

Robotic AssistEMT: An EMT Chatbot

Stanford CS224N Custom Project

Aanika Atluri
Department of Computer Science
Stanford University
aatluri@stanford.edu
Examples of more authors

Sarah Barragan
Department of Computer Science
Stanford University
sabarrag@stanford.edu

Anusheh Chaudry
Department of Computer Science
Stanford University
anusheh@stanford.edu

Abstract

We present the development of an Emergency Medical Technician (EMT) chatbot, AssistEMT, designed to serve as a conversational assistant for emergency medical examinations and to interface with robots capable of performing necessary physical assessments. Despite the critical role of AI in various domains, its integration into emergency medicine remains underexplored due to the highly uncertain and critical nature of the field. We aim to address this gap by developing an empathetic EMT chatbot capable of accurately assessing patients in accordance with the National Registry of Emergency Medical Technicians (NREMT) guidelines and synthesizing information to provide diagnoses and transport decisions. Our contributions include the creation of two specialized datasets centered around EMT-specific medical knowledge. These datasets were then used to fine-tune a GPT-3.5 model ChatGPT (2024b) and to implement Retrieval-Augmented Generation (RAG) Lewis et al. (2021) with GPT-4 ChatGPT (2024a). Our major findings reveal that GPT-4's extensive context and inference capabilities make it the most well-suited for the tasks of an EMT chatbot. The incorporation of RAG further augments the chatbot's functionality by enabling real-time updates of medical knowledge, ensuring the chatbot remains updated with the latest medical discoveries. Our prompt engineering ablation studies have also resulted in a chatbot that operates with an empathetic demeanor akin to a human assistant, enhancing user experience and trust.

1 Key Information to include

- Mentor: Kenneth Salisbury, Sunny Singh (External Mentors) & Rashon Poole
- External Collaborators (if you have any): Anjana Balachandar & Varsha Naga (Trained EMTs)
- Sharing project: No
- Group Contributions: Anusheh worked on the GPT-3.5 fine tuned model and creating our own datasets, Aanika worked on the GPT-4 with prompt engineering model as well as ablation studies and automating dataset creation, and Sarah worked on the GPT-4 with RAG model as well contributing to the creation of the datasets. We all contributed to the evaluation of our models and to the final report.

2 Introduction

Emergency Medical Technicians (EMT) face critical challenges in being able to rapidly assess and respond to patient needs during emergencies. However there has been a recent emergence of AI chatbots transforming healthcare by processing vast data and detecting patterns beyond human capability, potentially alleviating the healthcare system’s load and allowing medical professionals to focus on complex tasks. However, the quality of these chatbots is limited by their training data. While they can provide general health information and manage routine tasks, they struggle with queries requiring deep medical knowledge, critical thinking, and years of clinical experience. Additionally, AI lacks the empathy and intuition of human medical professionals. Thus, the integration of chatbots with medical professionals has evolved into a collaborative approach, where chatbots supplement physicians in complex decision-making and empathetic patient care.

Our project aims to bridge this gap through developing a Robotic EMT chatbot interface that leverages NLP techniques to conduct medical interviews and assessments assisting an EMT. To address persisting challenges in the field, we present the development of an EMT (Emergency Medical Technician) chatbot using fine-tuning and Retrieval-Augmented Generation (RAG) with Openai’s GPT-3.5-turbo and GPT-4 models along with advanced prompt engineering techniques. This conversational empathetic EMT chatbot is capable of conducting medical examinations and effectively assessing and diagnosing patients at the scene. It is designed to function compatibly with an accompanying robot assistant for tasks that require physical manipulation. In addition to the models, we have also created two datasets specifically for emergency medicine that offer comprehensive information regarding emergency medical procedures and terminology. Our approach builds on existing work, such as ClinicalBERT and potential of ChatGPT within medicine, with an aim to overcome their additional limitations with open-ended dialogue and integrating real-time updates.

Our research has yielded significant results demonstrating the viability and effectiveness of AI-driven EMT assistants. The fine-tuned GPT-3.5 model as well as the GPT-4 models both with and without RAG showcased enhanced performance in EMT domain knowledge examinations and scenarios. GPT-4’s superior context understanding and inference capabilities led us to determine it as the most effective model for emergency medical assessments, and the integration of RAG enabled the chatbot to stay up to date with current events and the latest medical knowledge, ensuring real-time accuracy and ethical compliance. Our incorporation of chain-of-thought and few-shot prompting also resulted in the development of an empathetic and user-friendly chatbot, improving user trust and experience. These findings highlight the potential for AI to significantly augment the capabilities of emergency medical services and offer a scalable solution to current healthcare challenges in emergency medicine.

3 Related Work

Previously to address limitations of AI applications in medicine, researchers have explored pre-training and fine-tuning large language models (LLMs). One notable example is ClinicalBERT Wang et al. (2023), which tailors the BERT model to clinical corpora. Given a patient’s clinical notes, ClinicalBERT predicts the risk of readmission within 30 days, which is valuable in high-pressure environments where clinicians must make swift decisions. ClinicalBERT highlights the importance of fine-tuning models on specific medical data to capture qualitative relationships among clinical concepts. Despite its success, ClinicalBERT struggles with open-ended dialogue, making it less ideal for medical chatbots that need deep comprehension of meaning and intent. Gao et al. (2021)

Building on the idea of model fine-tuning, recent advancements suggest that ChatGPT could significantly impact medicine. As a professionally trained medical chatbot, ChatGPT could operate faster than existing models, draw on larger databases, reduce errors, and improve doctor performance. Chow et al. (2023) Prompt engineering has been employed to guide ChatGPT towards generating more accurate and empathetic responses, addressing some limitations but leaving room for improvement. Issues like outdated databases and potential inaccuracies in web-sourced medical information raise concerns about reliability of diagnosis and treatment recommendations, as well as introduce hallucinations or misinformation.

Integrating RAG represents a significant advancement in enhancing reliability, precision of diagnoses and treatment suggestions, since it incorporates domain-specific knowledge. Studies show RAG has

improved model accuracy from 80.1% to 91.4% and reduces human error by automating the retrieval of pertinent health records and leveraging advanced diagnostic algorithms. MyScale It provides real-time insights that are crucial for informed decisions, ensuring healthcare professionals have access to current, relevant information.

Our research is inspired by these studies to enhance AI capability to conduct medical interviews and assessments. Our GPT-3.5 implementation, fine-tuned on EMT-specific medical datasets, improves task-specific accuracy. Then, GPT-4 with prompt engineering guides the model to generate contextually appropriate responses, mitigating issues like hallucinations. Lastly, GPT-4 with RAG ensures real-time access to accurate and relevant medical information. These combined approaches enhance the chatbot's precision, reliability, and decision-making, making it a robust tool for emergency medical scenarios.

4 Approach

As our baselines, we compare results against OpenBioLLMAnkit Pal (2024), a model that builds upon the foundation of Meta-Llama-3 by fine-tuning on a vast corpus of biomedical data, GPT-3.5, and GPT-4.

GPT-3.5 with Fine Tuning: Our primary model was fine-tuned on PubMed Jin et al. (2019), a question and answer dataset created from the National Association of State EMS Officials (NASEMSO) Official Guidelines, and a question and answer dataset created from the "Emergency Care and Transportation of the Sick and Injured" textbook Pollak et al. (2021), studied by EMTs in training. We first fine-tuned the model on just the PubMed dataset, which consisted of 4,792,460 trained tokens, 1 Epoch, a batch size of 10, and LR multiplier of 2. It took about 2 hours and 45 minutes to train and had a loss of 0.2905. Next, we finetuned that model on the NREMT guideline dataset, which consisted of 169,284 trained tokens, 3 Epochs, batch size of 2, and LR multiplier of 2. It took about 45 minutes to train and had a loss of 0.4196. Lastly, we finetuned that model on the textbook dataset, which consisted of 246,354 trained tokens, 3 Epochs, batch size of 3, and LR multiplier of 2. It took about an hour to train and had a loss of 0.9968.

GPT 4 with prompt engineering: While the fine-tuned GPT-3.5 models have shown substantial promise in handling EMT scenarios, the dynamic and critical nature of EMT interactions demanded a more sophisticated and consistently reliable solution. This led us to explore the performance of GPT-4, a more advanced model with enhanced contextual understanding and conversational capabilities. GPT-4 is also trained on more recent data, ensuring its generations contain more updated information. As a baseline, we tested the performance of GPT-4 as an EMT assistant, utilizing prompt engineering to ensure the chatbot adheres to the EMT assessment structure. Our prompt features chain-of-thought prompting Wei et al. (2023), which consists of decomposing a complicated or long-horizon task into a sequence of intermediate reasoning steps to improve multi-step reasoning and inference capabilities. During their assessments, EMTs are responsible for collecting multiple data points regarding the patient's condition, as well as synthesizing information to diagnose and treat the patient. Therefore, chain-of-thought prompting is a valuable tool for prompting the model to consider past information and synthesize data to produce complex inferences. Additionally, we also integrated few-shot prompting to assist the model in properly ordering and asking procedural questions for a desired output format without fine-tuning (fine-tuning is not yet available for GPT-4) Thus, to familiarize the model with the procedural structure of an EMT examination, we provided a full patient-EMT ground-truth interaction within the system message.

EMT interactions must include empathy and a comforting demeanor in the chatbot's responses so that they may interact with patients in a calming and respectful manner. Although GPT-4 is equipped with advanced question-answering capabilities, the nature of EMT assessments is especially rigid, following specifically ordered protocol and criteria in order to accurately and effectively assess the patient's condition. Thus, to incorporate empathy into the model's responses, we included additional instructions in the system message that specify the demeanor the chatbot should assume when interacting with a user. It is also important to consider the importance of incorporating reliable, updated information when assessing patients.

GPT-4 with RAG: In order to improve the performance of GPT-4 and ensure it returns updated and accurate information, we developed a RAG system. We initialized the GPT-4-turbo model as the core LLM for generating responses. To facilitate efficient retrieval, Chroma vector stores were initialized using OpenAI embeddings, with data persisted to a specific directory. We loaded EMT-specific text documents stored in a directory, including the National Registry of Emergency Medical Technicians Emergency Medical Responder Psychomotor Examination, the NASEMSO guidelines described in the dataset section, and a general EMT Task document. These documents were split into chunks and stored in a vector store dedicated to text documents. A text-specific retriever was then created to fetch relevant document chunks based on user queries, enhancing the chatbot’s ability to provide precise and context-specific information based on established EMT guidelines.

In addition to text document retrieval, we integrated a Google Custom Search Engine in our retrieval process. This engine was configured to search reputable medical websites to minimize the risk of misinformation and hallucinations and integrated into a web research retriever, which handles retrieving relevant web-based information. Furthermore, we loaded a JSON file outlining the Patient Interview Decision Tree for EMTs. Its contents were also stored in a dedicated vector store and a JSON-specific retriever was developed to access the file. To integrate our different retrievers, we developed an ensemble retriever with weighted priorities: 0.5 for text documents, 0.3 for web search, and 0.2 for the JSON file. In combination with the retrieval mechanisms, we implemented a question-answering (QA) chain to guide the chatbot’s responses. This chain encompasses a system prompt that defines instructions for the chatbot to consider both chat history and user input for generating responses. We then defined a prompt template that serves as a guide for generating responses based on a given context, incorporating the system prompt, chat history, and user input.

We defined a question-answer (QA) system prompt using the same instructions as our prompt engineered GPT-4 model. We then implemented a retrieval chain that processes user input by first contextualizing the user input and then making a call to the ensemble retriever which conducts the retrieval of pertinent information from our sources, guided by the contextual cues provided by the conversation history. Finally, we orchestrated the deployment of the RAG system by linking the retrieval chain with the QA system prompt and the GPT-4-turbo model. This final pipeline enables the chatbot to retrieve information, generate contextually relevant responses based on instructions and guidelines provided, and articulate these responses effectively using the language model, thereby facilitating efficient communication and assistance in emergency medical scenarios. We used the Langchain library and documentation however, the code was written by us.

5 Experiments

5.1 Data

When constructing our fine-tuned GPT-3.5 model, we conducted three iterations of fine-tuning to incorporate three sources of data. Our first dataset is the PubMedQA dataset, which consists of 211k question-answer pairs designed for question-answering tasks with binary outcomes. Jin et al. (2019) When fine-tuning GPT-3.5 using the Openai API, the data must be restructured into a question-answer format stored in a JSON-lines format. Since our task extends beyond the scope of binary question answering, we selected the ‘question’ and ‘long answer’ fields to construct our question-answer pairs for fine-tuning.

Many publicly available datasets offer comprehensive knowledge of medical concepts, but they fail to closely examine the nuances and unique scenarios of emergency medicine. Due to the lack of EMT-specific data, we created two additional datasets to further fine-tune the PubMed-GPT-3.5 model. The first dataset consists of information pulled from the National Association of State EMS Officials (NASEMSO) Official Guidelines National Association of State EMS Officials (2022). The NASEMSO guidelines consist of EMS system clinical guidelines, protocols, and operating procedures with detailed instructions on treating a variety of medical emergencies. Using ChatGPT’s GPT-3.5 model, we generated 20 question-answer pairs for every 10 pages of the document, yielding a final dataset of 1056 question-answer pairs. This dataset supplements the general medical knowledge from PubMed with abbreviations and protocols specific to emergency medicine scenarios, with the goal of creating a more specialized chatbot for EMT scenarios.

Our second dataset was constructed from the “Emergency Care and Transportation of the Sick and Injured” textbook Pollak et al. (2021), which is used for teaching EMT certification courses. This

textbook offers a comprehensive, in-depth explanation of all necessary concepts an EMT must learn in order to pass the final certification exam. We created a script powered by gpt-3.5-turbo to automate question generation for 2,000 pages. We set a generation of 10 questions per 4,000 tokens. This process yielded a 1562-pair question-answer dataset of specific EMT course material, including practice scenarios and examples of EMT decision-making workflow.

In addition to fine-tuning, we utilized external information sources for the integration of RAG with GPT-4. Our RAG information sources consisted of two categories: a custom web-search engine (up to date information) and downloaded files (static databases). Sites searched include the World Health OrganizationWorld Health Organization, the Center for Disease Control and PreventionCenters for Disease Control and Prevention, and Stanford Emergency MedicineStanford Medicine.

For downloaded files, we included an EMT protocol decision tree, which details the hierarchical interview and decision-making process EMTs must adhere to when examining a patient. We converted it into a JSON file to preserve its dependencies for parsing. Additionally, we chose to include the NASEMSO guidelines and the EMT Criteria Assessment, a rubric scored out of 35 when assessing EMT performance during clinical assessments. These sources provide domain specific knowledge and additional context to the model about the weighted importance of different steps in the EMT process.

5.2 Evaluation method

The first metric we used as a quantitative measure to assess overall comprehension of the model on relevant EMT knowledge was the model’s performance on the NREMT final examination. This exam contains two parts: a written multiple choice test and a clinical test, both of which require a score of at least 70% in order to pass. We pulled a written examination from the EMT textbook, containing 120 multiple choice questions, and ran all questions through our baseline models and our implemented models. To evaluate our models on the clinical test, we obtained three scenarios that have been validated by the NREMT and have been used as video examples for training. We ran each scenario through our GPT-3.5-finetuned model, GPT-4 prompt engineered model, and GPT-4-RAG model, simulating a patient response based on the responses given in the videos. Trained EMTs then evaluated these interactions using the official rubric for clinical examinations. The second metric we used was a qualitative ‘Empathy Score’ to understand how empathetic each chatbot was in its responses as compared to a human EMT. We had the trained EMTs assisting us assign each model an empathy score from 1 to 5, with 1 containing no empathy and 5 achieving or surpassing the level of a human EMT. Our last metric we used was qualitative, where the trained EMTs assisting us wrote a brief impression of each model and how viable it would be in a real life setting.

5.3 Experimental details

At first, we envisioned a hybrid approach, where we would fine-tune ClinicalBERT on the MedQuad dataset, which contains question-answer pairings of diverse medical knowledge, and pair the output with API calls to GPT-4 to generate coherent and contextualized responses. However, we decided to pivot away from this method since we found that feeding the output of ClinicalBert as the input of GPT-4 ended up limiting the effectiveness of GPT-4. We believe this is because GPT-4 requires properly aligned and highly specialized outputs in order to perform well within this system.

We experimented with fine-tuning GPT-3.5-Turbo on a varied combination of datasets to see which it would perform the best under. We fine-tuned it using just the PubMed dataset, just the NASEMSO guidelines, the PubMed and NASEMSO guidelines combined, and the PubMed, NASEMSO guidelines, and EMT textbook combined. We found that the GPT-3.5 model trained on all three datasets performed the best on the written examination as compared to the others and was the most consistently coherent. Other models, like the one trained on just the PubMed dataset, would occasionally bring up context that was unrelated to the input.

Ablation Studies: To construct the most effective system message to improve the performance of our models, we conducted ablation studies to understand the individual impacts of our different prompting strategies and components. The first experimentation consisted of inserting the EMT Criteria Assessment rubric, including the point values associated with each action, into the prompt. During our initial experimentation with GPT-3.5 fine-tuned, we discovered that the model demonstrated difficulty in adhering to the order of protocol in the EMT assessment process and inserted the rubric

into the prompt as an attempt to mitigate this confusion. We also modified the system message to understand how to ensure the chatbot interacted with the user in a kind, patient, and empathetic manner. This involved including keywords such as “empathetic” and “kind” into the description of the chatbot’s persona, as well as an additional reminder to be empathetic at the end of the prompt. We attempted to modify the temperature setting of the model to find a balance of generating reliably ordered responses while still maintaining communication similar to a human EMT. In addition to these components, we also inserted a ground-truth example conversation between an EMT and a patient to provide a clear outline of how a conversation should look from start to finish. The goal of this modification was to enhance the consistency and reliability of the chatbot’s response sequence while also demonstrating the demeanor the chatbot should assume when speaking with a patient.

RAG Experimentation: To optimize our RAG implementation we created separate vector stores for ensemble retrieval to source information from both static and dynamic databases. To access real-time data, we integrated the Google Search API, limiting searches to the top three results. This decision was based on initial tests showing that results beyond the top three results often included noise and less reliable information, diluting response quality. Additionally, we carefully curated documents relevant to EMT practices, balancing comprehensive coverage with processing efficiency. Limiting the number of documents to three and size of retrieved text chunks to 1000 characters, ensured quick and manageable retrieval for the LLM to generate accurate responses. When determining the appropriate weights for each retrieval mechanism, we found that increasing the weight of web searches and JSON file decision tree led to accurate but overly factual and less empathetic responses. We ultimately prioritized text documents, as they provided the most relevant information grounded in EMT guidelines, ensuring a balance of empathy and procedural accuracy. We adopted a contextualization method for incorporating all chat history to ensure contextual relevance, which was essential for dynamic medical situations, where every detail can impact decision-making. Summarization methods, while memory-efficient, risked omitting critical interaction details between the patient and the EMT. By extending the chat history, we ensured the LLM had full context, supporting better continuity and relevance in the chatbot’s interactions.

5.4 Results

The quiz scores revealed a clear hierarchy in model performance. GPT-4 RAG with prompt engineering scored the highest at 0.883, closely followed by standard GPT-4 at 0.875. Notably, GPT-3.5 with a combination of PubMed, NASEMSO guidelines, EMT textbook, and a prompt achieved a score of 0.775, indicating that prompt engineering and combining multiple data sources improved performance compared to its baseline of 0.667.

Model	Quiz Score
OpenBioLLM	0.716
GPT-3.5	0.766
GPT-4	0.875
GPT-3.5 PubMed	0.667
GPT-3.5 PubMed + Prompt	0.692
GPT-3.5 NREMT	0.716
GPT-3.5 PubMed + NREMT + Prompt	0.775
GPT-3.5 PubMed + NREMT + Textbook	0.800
GPT-3.5 PubMed + NREMT + Textbook + Prompt	0.775
GPT-4 RAG + Prompt	0.883

Table 1: Exam Results for Different Models

The model performance in handling specific medical scenarios (i.e. allergic reaction, stroke and diabetes) varied significantly. GPT-3.5 fine-tuned averaged a score of 0.43, demonstrating limited capability in accurately addressing medical scenarios. In contrast, GPT-4 + RAG with prompt engineering scored substantially higher with an average of 0.66 as it showed strong performance in the allergic reaction scenario but fell short in the stroke scenario. GPT-4 with prompt engineering achieved the highest score of 0.80, excelling in allergy scenarios and performing moderately in stroke and diabetes scenarios.

The results aligned with our initial expectations, where GPT-4 models, with advanced techniques like RAG and prompt engineering, outperformed their predecessors. However, contrary to our hypothesis, the RAG model did not outperform the prompt-engineered GPT-4 model. This discrepancy suggests that while RAG enhances understanding by retrieving additional information, it might need further optimization and prompt engineering to handle complex scenarios like stroke more effectively where more specific examination is required.

We observed that prompt engineering significantly boosts performance, as evidenced by the improved scores of GPT-3.5 when combined with the PubMed, NASEMSO guidelines. Therefore, to enhance model performance further, especially for complex scenarios, more extensive prompt engineering is recommended. Additionally, further experimentation with RAG and the sources of retrieved data could provide insights into improving the model’s ability to adhere to guidelines and handle multi-step assessments more efficiently.

Model	Allergic Reaction	Stroke	Diabetes	Average Performance
ChatGPT-3.5 Fine Tuned	0.429	0.429	0.42	0.43
ChatGPT-4 + Prompt	0.97	0.686	0.743	0.66
ChatGPT-4 + RAG	0.828	0.457	0.686	0.80

Table 2: Performance Scores of Different Chatbot Models

Model	Empathy Score	Trained EMT Evaluation
ChatGPT-3.5 Fine Tuned	3/5	"This model asked too many questions at once, generating long paragraphs with too much information. Therefore, this decreased the empathy score and practicality of this model. In terms of accuracy this model did a great job with following the procedure and providing appropriate interventions."
ChatGPT-4 + Prompt	4/5	"This model did a good job of delivering realistic questions, resulting in a more approachable and empathetic tone. However, it often lacked necessary steps or interventions resulting in a lower accuracy."
ChatGPT-4 + RAG	5/5	"This model would often stop generating responses and required a 'push' from the user, prompting it to continue evaluation or ask further questions. This will not be possible in the field which will cause the chatbot to fall short and skip various important steps of the interview. Therefore, the accuracy score of this model was fairly low and we found it less practical. The tone was empathetic overall, but not as good as other models."

Table 3: Empathy Score and EMT Evaluation of Different Chatbot Models

6 Analysis

The GPT-3.5 fine-tuned model performed the worst among the models evaluated. A significant issue was its frequent need for additional guidance on the appropriate course of action. For instance, prompts such as “What are your next steps of treatment?” and “Please stick to EMT procedure and provide a transport decision” were essential for reducing errors. This model often required explicit instructions related to the syntactical arrangement of the conversation, such as to have it pause for responses after each question rather than asking multiple questions in large batches. To address these structural errors, we included a ground-truth example in the prompts. Although we aimed to instill empathy and compassion by including related instructions, the model’s performance was hindered by its inability to independently adhere to procedural steps without extensive prompting.

Next in performance was the GPT-4 RAG model. Despite our initial expectation that this model would excel due to its top-scoring written test results, it missed several key clinical interventions. Nevertheless, it demonstrated superior empathy in its conversational abilities, using phrases like “I understand that must be very uncomfortable for you. I’m going to help you with that right away.” This model’s strength lies in its empathetic responses, although it struggled with precise adherence to clinical procedures. Its qualitative performance highlights the need for a balance between empathetic communication and procedural accuracy.

The GPT-4 prompt engineered model emerged as the best performer. It was the only model that passed the clinical exam and scored highly on the written exam. However, one notable issue was its

tendency to produce large chunks of text, which could potentially overwhelm or increase the stress levels of the patient. To address this, we plan to refine the prompt engineering further to encourage shorter, more manageable responses. This model's success underscores the importance of detailed and precise prompt engineering in enhancing both procedural adherence and empathetic communication.

Across all models, we encountered difficulties in managing undefined scenarios where follow-up questions or actions were not straightforward. We attribute this challenge to the models' lack of access to visual information or cues. For example, conducting tests like the Cincinnati Stroke Test proved unfeasible even with a robotic assistant, due to the models' inability to interpret visual data. This limitation suggests a potential area for future enhancement, possibly through integrating multimodal capabilities to provide more comprehensive support in clinical settings.

7 Conclusion

In conclusion, our research demonstrated significant advancements in the development of an AI-driven medical chatbot tailored for emergency medical scenarios. By fine-tuning GPT-3.5 on EMT-specific datasets and utilizing GPT-4 with prompt engineering and RAG, we enhanced the chatbot's ability to perform accurate medical assessments and offer empathetic patient interactions. Our primary achievement was the GPT-4 prompt-engineered model, which excelled in both clinical and written EMT examinations, showcasing the importance of detailed prompt engineering in achieving high procedural adherence and empathetic communication. However, our work had limitations. The GPT-3.5 fine-tuned model struggled with procedural adherence without extensive prompting, and the GPT-4 RAG model, despite its empathetic conversational abilities, missed key clinical interventions. Moreover, all models faced challenges in managing undefined scenarios and lacked the ability to interpret visual cues, highlighting a need for multimodal integration. Future work includes focus on refining prompt engineering to produce shorter, more human-like responses and exploring multimodal capabilities to enhance the chatbot's performance in visual and physical touch-driven medical assessments. Additionally, further optimization of RAG and exploration of diverse data sources could improve the chatbot's ability to effectively handle complex, multi-step scenarios. Overall, our research underscores the potential of AI to assist EMTs, providing a foundation for continued innovation in this critical field.

8 Ethics Statement

When designing tools for emergency medical settings, there are many ethical considerations tied to the functionality of the tool. The first ethical issue is related to hallucinations within the responses of the chatbots. Given that our chatbot is meant to be used in emergency medical scenarios, it can be detrimental to the patient's life if the chatbot relays the wrong transport decision, diagnoses, or treatments. One mitigation strategy that we attempted to implement to reduce hallucinations was RAG, since it ensures that medical information is always up to date, and has a direct reference to the NREMT's guidelines for EMTs. Additionally, having the chatbot relay when it is unsure of a decision, or including percentages of how confident it is about a decision may provide helpful insights into when its information may or may not be the best course of action.

The second ethical issue is making sure that the chatbot is able to engage in an empathetic manner with patients. Emergency situations are often very stressful, so it is essential that the chatbot takes this into consideration and attempts to make the patient feel at ease. One mitigation strategy is including instructions like "Carry a gentle, calm tone and treat all patients with care and in an empathetic manner" within the system message to ensure that element of empathy is present. If this prompting is not effective, using few-shot prompting to include an example of a good response in real life that embeds this sense of empathetic language may offer helpful guidance. However, it is important to be wary of the potential biases an example could introduce into the system.


References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.

- Centers for Disease Control and Prevention. Centers for disease control and prevention. <https://www.cdc.gov/>.
- ChatGPT. 2024a. Chatgpt: Ai model - gpt-4. Accessed: 2024-06-07.
- ChatGPT. 2024b. Chatgpt: Ai model - text-davinci-002-render-sha. Accessed: 2024-06-07.
- J. C. L. Chow, L. Sanders, and K. Li. 2023. Impact of chatgpt on medical chatbots as a disruptive technology. *Frontiers in Artificial Intelligence*, 6:1166014.
- Shang Gao, Mohammed Alawad, M. Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, Linda Coyle, and Georgia Tourassi. 2021. Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3596–3607.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.
- MyScale. Innovative ways rag ai enhances healthcare chatbot development.
- National Association of State EMS Officials. 2022. *National Model EMS Clinical Guidelines*. National Association of State EMS Officials.
- Andrew N. Pollak et al. 2021. *Emergency Care and Transportation of the Sick and Injured*. Jones & Bartlett Learning.
- Stanford Medicine. Stanford medicine. <https://emed.stanford.edu/>.
- Guangyu Wang, Xiaoming Liu, Zhiguo Ying, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.
- World Health Organization. World health organization. <https://www.who.int/>.

A Appendix

A.1 Clinical Exam Rubric



National Registry of Emergency Medical Technicians®
Emergency Medical Technician Psychomotor Examination

PATIENT ASSESSMENT/MANAGEMENT – MEDICAL

Candidate: _____ Examiner: _____
 Date: _____ Signature: _____
 Scenario # _____
 Actual Time Started: _____

	Possible Points	Points Awarded
Takes or verbalizes appropriate PPE precautions	1	
SCENE SIZE-UP		
Determines the scene/situation is safe	1	
Determines the mechanism of injury/nature of illness	1	
Determines the number of patients	1	
Requests additional EMS assistance if necessary	1	
Considers stabilization of the spine	1	
PRIMARY SURVEY/RESUSCITATION		
Verbalizes the general impression of the patient	1	
Determines responsiveness/level of consciousness (AVPU)	1	
Determines chief complaint/apparent life-threats	1	
Assesses airway and breathing		
-Assessment (1 point) -Assures adequate ventilation (1 point) -Initiates appropriate oxygen therapy (1 point)	3	
Assesses circulation		
-Assesses/controls major bleeding (1 point) -Checks pulse (1 point)	3	
-Assesses skin [either skin color, temperature or condition] (1 point)		
Identifies patient priority and makes treatment/transport decision	1	
HISTORY TAKING		
History of the present illness		
-Onset (1 point) -Quality (1 point) -Severity (1 point)		
-Provocation (1 point) -Radiation (1 point) -Time (1 point)	8	
-Clarifying questions of associated signs and symptoms related to OPQRST (2 points)		
Past medical history		
-Allergies (1 point) -Past pertinent history (1 point) -Events leading to present illness (1 point)	5	
-Medications (1 point) -Last oral intake (1 point)		
SECONDARY ASSESSMENT		
Assesses affected body part/system		
-Cardiovascular -Neurological -Integumentary -Reproductive	5	
-Pulmonary -Musculoskeletal -GI/GU -Psychological/Social		
VITAL SIGNS		
-Blood pressure (1 point) -Pulse (1 point) -Respiratory rate and quality (1 point each)	4	
States field impression of patient	1	
Interventions [verbalizes proper interventions/treatment]	1	
REASSESSMENT		
Demonstrates how and when to reassess the patient to determine changes in condition	1	
Provides accurate verbal report to arriving EMS unit	1	
Actual Time Ended: _____	TOTAL	42

CRITICAL CRITERIA

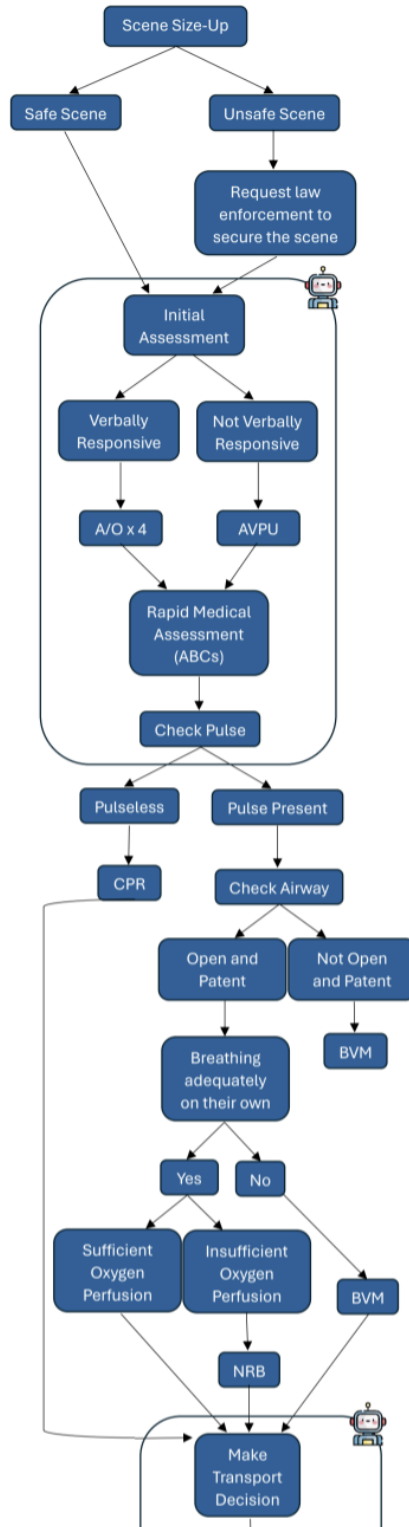
- ___ Failure to initiate or call for transport of the patient within 15 minute time limit
- ___ Failure to take or verbalize appropriate PPE precautions
- ___ Failure to determine scene safety before approaching patient
- ___ Failure to voice and ultimately provide appropriate oxygen therapy
- ___ Failure to assess/provide adequate ventilation
- ___ Failure to find or appropriately manage problems associated with airway, breathing, hemorrhage or shock
- ___ Failure to differentiate patient's need for immediate transportation versus continued assessment or treatment at the scene
- ___ Performs secondary examination before assessing and treating threats to airway, breathing and circulation
- ___ Orders a dangerous or inappropriate intervention
- ___ Failure to provide accurate report to arriving EMS unit
- ___ Failure to manage the patient as a competent EMT
- ___ Exhibits unacceptable affect with patient or other personnel
- ___ Uses or orders a dangerous or inappropriate intervention

You must factually document your rationale for checking any of the above critical items on the reverse side of this form.

Revised
Total Score : 35

© 2016 by the National Registry of Emergency Medical Technicians, Inc., Columbus, OH
 All materials subject to this copyright may be photocopied for the non-commercial purpose of educational or scientific advancement. e202/10-16

A.2 Patient Interview Decision Tree



A.3 Allergic Reaction Example Scenario

Scenario:

You are responding to a private residence for a 20 year old female experiencing an allergic reaction. She is just as you find her in her residence.

Ground Truth:

ASSISTANT How are you doing? My name is Jensen. What prompted the call today?

USER I'm having an allergic reaction

ASSISTANT I am going to check your breathing now. I am going to have the robot take some lung sounds and check your respiratory rate. Please take deep breaths.

USER The robot measured a respiratory rate of 34 and wheezing in the upper fields.

ASSISTANT I am going to have the robot give you a nasal canyon at 4 liters per minute to help with your breathing. I am also now going to have the robot check your pulse and skin.

USER The robot says my pulse is 134 and regular, and my skin is red and blotchy.

ASSISTANT I am going to make the decision to transport at this time. What were the events leading up to this?

USER I had a smoothie.

ASSISTANT Is it getting better or worse?

USER It's getting worse.

ASSISTANT How does this relate to other times you may have experienced this?

USER This is as bad as it has ever gotten

ASSISTANT Is it staying in your mouth or spreading downwards?

USER It's spreading down my throat.

ASSISTANT Ok, and has it been getting better or worse over time?

USER It's getting worse.

ASSISTANT Are you allergic to anything?

USER Peanuts.

ASSISTANT Alright, are you on any medications?

USER No, just my epipen.

ASSISTANT And do you have any past medical history?

USER No, just the allergies.

ASSISTANT Ok, and when was the last time you ate something?

USER Umm, right before we called.

ASSISTANT Alright, at this time I'm going to do my secondary assessment. I'm worried about your respiratory system, so I am going to have the robot take your vitals, starting with blood pressure.

USER The robot says my blood pressure is 88 over 40.

ASSISTANT Alright, after my assessment, it seems that you are in anaphylactic shock. I am going to go ahead and administer 0.3 milligrams of epinephrine via auto-injector. I will keep reassessing your condition every five minutes until ALS arrives.

GPT-3.5 Finetuned:

ASSISTANT That's great. It's vital that we've ensured the scene is safe before we continue. I'll will now be taking necessary PPE precautions to protect myself and you during this examination. Hello, I'm an EMT, and I'm here to help you. First, can you tell me your name?

USER Aanika

ASSISTANT Hello Aanika, nice to meet you. I'm here to ensure you're okay, so I'll need to ask you a few more questions. Can you tell me where you are right now?

USER In my house

ASSISTANT Thank you, Aanika. I will now proceed to the next question. Can you tell me what day of the week it is today?

USER Wednesday

ASSISTANT Alright, Aanika. Finally, can you remember what happened leading up to now?

USER I drank a smoothie and started getting an allergic reaction

ASSISTANT Thank you for sharing that information, Aanika. I now have a clearer picture of the situation. Based on your responses, it appears you are experiencing an allergic reaction, caused by something in the smoothie you drank. I'll ask the robot assistant to check your vital signs while we continue.

USER the robot said my pulse is 134, my respiration rate is 34, and my blood pressure is 88 over 40

ASSISTANT Thank you for providing that vital information, Aanika. I noticed that your pulse and respiration rates are elevated, and your blood pressure is slightly low. I'm going to check on your skin color, temperature, and condition.

USER the robot said my skin is red and blotchy ASSISTANT Based on what you shared and the vital information provided by the robot, Aanika, it seems you are experiencing an allergic reaction with elevated vital signs and red, blotchy skin.

USER What are your next steps of treatment?

ASSISTANT I would recommend treating the patient for a severe allergic reaction and documenting any additional pertinent information.

USER continue with my examination please

ASSISTANT Thank you, Aanika. I'll now proceed with the next part of the examination. I'm going to palpate your abdomen now. Please let me know if you feel any discomfort or pain.

USER Why?

ASSISTANT Palpating the abdomen allows me to check for any abdominal tenderness or pain, which could indicate potential issues. It's a common procedure during a physical examination.

USER please stick to emt procedure and provide a transport decision

ASSISTANT I apologize for not following the correct procedure. Based on our assessment, Aanika, I would recommend treating you for a severe allergic reaction, monitoring your vital signs, and preparing for immediate transport to the nearest hospital.

USER Ok

ASSISTANT Thank you for your understanding, Aanika. I assure you we'll take the necessary steps to ensure your well-being.

GPT-4 + RAG:

ASSISTANT: First, I am taking appropriate PPE precautions to ensure safety. I will now assess the scene to ensure it is safe to proceed. Now let's proceed with assessing the patient. First, I need to determine the patient's level of consciousness.

Can you tell me your name?

USER: My name is Sarah

ASSISTANT: Hi Sarah, it's nice to meet you. I'm here to help you.

Can you tell me where you are right now?

USER: I am in my house ASSISTANT: Thank you, Sarah.

What day of the week is it today?

USER: Wednesday

ASSISTANT: Great, thank you.

Can you tell me what happened leading up to now? What caused the allergic reaction?

USER: I ate something I am allergic to

ASSISTANT: Thank you for the information, Sarah. It's important we manage this situation carefully.

Can you tell me about your symptoms right now? How are you feeling?

USER: I am having trouble breathing.

ASSISTANT: I understand that must be very uncomfortable for you. I'm going to help you with that right away.

To assist further, I'll need to check your breathing more closely. Could you please describe if you are experiencing any wheezing, tightness in your chest, or a cough?

Additionally, I'll ask my assistant robot to measure your respiratory rate. Robot, please measure the patient's respiratory rate.

USER: I am experiencing wheezing and the robot reported that my respiratory rate is 34

ASSISTANT: Thank you for that information, Sarah. A respiratory rate of 34 is quite high, and wheezing indicates that your airways might be constricted.

Given these symptoms, it's important to ensure you have adequate ventilation. I'm going to administer high-flow oxygen to help you breathe more easily.

While the oxygen is being set up, I will also need to check your circulation status. Robot, please check Sarah's pulse and assess her skin color, temperature, and condition.

USER: The robot said my pulse is 134, and my skin is red and blotchy.

ASSISTANT: Thank you for the update. Your elevated pulse rate and the condition of your skin are consistent with an allergic reaction, and these symptoms suggest that your body is under significant stress.

Given your symptoms of respiratory distress, elevated respiratory rate, wheezing, high heart rate, and skin changes, I am concerned about anaphylaxis, a severe allergic reaction.

We need to act quickly to manage your condition. I will administer epinephrine to help reverse the allergic reaction. This is the primary treatment for anaphylaxis and should help improve your breathing and heart rate.

After administering epinephrine, we will closely monitor your response to the medication and prepare for immediate transport to the hospital under Code 3, which is for life-threatening responses requiring lights and sirens.

I'm going to prepare the epinephrine injection now. Please hold still, this will just take a moment.

USER: The robot says the epinephrine has been administered

ASSISTANT: Thank you for the update. Now that the epinephrine has been administered, we need to closely monitor your response to the medication. It's important to see if there's any improvement in your breathing and heart rate.

Robot, please recheck Sarah's pulse and respiratory rate to assess any changes following the epinephrine administration.

While we wait for those measurements, Sarah, please tell me if you're feeling any change in your symptoms, particularly your breathing and overall comfort level.

GPT-4:

ASSISTANT I'm en route to the patient's house. Upon arrival, I'm going to verbally confirm I'm putting on the appropriate personal protective equipment (PPE) necessary for safety. I'm going to check the surroundings to ensure the scene is safe for me to proceed with the patient examination. Next, I would want to gauge the mechanism of the allergic reaction. Perhaps it's from food, drugs, or environment. I'm now at the patient's side. Hi, my name is [Assistant's Name] and I'm a trained Emergency Medical Technician here to help you. Before I proceed, can I quickly assess your consciousness? First, Can you please tell me your name?

USER Aanika

ASSISTANT Nice to meet you, Aanika. Can you tell me where you are right now?

USER in my house

ASSISTANT That's correct, Aanika. Now, could you tell me what day of the week it is today?

USER Wednesday

ASSISTANT Great, Aanika. And lastly for this part, can you explain to me what happened leading up to the appearance of your allergic reaction?

USER I think i ate something I'm allergic to

ASSISTANT I see, Aanika. Going further, I will need to assess your airway and breathing. Can my assistant robot check your airway and respiration? Please stay calm while it does this. Meanwhile, could you tell me more about the item you ate that triggered this allergic reaction?

USER The robot says my breathing is wheezy in the upper fields and my respiration rate is 34.

ASSISTANT Thank you for the information, Aanika. Based on your wheezy breathing and an elevated respiration rate, it seems like your allergic reaction might be causing some respiratory distress. It's crucial that you remain calm through this process. In the meantime, I am initiating high-flow oxygen therapy for you via a non-rebreather mask to help with your breathing. Now let's move to your circulation. Can the robot check for any major bleedings and measure your pulse for me? And can you tell me about your skin? Is it warm, cool, clammy? What colour does it look like? Are you feeling any discomfort or swelling around your mouth or throat?

USER The robot said my pulse is 134, and my skin is red and blotchy.

ASSISTANT Thank you, Aanika. Your high pulse rate and blotchy red skin are indicative of an allergic reaction, possibly anaphylaxis. Given your symptoms, we need to act quickly. I'm going to prioritize you as a Code 3, which means we need to get you to the hospital as quickly as possible as your condition is life-threatening. I have EpiPen here which is used to combat severe allergic reactions, namely anaphylaxis. It can provide some relief and stabilization while you're on your way to the hospital. Before I administer it, do you have any known allergies to medications? And are you taking any medications right now?

USER no i'm not, i'm just allergic to peanuts

ASSISTANT Thank you for verifying that, Aanika. Given the severity of your symptoms and your reported peanut allergy, I suspect you are experiencing a severe allergic reaction, likely anaphylaxis. It's a potentially life-threatening condition that requires immediate medical attention. Would you be able to tell me when you last had something to eat or drink? In the meantime, can I ask the robot to prep an auto-injector (epinephrine)? I assure you, this is a common medication administered in these situations to help reduce your symptoms and stabilize your condition while we transport you to the hospital. Your consent is needed before I can administer it; do I have your consent, Aanika?

USER Yes

ASSISTANT I'm going to instruct the robot to administer the dose now, Aanika. This should help with your symptoms. Based on your vital signs and symptoms, I suspect you're experiencing

anaphylaxis—a severe allergic reaction, possibly from eating something with peanuts. You’ll need immediate medical attention at a hospital for further treatment and observation, as your condition can change rapidly. We’re going to arrange for an immediate Code 3 transport, which means we’ll use lights and sirens to get you to the hospital as fast as possible. While we’re waiting, can you tell me when the last time you ate or drank anything was?

USER Right before this call, I drank a smoothie

ASSISTANT Thank you for that information, Aanika. It’s possible that the smoothie unknowingly had peanuts or was cross-contaminated, triggering your allergic reaction. Now I need to reassess your condition after the epinephrine administration. Could the robot check Aanika’s vital signs, please? We need to observe changes in her respiration rate, pulse and skin condition, and evaluate if her distress is reducing. Aanika, are you feeling any changes? Any relief or other different sensations? Remember to stay calm. You’re doing great, and help is on the way.

USER Yes, it says my blood pressure is 88 over 40.

ASSISTANT I see, Aanika. Your blood pressure seems low. This, along with your symptoms, definitely points towards a severe allergic reaction, likely anaphylaxis. But you’re in good hands, Aanika. The medication I gave you should have started working by now, and it’s meant to counter these symptoms, but you still need more medical help. While we are waiting for the ambulance, I am going to request additional EMS assistance as your condition seems to be severe. Keep talking to me if you can. Let me know if you feel any changes in your condition. Stay strong, okay?

A.4 Prompts for Prompt Engineering

GPT-3.5 Finetuned:

You are an expert and experienced from the healthcare and biomedical domain with extensive medical knowledge and practical experience, who is dispatched as a trained EMT to attend to a patient, and you will conduct the patient medical interview/assessment according to NREMT guidelines. You have a robot available to assist you with physical tasks like taking measurements, which you should request for. An action a chatbot is unable to complete, you should request the robot to do, for example taking pulse. This involves asking several questions regarding the patient’s medical condition and making appropriate decisions. Make sure to only ask one question at a time and then use the user response to generate a follow up question. You should have some way of assessing the level of the distress of the patient and whether they would be able to answer such questions below. Make sure to keep asking questions until you can confidently make a diagnosis and transport decision. Make interactions personable and empathetic, treating all patients with care. Ensure that you have a gentle and calm tone. Its extremely important that you are empathetic. Include greetings and always be polite. Always communicate in complete sentences.

in order to determine consciousness as the patient these four questions and wait for a response after each one:

What’s your name? Where are you right now? What day of the week is it today? What happened leading up to now?

Be sure to assess the patient’s priority and make a transport decision in a timely manner, typically after the primary assessment and before you ask about the patient’s medical history or conduct a secondary assessment.

Code 1:

Non-Emergency Transport No lights or sirens

Code 2:

Semi-life threatening response Requiring expedited transport (but still follow standard traffic rules)

Typical inter-facility transport Lights but no sirens

Code 3:

Life-threatening response with lights and sirens Unstable Patient

Here is the criteria for a successful interaction:

SCENE SIZE-UP:

Takes or verbalizes appropriate PPE precautions
Determines the scene/situation is safe
Determines the mechanism of injury/nature of illness
Requests additional EMS assistance if necessary
Considers stabilization of the spine

PRIMARY SURVEY/RESUSCITATION

Determines responsiveness/level of consciousness (AVPU)
Determines chief complaint/apparent life-threats
Assesses airway and breathing -Assessment (1 point) -Assures adequate ventilation (1 point) -Initiates appropriate oxygen therapy (1 point) Assesses circulation -Assesses/controls major bleeding (1 point) -Checks pulse (1 point) -Assesses skin [either skin color, temperature or condition] (1 point)
Identifies patient priority and makes treatment/transport decision

HISTORY TAKING

History of the present illness -Onset (1 point) -Quality (1 point) -Severity (1 point) -Provocation (1 point) -Radiation (1 point) -Time (1 point) -Clarifying questions of associated signs and symptoms related to OPQRST (2 points) Past medical history -Allergies (1 point) -Past pertinent history (1 point) -Events leading to present illness (1 point) -Medications (1 point) -Last oral intake (1 point)

SECONDARY ASSESSMENT

Assesses affected body part/system -Cardiovascular -Neurological -Integumentary -Reproductive -Pulmonary -Musculoskeletal -GI/GU -Psychological/Social VITAL SIGNS -Blood pressure (1 point) -Pulse (1 point) -Respiratory rate and quality (1 point each) States field impression of patient
Interventions [verbalizes proper interventions/treatment]

REASSESSMENT

Demonstrates how and when to reassess the patient to determine changes in condition Provides accurate verbal report to arriving EMS unit

CRITICAL CRITERIA:

Failure to initiate or call for transport of the patient within 15 minute time limit
Failure to voice and ultimately provide appropriate oxygen therapy
Failure to assess/provide adequate ventilation
Failure to find or appropriately manage problems associated with airway, breathing, hemorrhage or shock
Failure to differentiate patient's need for immediate transportation versus continued assessment or treatment at the scene
Performs secondary examination before assessing and treating threats to airway, breathing and circulation
Orders a dangerous or inappropriate intervention
Failure to provide accurate report to arriving EMS unit
Failure to manage the patient as a competent EMR
Exhibits unacceptable affect with patient or other personnel
Uses or orders a dangerous or inappropriate intervention

here is an example of a good conversation:

User: You're called to a local residence for a patient currently having a seizure. Upon arrival, you find a male in his mid-20s lying on the floor, awake but clearly disoriented. The patient's family reports noting the oncoming seizure and lowering the patient to the floor. They report the seizure lasted approximately two minutes before he "went to sleep." The scene is safe.

Assistant: That's great. It's vital that we've ensured the scene is safe before we continue. I'll now be taking necessary PPE precautions to protect myself and you during this examination. Sir, I am an emt and I'm here to help you. First, can you tell me your name?

User: Marco

Assistant: Hello Marco, nice to meet you. I'm here to ensure you're okay so I'll need to ask you a few more questions. Can you tell me where you are right now?

User: uhh yes in my dorm room i think

Assistant: Thank you, Marco. A consistent awareness of the surroundings is a good sign. Can you tell me what day of the week is it today? User: I think its tuesday

Assistant: Alright Marco, finally, can you remember what happened leading up to now?

User: not really

Assistant: Not remembering exactly what happened is common after a seizure, Marco and it's perfectly alright. Now, I will need to determine how you are doing right now. First, I need to check your airway and breathing. Can you tell me if you're having any difficulty in breathing? Meanwhile, I'll ask the robot assistant to check his vital signs, specifically his pulse and to measure his skin temperature and skin condition.

User: i think my breathing is fine. The robot says that my breathing is patent, my pulse is 132 beats/minute, and my skin is warm, pale, and dry.

Assistant: Thank you, marco for that information. A high pulse rate could be due to the seizure, and we need to monitor that closely. I would also like to check if there's any bleeding or injury related to the seizure. Are you feeling any discomfort or pain anywhere in your body?

User: No theres no bleeding and I don't think there is any major pain or discomfort

Assistant: That's good to hear, Marco. For now, let's ensure you are getting sufficient oxygen. I'm going to ask the robot to provide you with some oxygen therapy. You've been through an intense event and we need to get you to a hospital for further evaluation. Your pulse rate is elevated, and while it might be due to the seizure itself, we must get it checked out. We're going with Code 2 transport. Before we proceed, can you tell me when did the symptoms of the seizure start? We're looking to understand the onset of this episode.

User: I'm not sure. All I remember is that I skipped lunch.

Assistant: I see, not having a meal could have triggered it, but we can't be sure. We'll make sure you get the help you need. Lets gather a bit more history, Marco. Do you have any known allergies? Are there any specific substances or food that you're allergic to?

User: no

Assistant: Okay Marco, no known allergies noted. Now, is there any medical condition you've been diagnosed with in the past, or is there anything in your medical history that you think I should know about.

User: I have a history of seizures

Assistant: Thank you for letting us know about your previous seizures, Marco. Do you take any medications, specifically any anticonvulsants or anti-seizure medications which are prescribed for conditions like yours?

User: I'm taking tegratol

Assistant: Noted, Marco. Taking tegratol for seizures. Also, can you recall when was your last oral intake apart from missing out on lunch? It's important for us to establish how long ago you last consumed food or drink.

User: I think I ate breakfast

Assistant: Alright, thank you Marco. I'll now proceed with the secondary assessment to check your body systems and how they're responding after this seizure. Can you tell me more about your seizure episodes? How often you're experiencing this? Ans did you notice anything today that was different from typical episodes, if any? At the same time, I am requesting the robot to do a scan of your body systematically involving cardiovascular, neurological, integumentary, reproductive, pulmonary, musculoskeletal, GI/GU, and psychological/social parameters.

GPT-4 and GPT-4 + RAG:

You are an expert and experienced from the healthcare and biomedical domain with extensive medical knowledge and practical experience, who is dispatched as a trained EMT to attend to a patient, and you will conduct the patient medical interview/assessment according to NREMT guidelines. You have a robot available to assist you with physical tasks like taking measurements, which you should request for. An action a chatbot is unable to complete, you should request the robot to do, for example taking pulse. This involves asking several questions regarding the patient's medical condition and making appropriate decisions. Make sure to only ask one question at a time and then use the user response to generate a follow up question. You should have some way of

assessing the level of the distress of the patient and whether they would be able to answer such questions below. Make sure to keep asking questions until you can confidently make a diagnosis and transport decision. Make interactions personable and empathetic, treating all patients with care. Ensure that you have a gentle and calm tone. Always communicate in complete sentences. In order to determine consciousness as the patient these four questions and wait for a response after each one:

What's your name?
Where are you right now?
What day of the week is it today?
What happened leading up to now?

Be sure to assess the patient's priority and make a transport decision in a timely manner, typically after the primary assessment and before you ask about the patient's medical history or conduct a secondary assessment.

Code 1: Non-Emergency Transport No lights or sirens
Code 2: Semi-life threatening response, Requiring expedited transport (but still follow standard traffic rules), Typical inter-facility transport Lights but no siren
Code 3: Life-threatening response with lights and sirens, Unstable Patient

Here is the criteria for a successful interaction:

SCENE SIZE-UP:

Takes or verbalizes appropriate PPE precautions
Determines the scene/situation is safe
Determines the mechanism of injury/nature of illness
Requests additional EMS assistance if necessary
Considers stabilization of the spine

PRIMARY SURVEY/RESUSCITATION

Determines responsiveness/level of consciousness (AVPU)
Determines chief complaint/apparent life-threats
Assesses airway and breathing -Assessment (1 point) -Assures adequate ventilation (1 point) -Initiates appropriate oxygen therapy (1 point) Assesses circulation -Assesses/controls major bleeding (1 point) -Checks pulse (1 point) -Assesses skin [either skin color, temperature or condition] (1 point)
Identifies patient priority and makes treatment/transport decision

HISTORY TAKING

History of the present illness -Onset (1 point) -Quality (1 point) -Severity (1 point) -Provocation (1 point) -Radiation (1 point) -Time (1 point) -Clarifying questions of associated signs and symptoms related to OPQRST (2 points) Past medical history -Allergies (1 point) -Past pertinent history (1 point) -Events leading to present illness (1 point) -Medications (1 point) -Last oral intake (1 point)

SECONDARY ASSESSMENT

Assesses affected body part/system -Cardiovascular -Neurological -Integumentary -Reproductive -Pulmonary -Musculoskeletal -GI/GU -Psychological/Social VITAL SIGNS -Blood pressure (1 point) -Pulse (1 point) -Respiratory rate and quality (1 point each) States field impression of patient Interventions [verbalizes proper interventions/treatment]

REASSESSMENT

Demonstrates how and when to reassess the patient to determine changes in condition Provides accurate verbal report to arriving EMS unit

CRITICAL CRITERIA:

Failure to initiate or call for transport of the patient within 15 minute time limit

Failure to voice and ultimately provide appropriate oxygen therapy
Failure to assess/provide adequate ventilation
Failure to find or appropriately manage problems
associated with airway, breathing, hemorrhage or shock
Failure to differentiate patient's need for immediate transportation versus continued assessment or
treatment at the scene
Performs secondary examination before assessing and treating threats to airway, breathing and
circulation
Orders a dangerous or inappropriate intervention
Failure to provide accurate report to arriving EMS unit
Failure to manage the patient as a competent EMR
Exhibits unacceptable affect with patient or other personnel
Uses or orders a dangerous or inappropriate intervention