did the pain appear? You can choose severity or score as following rules: No(0); Slow(1-3): more than 1 day; Moderate(4-6): less than 1 day; Fast(7-9): within several miniutes; Immediately(10)

Figure 3.7 An example shows how to rewrite and optimize term descriptions and query questions.

The DDXPlus dataset is the most important dataset for our project. Although it has no dialogue data, each patient case includes relative symptoms, including initial symptoms, and provides

differential diagnostic possibilities that do not exist in other datasets. An artificial conversation dataset between potential patients and symptom checkers, Chat-DDXPlus, was generated with these comprehensive symptoms.

Table 3.1 Datasets for evaluation

Dataset	Description
data_5k	5k multi-round conversation samples were totally rewritten by ChatGPT 3.5.
data_8k_without_no	8k multi-round conversation samples that do not include the only initial symptom conversation samples. And without any negative answered symptoms and antecedents added.
data_9k_without_no	8k multi-round conversation samples plus 1k only initial symptom conversation samples. And without any negative answered symptoms and antecedents added.
data_11k_without_no	8k multi-round conversation samples plus 1k only initial symptom conversation samples, then plus 2k one-round conversation samples. And without any negative answered symptoms and antecedents added.
data_8k	8k multi-round conversation samples that do not include the only initial symptom conversation samples.
data_9k	8k multi-round conversation samples plus 1k only initial symptom conversation samples.
data_11k	8k multi-round conversation samples plus 1k only initial symptom conversation samples, then plus 2k one-round conversation samples.

The basic symptom inquiry questions of DDXPlus contain medical terms that patients may find difficult to understand. To promote more patient-friendly and humanized conversation capabilities. To make it easier to understand and have some diversity, we rewrite and optimize each description of symptoms and their inquiry questions and put them in a text library (Figure 3.7). When constructing artificial conversation samples, all questions and answers are selected from the rewritten text library. As shown in the DDXPlus dataset(Figure 3.6), each piece of evidence only has positive symptoms and antecedents, so we add 5-10 most common symptoms and antecedents in the dataset with negative answers in each conversation, for those patients who consult with symptom checker may have only one main symptom. So, we created 10% of conversation samples with only one initial symptom. Moreover, In real medical conversation scenarios, it is possible that all symptoms can be described in one turn, so we compress 20% of the samples into a one-round conversation. Finally, to save training resources and time, we systematically sampled 110k samples based on the evidence for the final training dataset.

Furthermore, datasets containing different text structures were generated to evaluate data quality. A smaller 11k subset was systematically sampled from the 110k dataset based on the evidence. Then, all datasets were sampled from the 11k subset to maintain consistency(Table 3.1).

3.3. TRAINING FRAMEWORK

3.3.1. Alpaca

Alpaca(Taori et al., 2023; tatsu-lab, 2023) should have been an early and well-known model. As no open-source model is equivalent to OpenAI, the researchers released an Alpaca model based on Meta's LLaMA-7B model. The training methods and data have also been released to the public by the researchers. Currently, many models like ChatDoctorField (Li et al., 2023) train their model using Alpaca code since it provides a lightweight model training framework. Through the self-instruct training method, Alpaca performs similarly to text-DaVinci-003, but with a smaller size and easy to replicate. However, Alpaca demonstrates several common language model faults, such as the generation of disinformation, poisonous words, and stereotypes, for it is trained based on LLaMA without any commercial permit and lacks proper security features, which are exclusively for academic study only.

3.3.2. LMFlow

Most General Language Models have significant limitations in specific field tasks, which require additional fine-tuning in these domains to reach optimal performance, particularly in the medical industry. Therefore, we import the LMFlow (Diao et al., 2023; OptimalScale, 2023) framework in our project, simplifying fine-tuning and inference of foundation LLMs. It supports continuous pre-training, instruction tuning, efficient parameter fine-tuning, alignment tuning, large model inference, and a well-designed and extendable API for LLMs. It is worth noting that it has both task tuning and instruction tuning methods, which are really critical in our project. Task Tuning enhances LLM expertise in certain fields, such as mathematics or medicine. Language instruction tuning is a strategy for improving a language model's performance by training it to follow natural language instructions or commands.

3.3.3. FastChat

FastChat(lm-sys, 2023; Zheng et al., 2023) is an open-source system for training, evaluating, and serving as chatbots based on LLMs. The platform has complete training and evaluation algorithms based on multiple fundamental models, such as the LLaMA model, a Web-based user interface, and a distributed multi-model service system compatible with the OpenAI API, through which Vicuna-13B was trained and released. FastChat is quite valuable for our project because it provides a training framework for multi-round conversation data structures, which is not typically easy to find in other training frameworks.

3.4. TRAINING METHODOLOGY

3.4.1. Fine-tuning

Fine-tuning(Zhang, Lipton, Li, & Smola, 2021) is a transfer learning strategy in machine learning. If a model does not have enough training data, its accuracy typically does not meet the requirements. However, gathering and labeling data may require time, money, manpower, and material resources. Therefore, fine-tuning can transfer the knowledge learned by the model from the original data set to the target data set. The weights of the pre-trained model are trained based on new data, which means fine-tuning can make the pre-trained model better adapt to new data. Fine-tuning can be performed on the whole neural network or only on a subset of its layers. It helps increase the model's generalization capabilities when the target data set is much smaller than the source data set. Fine-tuning is still the best technique to redeploy these pre-trained models to complete specific tasks. It usually consists of four steps:

1. Pre-train the neural network model on the source data set to obtain the source model. In our Project, we use LLAMA as the source model.
2. Retain all model designs and parameters on the source model, except the output layer. That is, preserve the pre-trained model's parameters.
3. Add an output layer to the target model with the same number of outputs as the target dataset's categories. The model parameters of this layer are then initialized at random.
4. Train the target model on the target data set. The output layer will be trained from scratch, while the parameters of all other layers are fine-tuned based on the parameters of the source model.

3.4.2. ICL (In Context Learning)

The core concept of In Context Learning (ICL) (Dong et al., 2022) is to learn via analogies. ICL requires some examples for creating a demonstration context. Natural language templates are frequently used to write these examples. The query and a contextual demonstration (typically several related cases) are combined to generate an input and then fed into the LLM for prediction. ICL is similar to the decision-making process of humans who learn by analogy. It makes incorporating human knowledge into LLM simpler since demonstrations are written in natural language. It is also a training-free learning framework as compared to supervised training.

- **Zero-shot learning**

Zero-shot learning refers to predicting new tasks using information gained from previous tasks without using any labels or examples of new task samples. In other words, zero-shot learning is the capacity to infer new knowledge from existing knowledge.

- **One-shot learning**

When only one sample is provided for learning, One-shot learning refers to using this sample to predict new tasks. In this situation, the model must understand features or patterns from a single example while adapting to new tasks.

- **Few-shot learning**

Few-shot learning is a comparable learning model to single-shot learning. Several or tens of samples tend to be used in few-shot learning for the model to learn to adapt to new tasks.

3.4.3. Training methods in the project

As conversations often involve a large amount of context and understanding, training methods working on context have been used in our project. Instruction-tuning and chain-of-thought all contribute to the model's generation and context understanding capacities, but their methods and goals are slightly different.

- **Instruction-tuning**

Instruction-tuning (Wei et al., 2021) is a method of guiding model learning by providing task-related instructions to the model. The purpose of this approach is to enable the model to understand the requirements of the task better and improve its generative and contextual understanding capabilities. Instruction-tuning typically does not provide learning examples and therefore requires less training data to enhance the generalization performance of a model.

- **Parameter-Efficient Finetuning (PEFT)**

Although full-parameter instruction fine-tuning is more effective, it still requires more computing resources. The parameter efficient fine-tuning (PEFT)(H. Liu et al., 2022) method can effectively adapt to various downstream tasks without fine-tuning the parameters of all models. Numerous effective parameter fine-tuning options available right now can significantly decrease fine-tuning costs while offering the same performance as full-parameter fine-tuning.

- **Low Rank Adaptation (LoRA)**

Currently, for large language models, one of the most effective methods is the LoRA (Low-Rank Adaptation) (Hu et al., 2021). In practice, LoRA can reduce the number of trainable parameters by 10,000. LoRA is a technique designed to simplify the complexity of large models by approximating their high-dimensional structure with a low-dimensional structure, and it can still perform well in a specific task or domain. The idea behind low-rank adaptation is that the high-dimensional structure of large models may contain redundant or irrelevant information for numerous tasks. By recognizing and removing this redundancy, we can create a model that keeps its original performance but requires fewer resources to train and deploy. The key LoRA adaptation procedure steps are as follows:

Initialization: Begin with a pre-trained language model and add a low-rank adaptation layer to its weight matrix. This layer is represented by a low-rank matrix with randomly generated values.

Fine-tuning: Train the model on a new task or domain, updating only the low-rank adaptation layer while keeping the pre-trained full model's weights constant. This enables the model to learn task-specific information while maintaining its general knowledge.

Prediction: Use a fine-tuned model to make predictions based on new, previously unseen data from a specific task or domain.

LoRA achieves more efficient fine-tuning by focusing the adaptation process on low-rank matrices. When compared to full fine-tuning, computational time and memory overhead are reduced. Furthermore, LoRA can be applied to a number of large-scale language models, such as GPT or BERT (Bidirectional Encoder Representations from Transformers), and it is easily adaptable to various tasks or fields. However, using low-rank matrices to approximate the original model's high-dimensional structure may introduce approximation errors, affecting the model's performance on specific tasks, especially for those tasks or domains that require a high level of granularity in understanding or are significantly different from the training distribution of the pre-trained model. And, though LoRA reduces the complexity of fine-tuning, dealing with very large models or adapting to a large number of tasks or domains may still require a significant amount of resources.

- **Conversation Tuning**

Conversation tuning is a special type of instruction tuning. The purpose is to allow large language models to unlock "conversation" capabilities on the basis of "completion" capabilities. It requires not only context understanding but also the completion of multiple rounds of tasks, so the format requirements are stricter. Otherwise, even a properly trained model will result in a significant number of performance errors. Here are several standard formats for model training:

ChatGLM

Input

```
Prompt: """[Round 0]
问：你好。
答：你好。
[Round 1]
问：你是谁？
答：我是人工智能ChatGLM。
[Round 2]
问：你会做什么？
答："""
```

Output

```
Complete: "我可以回答您提出的问题并且解决您的疑惑。"
```

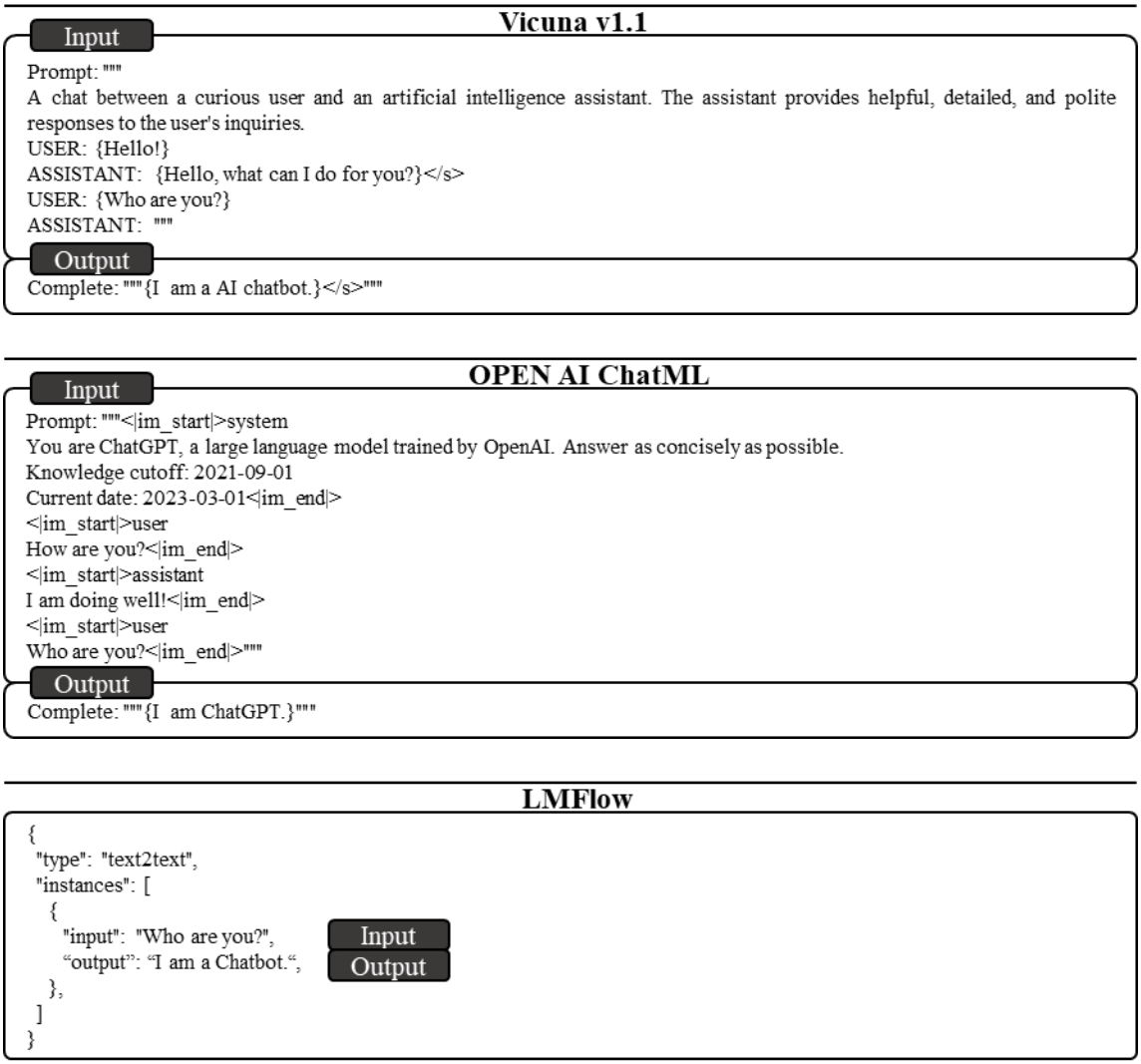


Figure 3.8 Data formats for different conversation tuning frameworks

The most significant distinction between conversation tuning and regular text finetuning lies in the end mark used after each dialogue, like the </s> in Vicuna and <|im_end|> in Open AI ChatML. It determines when the model should stop after a suitable reply. In this project, we experimented with the built-in Vicuna structure of FastChat and built our conversation prompt dataset structure using LMFlow. Since LMFlow doesn't have a specific training structure for conversation, we decompose the complete conversations and add conversation pairs round by round, through which the whole contextual information can be preserved entirely:

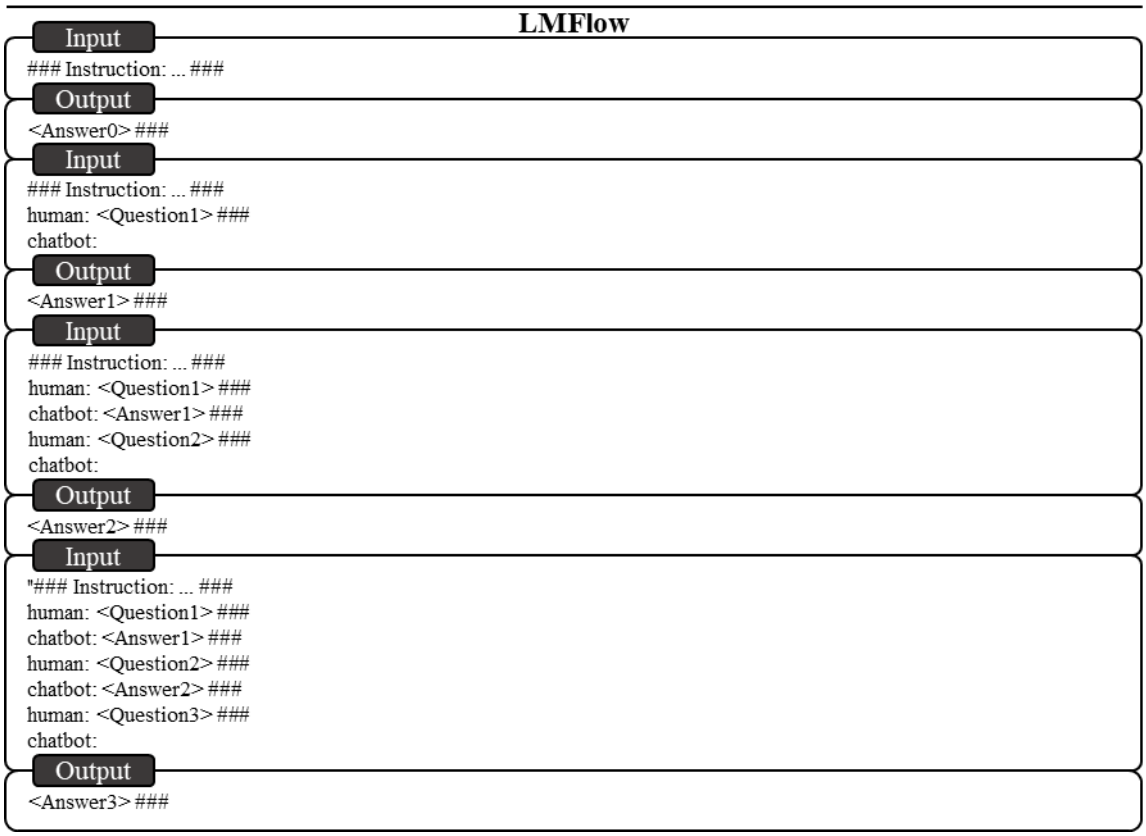


Figure 3.9 Data structure for conversation tuning by learning context in the LMFlow framework

3.5. TRAINING WORKFLOW

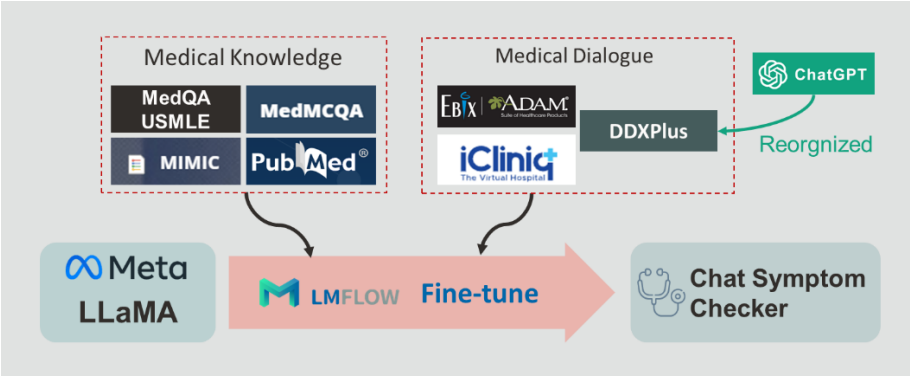


Figure 3.10 A summary of the training process of the Chat-Symptom Checker

During our model training process, LLaMA was selected and trained in two steps. First, the datasets of all medical knowledge were fed to the model to make the model have sufficient medical knowledge. Then, based on this, the model was fine-tuned with medical conversation datasets to unlock the multi-round conversation ability. The final model is trained using LMFlow workflow, and the training process is carried out on Orcale's cloud server using two A10 Nvidia graphics cards. The specific hyperparameter settings and descriptions are as follows:

Table 3.2 Hyperparameter details for training

Hyperparameter	Description
--num_train_epochs 2 \	The epochs we train the model on the collected dataset, which is 2.
--learning_rate 2e-5 \	The learning rate used to fine-tune the model, which is 2e-5.
--block_size 128 \	An integer indicating the optional input sequence length after tokenization. The training dataset will be truncated in blocks of 128 for training.
--per_device_train_batch_size 6 \	The per-gpu batch size for the fine-tuning, which is 6.
--use_lora 1 \	Where use LoRA or not. '1' means LoRA is used.
--lora_r 8 \	The rank of the lora parameters. The smaller lora_r is, the fewer parameters lora has. Here we set 8.
--bf16 \	Mixed precision mode, whether to use 'bf16' or 'fp16', here we use 'bf16' which has more precision.
--dataloader_num_workers 1 \	The number of processes to use for the preprocessing. Here we set 1.

3.6. EVALUATION

Evaluating the quality of LLMs typically involves multiple factors, such as the text's fluency, accuracy, relevance, and innovation. There are some widely used methods for evaluation and metrics, like BLEU, ROUGE, etc. These metrics are frequently used to evaluate and compare the quality of text generated by various models. However, they are more concerned with the accuracy than the ability of the content generated. For the multi-turn conversation LLM established in this project, particularly in the medical field, the ability to generate content that users feel comfortable with and the accuracy of the final guidance based on the conversation history is more crucial. Unfortunately, there is a lack of mature strategies for evaluating medical LLMs. A common method is using medical QA databases, such as MedQA-USMLE, MedMCQA, etc., to evaluate them. Due to the fact that these databases only contain binary or multiple-choice questions, they can only evaluate the LLM's decision-making ability and only based on a single round conversation.

As the conversation moves on more than two rounds, it becomes challenging to predict the user's inputs and the variety of responses generated by LLM. Traditional evaluation metrics are insufficient to evaluate system performance in multi-turn conversation scenarios. Human evaluation still appears to be the best solution, in which human evaluators score the generation by comparing LLM-generated text to a set of predetermined criteria. In the medical field, for instance, physicians are necessary to evaluate the precision, fluency, and relevance of generated content and provide feedback. However, manual evaluation is expensive and requires the participation of a number of healthcare professionals, making it difficult to

implement. In this Project, it is not being carried out, but it may be carried out in the future with some cooperation from some doctors in the hospital.

To evaluate the potential performance of our model, we evaluated the quality of the DDXPlus dataset introduced before, which is crucial for unlocking the ability of multi-round conversations. As we all know, the performance of a machine learning model is fundamentally determined by data. Especially in the medical field, the quality of conversation will directly affect the final diagnosis results of the LLM. So, the evaluation of data quality on other models can indirectly reflect the possibility of our LLM's potential best performance. Using BERT and clinicalBERT, which are also based on the Transformer architecture, we evaluated various DDXPlus datasets with different data structures (Table 3.1).

The entire conversation is treated as the input text, with the department as the target. Each dataset has been split into a training set of 70%, a test set of 20%, and a validation set of 10%. Since our model aims to guide patients to particularly suitable departments, accuracy on the validation set is the metric with which we are most concerned.

4. RESULTS AND DISCUSSION

Figure 4.1 A conversation talked with Chat-Symptom Checker

4.1. CHAT-SYMPTOM CHECKER

After training with a large quantity of data, we succeeded in creating the Chat-Symptom Checker. LLM is able to process complex and diverse natural language inputs, enabling users to receive more accurate and individualized feedback when describing their symptoms. According to the example displayed (Figure 4.1), Chat-Symptome Checker interacts with users in a natural and understandable manner with few medical terms to provide a more user-friendly experience. For the health system like hospitals, this natural interaction method not only lowers the user threshold, but also attracts more users to the platform and increases their trust and satisfaction.

We also found that after training and acquiring a large amount of medical knowledge, Chat-Symptome Checker collected not only the patient's own information, but also the user's past medical history and family medical history, even as well as the patient's travel history. Combining all the information, it finally provided the recommendation to assist the user in going to the correct department. In advance, suggestions and potential diagnoses are provided as a reference.

Implementing the LLM model also suggests that automated symptom-checking services can significantly reduce labor costs while improving service efficacy. For instance, if our model is implemented on a mobile system, it is able to automatically process a large number of user queries, by which patients can go directly to the recommended department for medical treatment, dramatically reducing the waiting time.

After the user interacts with the model, we can also obtain a complete record of the conversation through the back-end log file, which can be used for the EHR system in the future. Meanwhile, primary doctors can quickly review this record and lower its recommendation error rate. Importantly, since there are few clinical conversation data records between doctors and patients, a large quantity of saved conversation data is helpful to further the continuous optimization and improvement of the LLM-based system, which is also essential for the long-term development of hospital and medical research.

Obviously, our model also experienced challenges with inference. The most common problems we face are asking endless questions without entering the final recommendation turn, occasionally asking questions that have been asked before, and occasionally asking illogical questions. For example, where the user's answer does not have any pain, the Chat-Symptom Checker still asks about the location and severity of the pain. Due to the above problems, this LLM still needs some improvement.

4.2. EVALUATION OF TEXT QUALITY

After testing various models on various types of datasets, we've discovered that for language models, it is not present that more data is better. As language models are incapable of learning logical reasoning, it depends more on the quality of the text. The results indicate that the

dataset containing only 5k GPT rewritten records achieved the highest level of accuracy and F1 score, indicating that the DDXPlus structured QA dataset constructed by selecting questions and answers from the text library does not appear capable of achieving the same text quality as GPT generated. This may be due to the random pairing of questions and answers in the structured DDXPlus QA dataset, which may result in a lack of contextual relevance and coherence. GPT3.5 is able to comprehend the context of the input and generate more relative responses. Due to this, the generated text is also frequently better in terms of logical structure. Since text is chosen from a limited number of sentences, the style and tone of the DDXplus QA dataset are quite limited, whereas the GPT can generate text with a greater variety of styles and tones. Nevertheless, a static QA library may become ineffective and unable to quickly adapt to altering environments or current events.

Table 4.1 Text quality evaluation of various datasets on BERT and ClinicalBERT models

Dataset	Test Model	Accuracy	Precision	Recall	F1-score
data_5k	BERT	**0.97**	**0.97**	**0.97**	**0.97**
data_5k	ClinicalBERT	**0.98**	**0.97**	**0.98**	**0.98**
data_8k_without_no	BERT	0.94	0.95	0.94	0.94
data_8k_without_no	ClinicalBERT	0.93	0.94	0.93	0.93
data_8k	BERT	0.83	0.84	0.83	0.83
data_8k	ClinicalBERT	0.75	0.76	0.75	0.74
data_9k_without_no	BERT	0.92	0.93	0.92	0.92
data_9k_without_no	ClinicalBERT	0.92	0.92	0.92	0.91
data_9k	BERT	**0.77**	**0.77**	**0.77**	**0.76**
data_9k	ClinicalBERT	**0.67**	**0.71**	**0.67**	**0.66**
data_11k_without_no	BERT	0.96	0.96	0.96	0.96
data_11k_without_no	ClinicalBERT	0.90	0.90	0.90	0.90
data_11k	BERT	0.89	0.89	0.89	0.89
data_11k	ClinicalBERT	0.83	0.82	0.83	0.82

A result demonstrates that language models depend highly on text description. From the table, we can find that the artificial DDXPlus QA datasets we created using the same original data perform very differently in various text types. When constructing the artificial dataset to simulate real-world medical conversation scenarios, we included some symptom check questions with negative answers. Because physicians cannot only inquire about positive symptoms in real scenarios. Nevertheless, based on the results of the evaluation, the accuracy and F1 scores decreased significantly in all datasets that included negative answered symptom check questions. Additionally, in these data sets, adding the 1k only one initial positive symptom dataset yielded worse results than the 8k dataset. Then, the accuracy and F1 score recovered with the addition of the 2k one-round conversation dataset. This demonstrates that the language model only makes the final determination based on the text content, that is, the number and types of all questions that appear in the entire conversation content, instead of making logical determinations based on the ‘Yes’ or ‘No’ answers. For those datasets that

lack negative answered symptom check questions, all questions refer to positive symptoms, so model performances will not be too bad, whose accuracies are all more than 90%. For the dataset containing only one initial positive symptom, more negative answered symptom check questions will interfere with the model, dramatically reducing the prediction accuracy.

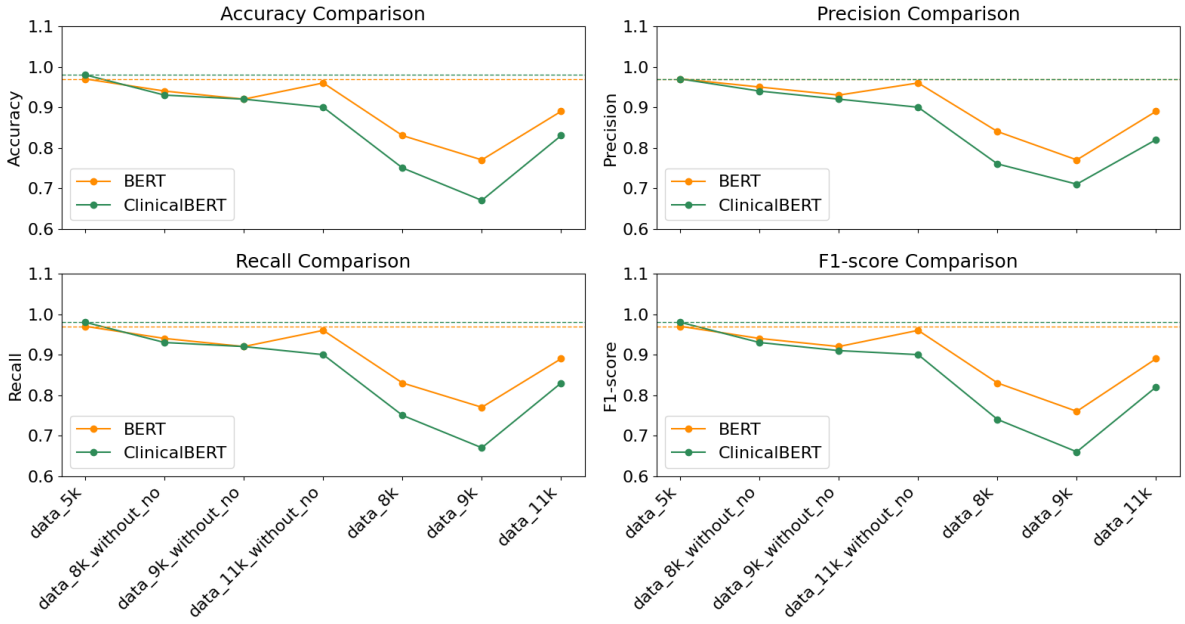


Figure 4.2 Model performance comparison results across datasets

We also found an interesting thing from this result. The ClinicalBERT model trained with a large amount of medical data does not perform better than the BERT model. In several data sets, its accuracy and F1 score are even lower than those of BERT. This phenomenon could be the result of gaps between the specific clinical notes trained by ClinicalBERT and actual usage scenarios. The complexity of the medical field isn't sufficiently explained in these training texts, which reduces generalization ability. In addition, clinicalBERT may be overfitted for certain tasks or data types, limiting its ability to generalize to a wider range of medical texts. While basic BERT, on the other hand, maybe more adaptable due to its training on a large number of basic corpora, particularly when the classification task does not heavily rely on medical knowledge. This also reflects the difference in building a symptom checker language model that interacts with users because its text content relies more on daily language than medical terminology but still involves knowledge and logic in the medical field.

5. CONCLUSIONS AND FUTURE WORKS

In this study, we succeeded in developing a symptom checker LLM that recognizes health symptoms and provides the medical department recommendation. The model can understand input in natural language and ask questions using non-medical terms to make patients easier to understand. Users who are not medical professionals are able to use the symptom checker to describe medical conditions. Based on the ability of multi-round conversation, the model collects more medical information and gives more accurate recommendations.

However, our models are currently limited not only by the variety and quality of the dataset but also by the shortage of LLMs. The main issue is that LLM, as a language model, has rather weak logic abilities. Because it mainly performs inference by learning statistical features from a large amount of text rather than simulating human logical thinking processes. For example, the "Reversal Curse" is a significant drawback of LLM(Berglund et al., 2023). An LLM cannot automatically generalize to "B is A." when it is trained on sentences of the form "A is B.”. The reverse of the curse demonstrates that following training, LLM is unable to correctly respond to inquiries represented in the opposite direction of the training data, and the accuracy is even close to 0%. This is a significant challenge that our model will face because we cannot predict every possible language format in such complex medical conversation scenarios. As a result, the model's accuracy is likely to drop significantly.

The hallucination of LLMs (Ji et al., 2023) describes the occurrence of false or incorrect information during the generation of LLMs. These errors or inaccuracies may be caused by noise in the training data, parametrized knowledge errors, attention mechanism errors, improper training strategies, or inference exposure biases. In our project, we discovered that we should pay particular attention to the punctuation marks in the training data like ":", "!", etc., since they might have quite different functions in various training frameworks. Take Fastchat as an example; the character ':' is used as the end token to denote the conversation role. Significant changes are likely to happen when the ':' appears in the training data. Hallucinations of LLMs are relatively simple to recognize and solve in summary generation and neural machine translation tasks. But in conversational systems, it's essential to achieve a balance between variety and consistency in conversational responses, which indicates the solution of how to reduce hallucinations is still in its early stages in this field.

Since it's important for the model to learn the statistical rules of language and to understand the requirements of particular tasks, the prompt of LLM's training data is quite crucial. When label data is limited, good prompts can significantly improve the model's performance on particular tasks while also enhancing its interpretability and controllability in usages in particular fields. As a result of our previous evaluations of various databases, it has been discovered and proved that different data structures significantly affect the outcomes and even improve the performance of the model on smaller datasets. In addition, as we are

developing a multi-turn conversation model, the model's output will depend on the context given before. The length and structure of the context should be optimized as well, given that a study has shown that longer contexts do not always lead to more information being packed into the model and better responses (Lewis et al., 2020).

The modern LLM training process includes RLHF as an essential component, which helps refine and generalize LLM so it can perform well in specific tasks and fields by improving the efficiency and safety of the model. Despite the fact that the costs could be very high, and the implementation will be difficult in the medical field, the model can be aligned with clinical use cases through RLHF carried out by medical professionals. This is essential for enhancing the efficiency and precision of models in actual medical scenarios and cannot easily be substituted by any other approach.

To increase the potential for generalization and efficiency of the model, our future work will be focused on expanding the data sources to include more symptoms and disease types. The Symptom Checker LLM will turn out to be a simple, affordable, and effective medical guidance tool with additional study and improvement.

BIBLIOGRAPHICAL REFERENCES

Abensur Vuillaume, L., Turpinier, J., Cipolat, L., Arnaud Dépil, D., Dumontier, T., Peschanski, N., . . . Galland, J. (2023). Exploratory study: Evaluation of a symptom checker effectiveness for providing a diagnosis and evaluating the situation emergency compared to emergency physicians using simulated and standardized patients. *PLoS One, 18*(2), e0277568. doi:10.1371/journal.pone.0277568

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. J. a. p. a. (2023). The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A".

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. J. a. p. a. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Diao, S., Pan, R., Dong, H., Shum, K. S., Zhang, J., Xiong, W., & Zhang, T. J. a. p. a. (2023). Lmflow: An extensible toolkit for finetuning and inference of large foundation models.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., . . . Sui, Z. J. a. p. a. (2022). A survey for in-context learning.

Fansi Tchango, A., Goel, R., Wen, Z., Martel, J., & Ghosn, J. J. A. i. N. I. P. S. (2022). Ddxplus: A new dataset for automatic medical diagnosis. *35*, 31306-31318.

GitHub. (2023). Copilot Retrieved from https://github.com/features/copilot

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., . . . Stanley, H. E. J. c. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *101*(23), e215-e220.

Google. (2023). PaLM2. Retrieved from https://ai.google/discover/palm2/

Gräf, M., Knitza, J., Leipe, J., Krusche, M., Welcker, M., Kuhn, S., . . . Klemm, P. J. R. I. (2022). Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. *42*(12), 2167-2176.

Hennemann, S., Kuhn, S., Witthöft, M., & Jungmann, S. M. (2022). Diagnostic Performance of an App-Based Symptom Checker in Mental Disorders: Comparative Study in Psychotherapy Outpatients. *9*(1), e32832. doi:10.2196/32832

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., . . . Chen, W. J. a. p. a. (2021). Lora: Low-rank adaptation of large language models.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., . . . Fung, P. J. A. C. S. (2023). Survey of hallucination in natural language generation. *55*(12), 1-38.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. J. A. S. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *11*(14), 6421.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. J. a. p. a. (2019). Pubmedqa: A dataset for biomedical research question answering.

Johnson, A., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV-Note: Deidentified free-text clinical notes. In: PhysioNet.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Rocktäschel, T. J. A. i. N. I. P. S. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *33*, 9459-9474.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. J. C. (2023). ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *15*(6).

Liu, A. W., Odisho, A. Y., Brown III, W., Gonzales, R., Neinstein, A. B., & Judson, T. J. (2022). Patient Experience and Feedback After Using an Electronic Health Record–Integrated COVID-19 Symptom Checker: Survey Study. *9*(3), e40064. doi:10.2196/40064

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. A. J. A. i. N. I. P. S. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *35*, 1950-1965.

lm-sys. (2023). FastChat. Retrieved from https://github.com/lm-sys/FastChat

Maturana, C. R., de Oliveira, A. D., Nadal, S., Bilalli, B., Serrat, F. Z., Soley, M. E., . . . Abelló, A. J. F. i. m. (2022). Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review. *13*, 1006659.

Mohammed, M. A., Mohammed, V., Mohammed, V. A. J. I. J. o. S. E., & Applications. (2022). Impact of Artificial Intelligence on the Automation of Digital Health System. *13*.

OpenAI. (2022). ChatGPT. Retrieved from https://chat.openai.com/

OptimalScale. (2023). LMFlow. Retrieved from https://github.com/OptimalScale/LMFlow

Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022). *Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering.* Paper presented at the Conference on Health, Inference, and Learning.

Peven, K., Wickham, A., Wilks, O., Kaplan, Y. C., Marhol, A., Ahmed, S., . . . Fenech, M. J. m. (2023). Assessing the accuracy of a digital symptom checker tool for suggestion of reproductive health conditions: a clinical vignettes study. 2023.2002. 2022.23286305.

Significant-Gravitas. (2023). AutoGPT. Retrieved from https://github.com/Significant-Gravitas/AutoGPT.git

Simpson, L. A., Macdonald, K., Searle, E. F., Shearer, J. A., Dimitrov, D., Foley, D., . . . Shenoy, E. S. (2022). Development and deployment of tools for rapid response notification of Monkeypox exposure, exposure risk assessment and stratification, and symptom monitoring. *Infect Control Hosp Epidemiol, 43*(8), 963-967. doi:10.1017/ice.2022.167

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., . . . Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. In.

tatsu-lab. (2023). stanford_alpaca. Retrieved from https://github.com/tatsu-lab/stanford_alpaca

Thissen, M. (2023). Vicuna-13B: Best Free ChatGPT Alternative According to GPT-4 | Tutorial (GPU). Retrieved from https://medium.com/@martin-thissen/vicuna-13b-best-free-chatgpt-alternative-according-to-gpt-4-tutorial-gpu-ec6eb513a717

Wang, G., Liu, X., Ying, Z., Yang, G., Chen, Z., Liu, Z., . . . Chen, Y. (2023). Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine, 29*(10), 2633-2642. doi:10.1038/s41591-023-02552-9

Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., & Liu, T. J. a. p. a. (2023). Huatuo: Tuning llama model with chinese medical knowledge.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., . . . Le, Q. V. J. a. p. a. (2021). Finetuned language models are zero-shot learners.

You, Y., & Gui, X. (2020). Self-Diagnosis through AI-enabled Chatbot-based Symptom Checkers: User Experiences and Design Considerations. *AMIA Annu Symp Proc, 2020*, 1354-1363.

Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., . . . Zhang, R. (2020). *MedDialog: Large-scale medical dialogue datasets.* Paper presented at the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. J. a. p. a. (2021). Dive into deep learning.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., . . . Xing, E. J. a. p. a. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena.

Zobel, M., Knapp, B., Nateqi, J., & Martin, A. J. P. o. (2023). Correlating global trends in COVID-19 cases with online symptom checker self-assessments. *18*(2), e0281709.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa