Conversational Model for Music Therapy Treatment for Preterm Children

Marine Buliard

November 25, 2023

This thesis is submitted in partial fulfillment of the requirements for a degree of Master in Systems and Computing Engineering.

Assessor:

Prof. Ruben Francisco MANRIQUE — Universidad de los Andes, Colombia

Abstract

This study aims at building and evaluating a chatbot fine-tuned for music-therapy for neonatal infants. Indeed, numerous sources show the positive impact of music-therapy on patients'health and well-being, despite this speciality being underemployed in some areas of the world, and with the development of conversational models and Large Language Models, the potential in medical field is immense and has already sparked interest among the scientific community. However, due to the sensitivity of the data and the potential risks for the patients, medical chatbots must be used with extreme caution to make sure the answers they provide are correct.

To tackle the issue, this research will focus on fine-tuning a general language model and its parameters. It will be done thanks to a RAG architecture, which uses an external database filled with music-therapy papers, to validate the accuracy of the answer. The goal is to compare the quality of the answers from our chatbot with other available models and to determine the impact of various features such as prompting and treatment strategies. We can then use these findings to incrementally improve the model enough to generate accurate and useful answers, thus assisting music therapists in real-life situations.

Contents

1 Introduction

1.1 Context

Conversational systems are meant to allow human users to interact with a machine in the most natural way possible. They are built on LLMs, linguistic models based on millions or even billions of parameters and represent a revolution when compared to classic AI models. They can achieve various type of general-purpose language understanding and generation exercises, including reasoning and connection between seemingly unrelated elements[3] across a wide range of areas from entertainment or education to research and finance.

Recently, there has been a huge augmentation in the use of such tools, particularly with the release of ChatGPT to everyone. This generated a gain of interest from both the scientific community and the general public. However, the most widely used tools are generalist and as such, are not really suitable for specific uses such as medicine.

Moreover, due to the sensitive nature of medical data and stakes, the use of LLMs in this field must be extremely controlled and prudent. Indeed, there can be several risks associated with this technology. First, the confidentiality of medical data must be preserved so that no one can find the medical history of a patient by using the model, whether intentionally or not. Second, a language model is the reflection of the training dataset it was fed and as such, it can contain and even accentuates bias and stereotypes based on gender, ethnicity or socioeconomic backgrounds for example. Finally, these models can be subject to hallucinations, a phenomenon where the given answer is completely wrong despite appearing coherent and believable.

1.2 Problem statement

Recent studies have shown that conversational models can yield significant results in a specific area of medicine if they are fine-tuned properly. However, to the best of our knowledge, none focused specifically on music therapy. That's why we propose to explore how a general conversational model can be adapted to fit best the need of the practitioners. In order to achieve this, we will test multiple conversational models and analyze their performance.

1.3 Justification

The positive impacts of music on people are getting an increasing recognition all over the world. This appreciation extends to all formats of music: listening or practicing and through personal initiative, wellness application or therapeutic settings offered by a professional. Among the main outcomes, we can cite an increase in socialization, an improvement of mental health and of some physical and physiological markers[4].

This tendency is also reflected in the more specific case of music-therapy (hence dispensed by specialists). It attracted more and more attention form the scientific community, as shown by the globally linear increase in paper in the last twenty years[6]. While most papers focus on a specific use of music therapy, similarities between the effects on patients are observed: reduction of stress and anxiety, reduction of the perception of pain (leading to a reduction in administrated painkillers)[13][2].

These results can also be seen in the case of neonatal music-therapy. Indeed, a Neonatal Intensive Care Unit (NICU) can be a distressing environment for babies because of the foreign sounds, harsh lights and scarcity of contact between the parents and the infant. This can generate anxiety and a feeling of helplessness in both the caregivers and the child. Bieleninik et al., while reviewing randomized controlled trials (RCTs) on the matter, evidenced an improvement in feeding skills, a reduction of heart rate, respiratory rate and both maternal and newborn stress[1] thanks to music-therapy.

1.4 Objectives

1.4.1 General objective

The general objective is to implement and evaluate a music-therapy specific chat-bot.

1.4.2 Specific objectives

- Implement a vectorial database for embeddings and a training dataset built with the support of experts in music therapy.
- Implement multiple back-end models based on various conversational models and

prompting strategies, especially trying to get both commercial and open-source models.

- Compare the performance of different conversational model and prompting strategy combinations on the training set.
- Analyze the impact of the different combinations of LLMs and prompting strategies to provide insights regarding which one is the most suitable.

2 Methodology

2.1 Methodology framework

The objective of this study is to implement a domain-specific chat-bot using a RAG architecture to accurately answer a question about music-therapy in a neonatal context, using sources from a specific dataset to support the answers and then evaluate the performance of said chatbot.

This is because despite the growing performances of general conversational models and the attention it attracted in recent years for tasks such as decision-making or summarizing, the medical-field is too sensible to use unrefined models, and a fortiori in neonatal care through music-therapy.

Some SOTA models tried to tackle this issue using the RAG architecture. A good example of that is ChatDoctor[7], a chatbot based on the open model LLaMa to which is adjoined external databases, to be used as the foundation of the answers. Our study uses a RAG architecture with a focus on music-therapy for neonatal infants and based on a general LLM (which model is the best for the purpose of this study is not yet fully determined, to this day it is based on GPT-3.5 but this can be subject to later modification). The offline dataset associated with our model is a compilation of all papers about music-therapy in neonatal care since the 1970s, presented as vectors that are called embeddings.

Our model propose to retrieve the chunks of this offline dataset that are relevant with respect to the query and feed them to the model which will use them as context, along with the question and a prompt designed to maximise the accuracy of the answer.

The quality of the obtained responses, and as a result the merit of the model, will be assessed using two different methods: a BERTscore, an automatic evaluation metric based on semantic similarity[20], and a human evaluation by a pool of experts that will rate several factors such as the relevance of the information, the understandability of the answer or its accuracy for a benchmark specific to music-therapy.

2.2 Experiment workflow: calendar overview

The experimental project will consists of the following stages:

1. **Embeddings construction:** The offline database will come from research papers published in the last 50 years about music-therapy for neonatal care and takes the form of vectors. While this part is mainly straightforward, the format of some data must be changed or ignored. Indeed, GPT-3.5 does not support tabular data and images, both of which are very present in scientific papers, hence requiring a special attention.

2. **Prompt design:** Prompts are necessary to make the model understand its task and how to do it best. It should include the instructions on how to treat the different kinds of questions and any useful information for the model, for instance the tone to use or the brevity of the answer. Prompts can have a major impact on the results obtained with a chatbot, to the point it is considered one of the most important element to fine-tune[19][11].

3. **Ranking system for context retrieval:** Multiple chunks from the embeddings can have a high similarity with a given query. However, it does not assuredly means all these chunks are actually useful to answer the question: sometimes, the similarity can be pure coincidence. Moreover, even when several chunks are relevant to answer the query, all of them cannot necessarily fit into the context and the system must be able to chose the fragments which will lead to generating the best answers. This type of improvements is part of the postprocessing.

4. **Version with an open model:** The current version of our chatbot uses OpenAI's ChatGPT-3.5, but this choice have consequences on various aspects of the project. The main one is the redistribution of the chatbot: with OpenAI, it is not possible to download and distribute the fine-tuned model. This is why it would be really interesting to transition to an open model such as LLaMa[17], or its newer version LLaMa2[16].

5. **Refinement of the user interface:** The front-end is build in React and communicate with the back-end through an API and json files. To improve the user experience, it is possible to make the interface more attractive and easily-accessible.

6. **Evaluation of the model:** The evaluation, done according to the BERTscore and expert human evaluators, is the step that will truly allow us to determine the validity and usability of the constructed model. It will be compared with at least one general LLM such as GPT-3.5 and a domain-specific model such as BioGPT.

7. **Feedback:** From the results obtained up to that point, we will adapt the models to get better results and fit the user's needs.

Experimental calendar							
Week	Embedding construction	Ranking system	Prompt design	Adaptation to an open model	User interface	Evaluation	Feedback
1	X						
2	X	X					
3		X	X				
4		X	X	X			
5				X			
6				X			
7					X	X	
8						X	X
9						X	X
10	X	X					
11		X	X				
12			X	X			
13				X	X		
14						X	
15						X	X
16						X	X

Figure 2: Estimated calendar overview

2.3 Progress

A prototype is already available for this project. It includes a simple but functional interface and can use a small embedding file generated from papers about to music-therapy to support its answer and print the references it used. In addition, it says "I don't know" when supporting elements are not found in the embeddings in order to reduce the risks of hallucinations.

More precisely, the front-end is coded in React and uses a json API to communicate with the back-end which is coded in Python. The back-end uses an OpenAI API to be able to create the answers from the context. After retrieving the cosine similarity between the question and the embeddings, the best ones are given back to GPT-3.5 with the question and the prompt "Can this fragment really help to answer the question?"

3 Related work

3.1 Referential framework

The aim of this referential framework is to provide an overview of the key concepts required for a proper understanding of the project's functioning.

3.1.1 Large Language Models

Large Language Models, or LLMs, are neural networks used to solve various natural language processing tasks. Their training dataset is extremely wide: it is composed of millions, billions or sometimes even trillions of parameters. These parameters often come from texts on internet such as Wikipedia articles, forums, or specialized databases. Training of the LLMs can be self-supervised (no annotations or human labeling beforehand) and semi-supervised (a small portion of the training data is labeled by human while the rest of the dataset remains unlabeled). The performance of an language model is often correlated with the size of its training dataset.

3.1.2 Generative versus discriminative LLMs

There are two main kinds of models: generative and discriminative ones. Generative models generate new data following the training set patterns while discriminative models aim at distinguishing various categories present within the data.

The main generative model is GPT. It can perform various types of tasks including, in a medical context, taking notes, bedside decision support or chatbot with the patients for example[18]. It is also the base of several specialized models such as GatorTronGPT[12], Radiology-GPT[9] or BioGPT[10].

On the other hand, BERT is a discriminative model and so it can be used for tasks such as sequence classification or labeling[10]. While it is considered as a good discriminative model, it does not yield very good results in generation tasks.

Finally, there are hybrid models that combine the generative and discriminative architecture such as PaLM, and subsequently MedPaLM and MedPaLM2[15], which generate

sequences of text using discrimination for better predictions.

3.1.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is the process of fetching data from an external database to reduce the risk of hallucination. Even when the information was seen during training, it can be misused by the model[8] because it is subdued by the quantity of data and it gets lost in all of the other documents that can be contradictory.

RAG architecture adds an extra-step to the classic chatbot sequence: when the user asks a question, it is not transmitted immediately to the general chatbot because relevant information is retrieved before from an external database, whether offline (like it is the case for our model), online or both (Wikipedia and an offline medical database for ChatDoctor[7]).

This presents several major benefits. Firstly, it reduces drastically the risk of hallucination by requiring the answer to be based on the content of the external database. Secondly, it gives the model access to up-to-date information since the last knowledge GPT-3.5 was in January 2022. Finally, it makes it possible for our model to says "I don't know" when the information is not available in the database.

3.1.4 Embeddings

Embeddings are very useful when fine-tuning a generalist LLM. Indeed, when we want to add domain-specific knowledge to a model, recompiling would require a lot of time and resources. That's why embeddings are used instead. They are a way of storing information as vectors. Then, the embedding can be passed directly to the model as context. To do so, we compute similarity between the text contained in the embeddings and the question. The more important the similarity, the more useful the text in order to answer the question. Then, we can pass the most relevant texts fragment to the model as context, which gives the model the necessary information to answer the user's question.

3.1.5 Prompt engineering

Prompt engineering is the process of refining input queries and prompt to obtain the best possible output from a LLM for a given task. It leads to an increase in the model's performance[19] -especially when dealing with complex reasoning[8]- while being computationally inexpensive since the parameters of the model are not changed. The main prompting strategies are:

- Zero-shot prompting, where a single completion step is performed with no annotation before or after.

- Few-shot prompting, it consists in providing a few example inputs and outputs to the LLM so that it can see how to solve efficiently the question and detect patterns in the examples.

- Chain-of-Thought (CoT) requires the model to explain its reasoning step-by-step. It takes more time for the model to generate an answer, but it helps avoid reasoning mistakes. It can be combined with few-shot (examples of what the reasoning should look like) or zero-shot prompting ("Let's think step-by-step" is the typical CoT prompt).

Various prompting strategies can be used together to get better results, such as few-shot prompting and CoT.

3.1.6 Preprocessing and postprocessing

Preprocessing and postprocessing are two important steps of fine-tuning.

Preprocessing is the treatment of the data before it is used by the model. It can include removing illegal characters or verifying the user is not trying to hijack the model by giving instructions that are contradictory with the ones we provided. For example, it can be someone trying to indicate to the model the answer should be written with a certain style or set a length limit.

Postprocessing is the step following the cosine similarity computation. When the embeddings are really big, the query vector can be similar to one chunk by accident, so we

want to make sure we will not use inadvertently a fragment which has nothing to do with the question. To do so, we can asks the LLM if the chunk and the query are really related for the top-chunks.

3.1.7 Metrics

The evaluation of a model is done according to multiple factors. One of them is the type of questions because they will not require the same metrics. Indeed, in the case of MCQ, where the answer can be labeled as true or false, one can use the following metrics:

- Accuracy (correct predictions among all)
- Precision (true positive among all predicted positive)
- Recall, also called sensitivity (proportion of properly predicted positive among actual positive)
- F1-score (metrics mixing recall and precision but favoring a balance between both over one with a very good score and the other very bad)

Whereas for open or long-form questions, it is not possible to use such Manichean metrics, because there is not a single correct response. In this case, the evaluation can be achieved with:

- Pairwise comparisons: for all not strictly numerical criteria, it can be hard to attribute a mark and the same mark will not necessarily mean the same thing for different people anyway. This is why it is more efficient to do pairwise ranking, where the answer are compared in two to determine the best[15]. Pairwise comparisons can be carried out according to all the criteria listed below.
- Uncertainty: it allows the model to determine how certain of its response it is, and if the value is below a determined threshold, withhold the answer to avoid hallucination[14].
- Clarity of the answer: no matter how good the content of the answer, if it is not clear to the user, then it will be useless.

- Helpfullness of the answer: To be really beneficial, an answer must not only expose correct facts and be clearly stated, but also bring added value: if it is too generalist or uselessly too long, it will not help the user and can even be counterproductive.

3.2 State of the art

3.2.1 Datasets

The datasets must be chosen very carefully since they are the starting point of both training and evaluation of the models. Most datasets are completely and partially in English, with regularly Chinese datasets[5] [18]. The most popular datasets for training and fine-tuning are built from real-world data, such as PubMed, online medical blogs or forums, and medical records. However, it is also possible to synthesize training datasets with a specialized LLM: GatorTronS was trained with GatorTronGPT[12] and while it had less parameters than its counterpart trained with real-world data, it obtained better results.

Some models use external online databases and knowledge graph[11] in addition to training, allowing them to have up-to-date information. This is the case for ChatDoctor[7]. These external datasets can also be formated as knowledge graphs (KG) with triplets (subject, relationship, object)[18].

Evaluation datasets are numerous and can have various format. There are multiple-choice question answering like PubMedQA[12] [10] or MedQA[18] [15], open-ended questions[15], the US medical exam USMLE[12], patient consultations and diagnostic analysis of medical records. Since there are numerous datasets, MultiMedQA was created [15] as the mix of six existing datasets and the addition of a new one, HealthSearchQA, which contains the most commonly searched health question on internet.

3.2.2 Medical fine-tuned model performance

While general-purpose LLMs such as GPT or BERT are good, the proportion of medical data in their training is not enough to obtained satisfying result in such a sensible area as medicine. To overcome this, we can use transfer-learning (fine-tuning with in-domain data).

This method allows us to keep using the general vocabulary but improve performances on specific tasks.

Fine-tuning can take different forms and greatly improve performance and allows to use various format documents for fine-tuning: BioGPT[10] use knowledge graphs and convert it to natural language to get as much information as possible. [11] shows it is a good strategy to use knowledge graphs.

3.2.3 Prompting strategies

Similarly to fine-tuning, prompt engineering can improve greatly a model's performance but without modifying parameters and weights. The advantage of this strategy is how computationally inexpensive it is. Simple prompts ("Doctor:" or "Doctor may think") are enough to substantially improve LLM's holistic thinking capabilities[19] and perform even better than domain-specific and high-level instruction prompts ("Let's reason like a medical expert:" or "Doctor: Let's go through the process of elimination to determine the possible causes. My hypothesis is"). However, while simple prompts improve the performance, some more complicated prompts worsen the results. This means that a bad prompt can be truly prejudicial. We can also note two developing strategies for prompting: hard prompting, when the prompt is manually-designed[10], and automatic prompting where the model finds on its own the best prompt. Finally, prompting strategies can go through several steps. For holistically thought (HoT)[19], we can distinguish three stages:

- Diffused thinking: the model generates various possible answer from the symptoms.
- Focused thinking: the model generates the patient's file (independently from the question) with as much precision as possible
- Response generation: the model combines the output of the first two steps to create the answers.

This method gives very good results but sometimes there is a deficiency in the understandability.

3.2.4 Limitations

All papers mentioned numerous risks associated with the use of LLMs in medical environments. The first issue is patient privacy: the anonymization cannot always remove all of the protected health information (PHI) from the datasets, leading to potential exposure of secret information.

Secondly, the training datasets can transmit bias to the model. The bias can concern a part of the population (gender, ethnicity, socioeconomic status) or a particular practice, like shown in [5], where the model is adapted to western medicine and thus perform poorly on Traditional Korean medicine questions. This can be partially overcome by fine-tuning, allowing the model to get more of the underrepresented data.

Finally, hallucinations represent a major threat to the use of LLMs in medical settings: because of the probabilistic nature of text generation, LLMs can sometimes generate seemingly accurate answers but in reality, they are based on no concrete data. Moreover, a model can also make logic inconsistencies, reasoning errors, incoherences or factual errors, all of which can lead to patient's harm.

3.2.5 State-of-the-art conclusion

LLMs performances are extremely dependent on the architecture, tuning and prompting used. In consequence, it is important to choose carefully these elements depending on the tasks we want the model to perform. Given the sensitive nature of medical domain, a particular attention must be given to the selection of the dataset and the anonymization of the data. To limit the risks of hallucination, errors, over-generalization and bias, it is also primordial to evaluate thoroughly and honestly the results. Finally, it is important to remember that no matter what, these LLMs are tools and as such, are meant to be used together with medical experts and not replace them.

Acronyms

BERT: Bidirectional Encoder Representations from Transformers

CoT: Chain-of-Thought

GPT: Generative Pre-trained Transformer

HoT: Holistically Thought

LLM: Large Language Model

MLM: Masked Language Modeling

NLP: Natural Language Processing

PaLM: Predictive and Learning Models

RAG: Retrieval Augmented Generation

RNN: Recurrent Neural Network

References

[1] Åucja Bieleninik et al. "Potential Psychological and Biological Mechanisms Underlying the Effectiveness of Neonatal Music Therapy during Kangaroo Mother Care for Preterm Infants and Their Parents". In: *International Journal of Environmental Research and Public Health* (2021). DOI: 10.3390/ijerph18168557.

[2] Joke Bradt, Cheryl Dileo, and Denise Grocke. "Music interventions for mechanically ventilated patients". In: *Cochrane Database of Systematic Reviews* (2010). DOI: 10.1002/14651858.cd006902.pub2.

[3] Mark Connor and Michael O'Neill. *Large Language Models in Sport Science Medicine: Opportunities, Risks and Considerations*. 2023. arXiv: 2305.03851 [cs.CL].

[4] Saoirse Finn Fancourt Daisy. "What is the evidence on the role of the arts in improving health and well-being? A scoping review". In: *Health Evidence Network synthesis* 67 (Nov. 2019). DOI: 10.1093/bib/bbac409. URL: https://www.who.int/europe/publications/i/item/9789289054553.

[5] Dongyeop Jang and Chang-Eop Kim. *Exploring the Potential of Large Language models in Traditional Korean Medicine: A Foundation Model Approach to Culturally-Adapted Healthcare*. 2023. arXiv: 2303.17807 [cs.CL].

[6] Kailimi Li, Linman Weng, and Xueqiang Wang. "The State of Music Therapy Studies in the Past 20 Years: A Bibliometric Analysis". In: *Frontiers in Psychology* (2021). DOI: 10.3389/fpsyg.2021.697726.

[7] Yunxiang Li et al. *ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge*. 2023. arXiv: 2303.14070 [cs.CL].

[8] Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. *Can large language models reason about medical questions?* 2023. arXiv: 2207.08143 [cs.CL].

[9] Zhengliang Liu et al. *Radiology-GPT: A Large Language Model for Radiology*. 2023. arXiv: 2306.08666 [cs.CL].

[10] Renqian Luo et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining". In: *Briefings in Bioinformatics* 23.6 (Sept. 2022). ISSN: 1477-4054. DOI: 10.1093/bib/bbac409. URL: http://dx.doi.org/10.1093/bib/bbac409.

[11] Shirui Pan et al. *Unifying Large Language Models and Knowledge Graphs: A Roadmap*. 2023. arXiv: 2306.08302 [cs.CL].

[12] Cheng Peng et al. *A Study of Generative Large Language Model for Medical Research and Healthcare*. 2023. arXiv: 2305.13523 [cs.CL].

[13] D. Rudin et al. "Music in the endoscopy suite: a meta-analysis of randomized controlled studies". In: *Endoscopy* (2007). DOI: 10.1055/s-2007-966362.

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

[15] Karan Singhal et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. 2023. arXiv: 2305.09617 [cs.CL].

[16] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[17] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[18] Guangyu Wang et al. *ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation*. 2023. arXiv: 2306.09968 [cs.CL].

[19] Yixuan Weng et al. *Large Language Models Need Holistically Thought in Medical Conversational QA*. 2023. arXiv: 2305.05410 [cs.CL].

[20] Tianyi Zhang et al. *BERTScore: Evaluating Text Generation with BERT*. 2020. arXiv: 1904.09675 [cs.CL].