Medicine based on Symptoms: Improving Large Language Models to Answer Multiple Choice Questions

Arnav Dedhia

Abstract

In recent years, Large Language Models (LLMs) have become known for answering multiple different types of questions. However, these models give the wrong answer many times, and that may be a problem in cases where mistakes may cost lives. In this paper, I present ways to help increase the accuracy of multiple models based on the MedQA dataset. I have trained models on textbooks, Q&As, and prompts to have a fuller understanding of what methods work the best. These AIs in the field of Biomedics have a prominent future and are currently understudied. If LLMs can learn how to accurately answer multiple choice questions, they will be able to help people from all over the world, and all from home. With the methods explained in the paper, the accuracy of the model has increased over 15 percent with the highest being a model 55 percent accuracy. Overall, the MedQA dataset presents great challenges to multiple models, and I hope to use the dataset to provide a deeper understanding of the methods that can be used to increase the accuracy of LLMs.

1. Introduction

Recent advances in LLMs have prompted a deeper dive into how these computer models work and the shortcomings that come with them. LLMs such as GPT-4, Llama 2, or Mistral 7 are very powerful natural language processors that are being developed more and more each day. These LLMs now require people to adapt them to the multiple different purposes, domains, that a model could be used in. These domains could be video game analysis, vacation planning, or based off of medical diseases as this paper covers.[1] With the backend training that each model has gone through, they can easily surpass the knowledge of a normal human, and in some cases, even professionals. LLMs have a very specific way they can answer a prompt that they have been given. In their memory, they store hundreds of millions of connections between words, phrases, and concepts. When given a prompt, the LLM has to look through these multi-layered neural networks to decipher what words have the top probabilities of answering the question at hand. Using its previous knowledge and related context, the LLM can almost accurately predict what the prompt is asking. In turn it can give an open

[1] https://thedatascientist.com/5-best-examples-of-domain-specific-llms-in-ai/#:~:text=Finance%20tops%20the%20list%20when,big%20on%20stability%20and%20accuracy.

```
{"question": "A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication. The attending tells the resident that the patient will do fine, and there is no need to report this minor complication that will not harm the patient, as he does not want to make the patient worry unnecessarily. He tells the resident to leave this complication out of the operative report. Which of the following is the correct next action for the resident to take?", "answer": "Tell the attending that he cannot fail to disclose this mistake", "options": {"A": "Disclose the error to the patient but leave it out of the operative report", "B": "Disclose the error to the patient and put it in the operative report", "C": "Tell the attending that he cannot fail to disclose this mistake", "D": "Report the physician to the ethics committee", "E": "Refuse to dictate the operative report"}, "meta_info": "step1", "answer_idx": "C"}
```


Figure 1: Sample question for the MedQA Dataset

ended answer that is similarly worded and fairly correct.[2]

However, contrary to popular beliefs, LLMs have a harder time answering multiple choice questions that are more niche or ones that require more complicated thought processes.[3] This is because the LLMs are restricted to only being able to choose out of the options given. When they are given the opportunity to answer an open ended question, they can write whatever they see fit and whatever flows the best. However, when there are only a set of multiple choice answers, the LLM has to do some extra steps to get to the answer.

With multiple choice answers, the LLMs must first think of what an open ended response to the question would be. Using this answer, they have to take it to the next level by finding the closest answer choice out of the ones given. When the LLM makes its own open ended answer, it has the freedom of choosing anything, however, the answer it has thought of may not be in the answer choices given. This means that the machine has to reason its way into picking the closest answer choice. This is usually where the largest proportion of errors occur.[4]

To test this theory, I have decided to use the MedQA dataset. Through the multiple experiments that have already been completed on this dataset, I have convincing evidence that the questions included in this set have a significantly lower accuracy than other datasets with similar questions.[5]

2. Data

The data used in this paper is the MedQA dataset created by Jin et al. This dataset is very accessible to the public by going to their github and downloading the link.[6] Within this data, there are multiple languages in which a choice is given to train and test a model. These languages include: Chinese, Taiwanese, and English. Depending on the goal of a model any of these languages can be used.

For the sake of this paper, I will be using all the data given from the English language section. However, when people will be using this model, they will have the choice of these 3 languages. This means that the 2.5 billion

[2] https://nexocode.com/blog/posts/generative-question-answering-llms/#:~:text=Answer%20Generation%3A,contextually%20accurate%20but%20also%20insightful.

[3] https://openreview.net/pdf?id=shr9PXz7T0

[4] https://openreview.net/pdf?id=shr9PXz7T0

[5] https://arxiv.org/pdf/2009.13081

[6] https://github.com/jind11/MedQA

people who speak these languages could have access to this service.[7]

2.1 Textbooks

The MedQA dataset includes 18 large medical textbooks. These textbooks are fairly popular throughout medical students, showing that training the model on these may prove to be useful. Each textbook is not only written at a very high level, but all of them contain tens of thousands of paragraphs. The combination of these factors help make it a very viable option when training the LLM. Since answering the questions in the data requires very vast and deep medical knowledge, these textbooks can provide the model with access to a good background and understanding on what many of the diseases and terms mean. Also included in the texts are common symptoms of diseases, side effects of certain medications, and actions one can take when faced with a certain problem.

The text in the textbooks were well organized and once opened up into a text document, it needed no pre-processing. The pre-processing of the data was already done by the publishers of the dataset. The text was ready to be sent into the model as training data.

In a previous study, 100 random questions were taken from the question set and given to a medical professor. When looking through the text, they were only able to find 88% of the answers were covered in at least one of the texts.[8]

2.2 Questions

The MedQA dataset additionally includes thousands of different quiz type questions. Each one of these are different from one another. The questions of the dataset are split up into 3 categories: Questions with 5 answer choices, Questions with 5 answer choices and "metamap_phrases," and Questions with 4 answer choices and "metamap_phrases." The metamap phrases referred to are 15 - 20 key words inside the prompt that humans or computers could use to get a better understanding of the prompt. This would be especially useful for LLMs as with the key words, they could narrow down their search and get better results.[9]

In this study, I chose to use the simple Questions with 5 answer choices. These are the first set of "train", "test", and "dev" files to appear in the questions folder. This set of questions were specifically chosen as it best fits the main question of this paper: how average people could use this to know what medicine to use. In real life, the people who would use this should not be required to add a list of keywords for this to work. Another reason this data set was chosen is because there are simply more answer choices for the users to add. A normal household would have more than 4 treatments, so it is beneficial to allow more answer choices into this new model.

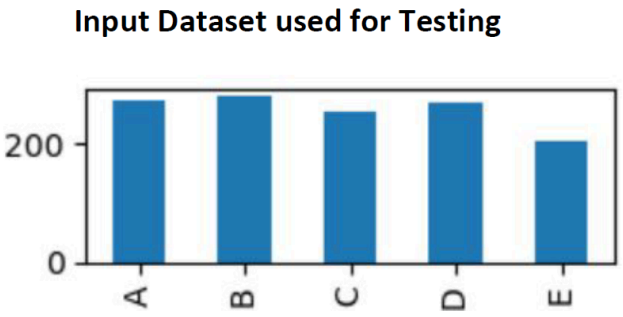
As shown in Figure 1, a question in the dataset includes multiple different keys and values that store information. The keys include question, answer, options, meta_info, and answer_idx. By having this pre-processed into a format like this makes it very simple to analyze initial data. For example, Graph 1 shows the distribution of correct answers. The graph shows B having the most correct answers, but each letter choice has a fairly similar amount. With this information, I know there is little to no bias in ordering of answer choices.

---

[7] https://gurmentor.com/what-is-the-most-spoken-language-in-the-world/

[8] https://arxiv.org/pdf/2009.13081

[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6309052/



Graph 1: Number of correct answer choice letters

The data set for each of the 3 types of questions included a total of 12,730 questions. These 12,730 questions were split up into a ratio of 8:1:1 of training : testing : and dev questions. I had no need to use the set of questions designated for hyper-parameter tuning, so I combined the dev with the training questions. Totalling to more than 11,000 training questions, there were enough to have conclusive results at the end of the experiment.

3. Methods

In this study I decided to use 5 different types of methods to train my 3 models: Llama 2 7b, Llama 2 70b, and Mistral 2 7b. Although Llama 2 7b and 70b are using the same base model, I chose to use both because Llama 2 70b is trained with 70 billion parameters while the other is only trained on 7 billion.[10] Mistral 2 7b was also included into this study so I can make a conclusion on how different base models compare with each other.[11]

The methods I used for tuning my models were Textbook, Instruction, Prompt, and Instruction & Prompt Training. There was also a control of the model itself without any training on my end. Each of these are a form of Fine-Tuning as different parameters were passed in in hopes of creating the most optimal version of the model.

3.1 Textbook Tuning

In this type of tuning I gave the models all 18 of the textbooks that were included in the MedQA database. These books, as previously mentioned, are very popular among students in studying for a multitude of different medical exams. With access to such a valuable resource that contains hundreds of thousands of words, there was a good chance that the models would adequately tune on the textbooks. However, one concern I had was that due to the large volume of items it takes to train a LLM, these books might have already been previously used as training data.

Nevertheless, due to the fact that around 88% of the answers to the testing question are in the textbook, I hoped that the model would get enough knowledge from these passages to pass around 70% of the questions.[12] With LLMs continuously getting better at finding answers to questions in texts, I thought that this method would have a very good chance of attaining a high accuracy.

3.2 Instruction Tuning

In this type of tuning, I gave the models all the questions and answers as training data. With the 11,400 questions in the training and dev dataset, there were more than enough questions for the models to understand how the questions are formatted and learn from them.[13] In addition to format, the model can work backwards to learn more about the diseases and treatments provided in the questions.

[10] https://blog.nimblebox.ai/choosing-the-right-llama2-model

[11] https://www.simplypsychology.org/experimental-designs.html

[12] http://arxiv.org/pdf/2009.13081

[13] https://arxiv.org/pdf/2308.10792

```
"You are a doctor whose job is to diagnose the patient with the correct symptom or tell them
what the next steps should be.  You will get a scenario with 5 answer choices, of which only
1 will be correct. The answer choice may resemble a disease or next steps doctor would take
based on symptoms in question. If the question is to find the disease, work backwards from
the answer choice. For each choice find the symptoms and then match them up with ones in the
question. Eliminate the 3 choices that are obiviously not matching up. For the 2 remaining
choices, think again step by step on the symptom and decide the best answer. If the answer
choices resemble a procedure on what to do next, then you have to find the disease first. And
then find the next steps for each disease. Finally compare those with the options. Eliminate
the 3 choices that are not appropriate and for the remaining 2 choices, do step by step
thinking. Pick one final answer. Restrict your response to only provide option Alphabet and
its value in the format as shown in the example. Do not provide explanation for choice or any
other text except the value in answer tag.\
Example:<symptom>What is
2+2</<symptom><option>\'A\':\'3\',\'B\':\'5\',\'C\':\'4\',\'D\':\'7\'</option>
<answer>C.4</answer> \
Question: <symptom> {symptom}</symptom> \
<option> {option}</option> "
```

Figure 3: Prompt given to all models for training purposes

{"question": "A 21-year-old sexually active male complains of fever, pain during urination, and inflammation and pain in the right knee. A culture of the joint fluid shows a bacteria that does not ferment maltose and has no polysaccharide capsule. The physician orders antibiotic therapy for the patient. The mechanism of action of action of the medication given blocks cell wall synthesis, which of the following was given?", "answer": "Ceftriaxone", "options": {"A": "Chloramphenicol", "B": "Gentamicin", "C": "Ciprofloxacin", "D": "Ceftriaxone", "E": "Trimethoprim"}, "meta_info": "step1", "answer_idx": "D"}

Figure 2: Sample question from the Dev dataset in MedQA

As shown in Figure 2, by giving this specific question to the models, I hoped that it could figure out more about Ceftriaxone. From this question, the model could have gotten the information that Ceftriaxone is given to people who have fevers, pain during urination, and inflammation in the right knee. Although not all these are accurate, the model will now have this information in the future to help make a better guess.

Given that the model's training set was 11,400 questions, there was bound to be a number of times where this specific medication was mentioned. By factoring in all the other instances of Ceftriaxone being a correct answer, the model would work backwards and learn more about the commonalities of said questions.[14] In turn, it would create a repertoire of use cases of this medicine.

3.3 Prompt Tuning

In this type of tuning, I gave the models a self created prompt that would explain how to solve the multiple choice questions in the testing dataset. These instructions would serve as guidelines for the models to follow to get a better chance of getting the answer correct.[15] The given prompt was also very basic so that it could apply to all types of questions and give the best guidance for each.[16]

As shown with the 12 sentences written in Figure 3, I split up the prompt into an "if else" statement. In a sample taken of 100 random questions from the training dataset, 89% of the questions were requiring an answer to be in the

[14] https://www.ibm.com/topics/instruction-tuning#:~:text=Instruction%20tuning%20is%20a%20technique%20for%20fine-tuning%20large,thus%20helping%20adapt%20pre-trained%20models%20for%20practical%20use.

[15] https://www.youtube.com/watch?v=yu27PWzJI_Y&t=1s

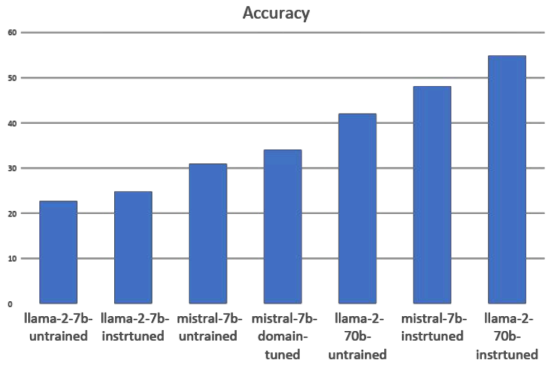
[16] https://www.superannotate.com/blog/llm-prompting-tricks#:~:text=26%20prompting%20techniques%201%201.%20No%20need%20to,8%208.%20Format%20your%20prompt%20...%20More%20items

form of a medication or the procedure to do next.[17] Because these 2 question types covered a very large percent of the data, it would only confuse the model if more instructions were added to the prompt.

The prompt also gives a good step by step process that would guide the model through each question. By following the line of steps, the models would have gotten a better understanding of both the question posed and the scenario given. Both of these help get a better total accuracy.[18]

3.3 Instruction & Prompt Tuning

In this type of tuning, I gave the models a combination of the prompt I made and all the questions and answers in the MedQA dataset. By giving the models both of these training



Graph 2: Accuracies of all 3 model with all the different types of initial tuning

items at the same time, it can learn from the prompt while using the questions to form a connection. The models would adjust the weightages of the neural network links as normal, except in this case, the model will start with more knowledge.[19]

By having a chain of thought to follow and learn from, the models can save hundreds of thousands of questions learning the basic process.[20] With the extra thousand questions now available, I hope that the model can tune itself till it can get a majority of the questions correct. Because this tactic combines the previous two, I believe that this will have the greatest accuracy and precision scores.

4. Results and Analysis

By performing the actual experiment on the training data of the MedQA dataset, I have data that is easily analysable. In Graph 2, it is clear that Llama 70b is the best model. However, the prompt I created had little to no impact on the accuracy of models tested.

As expected the Instruction + Prompt fine-tuning on both Mistral 7b and Llama 2 70b was the most accurate. The more instructions given to the model, the faster it can understand how to connect the links of the internal neural networks.

When looking at Graph 2, the numbers suggest that the most helpful fine tuning was in fact the instruction training. By giving the 3 models a set of questions with answers, the Llama 2 70b was able to increase its accuracy by 12% and Mistral 7b was able to increase its accuracy by 14%.

Without a fine-tuning on the types of questions that will be asked, all the models underperformed by quite a large margin. Although, with the instruction tuning, the accuracy went from 20%, by guessing, to over 50%. This is not perfect, but much better than

[17] https://aclanthology.org/W19-5039.pdf

[18] https://www.superannotate.com/blog/llm-prompting-tricks#:~:text=26%20prompting%20techniques%201%201.%20No%20need%20to,8%208.%20Format%20your%20prompt%20...%20More%20items

[19] https://www.geeksforgeeks.org/large-language-model-llm/

[20] https://www.techtarget.com/searchEnterpriseAI/definition/chain-of-thought-prompting

what a common person at home could get. I can conclude that the future tuning of this model will rely on different methods using some sort of instruction tuning algorithm.

One example, not covered in this paper, is running through the data and having the model classify the questions. This could either be classifying them by what the answer choices are asking, or splitting by situations with similar symptoms. By having the data split up prior to the training, the models might be able to better understand how to get more accurate answers on the testing set.

5. Conclusion

I have presented my findings of the models Llama 2 7b, Llama 2 70b, and Mistral 7b on the questions provided by the MedQA dataset. Although the expectations of nurses and doctors being able to use these improved models were not met, a lot of valuable information was gained.

The MedQA dataset has a lot of realistic use cases for more than just training Machine Learning models. With the accessibility to 3 different languages and the 15+ textbooks in each language, it is a great study resource for those taking medical exams.

The different methods in the 3 models also show us which ways of training were universally impactful versus which ones only worked for models with a high number of initial parameters. By using this new knowledge I hope further research is conducted on the multitude of different training methods currently known.

When models like these get to a high enough accuracy, they could potentially be able to be released to everyone so that they can get more information on the ointment they should be using, based on their current situation. With my research, I am 1 step closer in achieving that final goal.

References

Abacha, Asma, and Dina Demner-Fushman. "Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering Chaitanya Shivade 2." Association for Computational Linguistics, 2019. https://aclanthology.org/W19-5039.pdf.

Craig, Lev. "Chain-of-Thought Prompting." Enterprise AI. TechTarget, 2024. https://www.techtarget.com/searchenterpriseai/definition/chain-of-thought-prompting#:~:text=Chain%2Dof%2Dthought%20prompting%20is,way%20that%20mimics%20human%20reasoning.

dorota-owczarek. "Generative Question Answering over Documents with LLMs." nexocode. nexocode, December 11, 2023. https://nexocode.com/blog/posts/generative-question-answering-llms/#:~:text=Answer%20Generation%3A,contextually%20accurate%20but%20also%20insightful.

erika. "5 Best Examples of Domain-Specific LLMs in AI - the Data Scientist." The Data Scientist, April 17, 2024. https://thedatascientist.com/5-best-examples-of-domain-specific-llms-in-ai/#:~:text=Finance%20tops%20the%20list%20when,big%20on%20stability%20and%20accuracy.

GeeksforGeeks. “What Is a Large Language Model (LLM).” GeeksforGeeks. GeeksforGeeks, June 4, 2023. https://www.geeksforgeeks.org/large-language-model-llm/.

Gurmentor. “What Is the Most Spoken Language in the World in 2023.” Encore!!!, April 23, 2021. https://gurmentor.com/what-is-the-most-spoken-language-in-the-world/.

IBM Technology. “What Is Prompt Tuning?” YouTube Video. *YouTube*, June 16, 2023. https://www.youtube.com/watch?v=yu27PWzJI_Y&t=1s.

Ibm.com. “What Is Instruction Tuning? | IBM,” April 4, 2024. https://www.ibm.com/topics/instruction-tuning#:~:text=Instruction%20tuning%20is%20a%20technique%20for%20fine-tuning%20large,thus%20helping%20adapt%20pre-trained%20models%20for%20practical%20use.

Jin, Di, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. “What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams.” *Applied Sciences* 11, no. 14 (July 12, 2021): 6421. https://arxiv.org/pdf/2009.13081

———. “What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams.” *Applied Sciences* 11, no. 14 (July 12, 2021): 6421. https://doi.org/10.3390/app11146421.

jind11. “GitHub - Jind11/MedQA: Code and Data for MedQA.” GitHub, 2020. https://github.com/jind11/MedQA.

Nimblebox.ai. “Choosing the Right Llama 2 Model - Plug-And-Play MLOps Platform | NimbleBox.ai,” 2023. https://blog.nimblebox.ai/choosing-the-right-llama2-model.

Simply Psychology. “Experimental Design: Types, Examples & Methods,” July 31, 2023. https://www.simplypsychology.org/experimental-designs.html.

Superannotate.com. “26 Prompting Tricks to Improve LLMs | SuperAnnotate,” 2018. https://www.superannotate.com/blog/llm-prompting-tricks#:~:text=26%20prompting%20techniques%201%201.%20No%20need%20to,8%208.%20Format%20your%20prompt%20...%20More%20items.

Zeng, Zexian, Sasa Espino, Ankita Roy, Xiaoyu Li, Seema A Khan, Susan E Clare, Xia Jiang, Richard E Neapolitan, and Yuan Luo. “Using Natural Language Processing and Machine Learning to Identify Breast Cancer Local Recurrence.” *BMC Bioinformatics* 19, no. S17 (December 1, 2018).

https://doi.org/10.1186/s12859-018-2466-x.

Zeng, Zexian, Sasa Espino, Ankita Roy, Xiaoyu Li, Seema A Khan, Susan E Clare, Xia Jiang, Richard E Neapolitan, and Yuan Luo. “Using Natural Language Processing and Machine Learning to Identify Breast Cancer Local Recurrence.” BMC Bioinformatics 19, no. S17 (December 1, 2018). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6309052/

Zhang, Shengyu, Dong Linfeng, Xiaoya Li, Sen Zhang, Xiaofei, Shuhe Wang, Jiwei Li, et al. “Instruction Tuning for Large Language Models: A Survey.” Accessed June 4, 2024. https://arxiv.org/pdf/2308.10792.

Zheng, Chujie, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. “LARGE LANGUAGE MODELS ARE NOT ROBUST MULTIPLE CHOICE SELECTORS,” n.d. https://openreview.net/pdf?id=shr9PXz7T0.