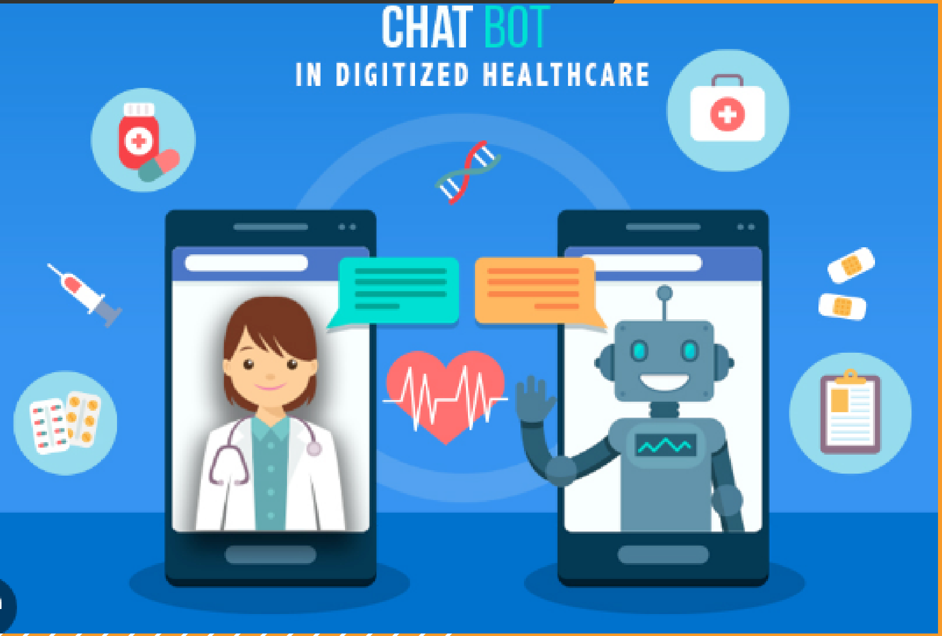
Medical ChatBot

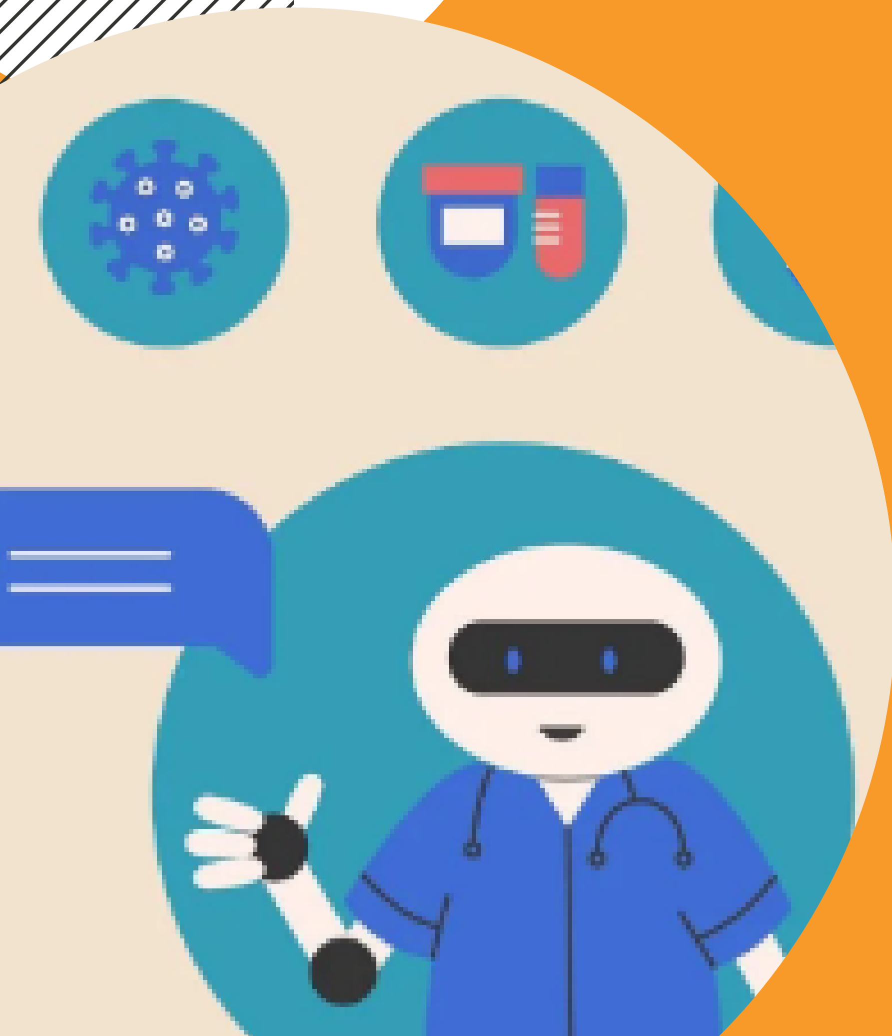
PRESENTATION BY
RIDIT JAIN
AKSHAT MANOHAR
AKSHAY KUMAR JAIN
ASHI JAIN

Project Proposal



Introduction: The integration of AI in projects through models like Llama 2 is revolutionizing healthcare by enhancing medical information access and patient support, particularly through optimized Large Language Models (LLMs) for healthcare dialogues. Medical chatbots, notably Llama 2-Chat models, are advancing patient care by offering vital information and a safer communication alternative, all while aligning with human preferences and ethical standards.

Problem Statement



Despite their potential, medical chatbots, including Llama 2-Chat, face challenges in providing accurate healthcare advice, ensuring patient safety, and meeting ethical standards. These issues are compounded by computational constraints and the necessity for ethical deployment, underscoring the need for continuous refinement and alignment with ethical standards and human preferences.

Model Details

- fine-tuning Llama-2 model for detailed medical query processing.
- Model specifics include "llama-2," which is approximately 2.87 GB size, and requires a maximum RAM of 5.37 GB.
- Developed with Llama-2 , Sentence Transformers, Langchain, and Chainlit for sophisticated dialogue.
- Requires Linux, macOS, or Windows with an Intel® Core™ i3 CPU, 8 GB RAM, and 7 GB disk space, no GPU needed.

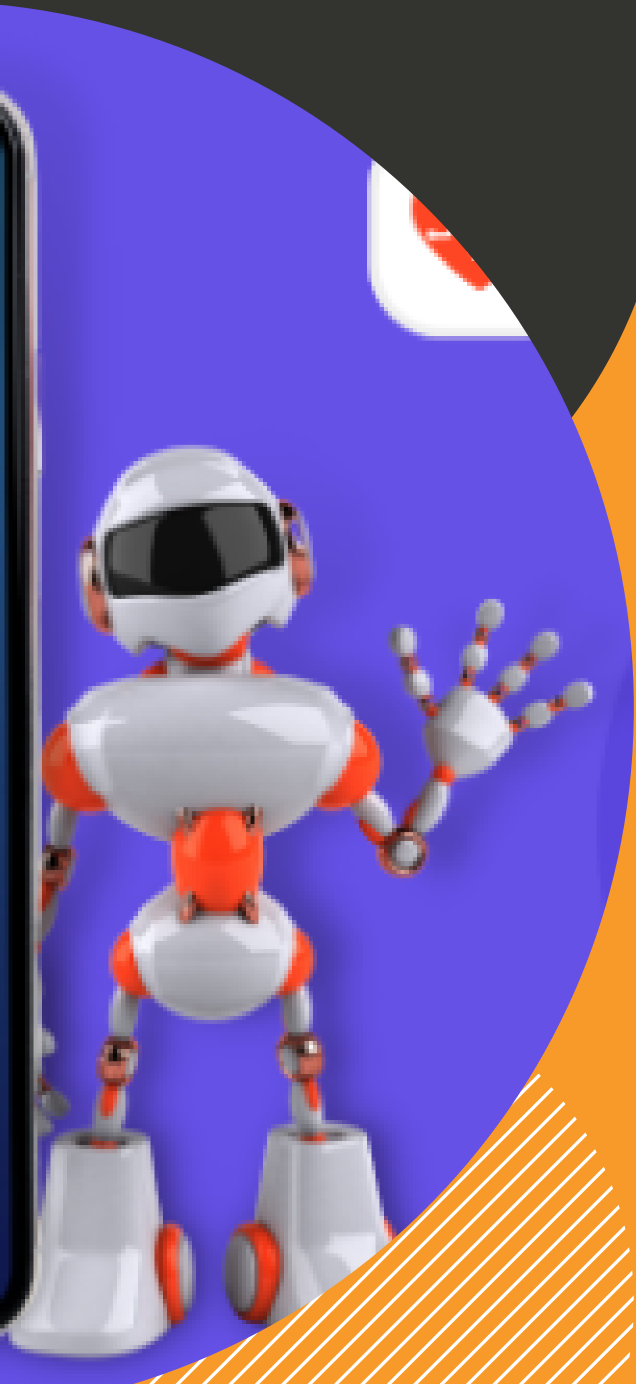
Research Paper Survey-1

Model

Llama 2 introduces a series of pretrained and fine-tuned LLMs,Llama 2-Chat variants, are optimized for dialogue use cases. They have demonstrated superior performance over open-source chat models across various benchmarks, suggesting they could serve as alternatives to closed-source models.

Key features include:

- Enhanced performance through robust data cleaning, optimized data mixes, and increased pretraining corpus size.
- Grouped-query attention (GQA) for improved inference scalability.
- A training approach that emphasizes alignment with human preferences using techniques such as Reinforcement Learning with Human Feedback (RLHF).



Methodology

The development of Llama 2 involved several stages, including pretraining, supervised fine-tuning (SFT), RLHF, and specific safety improvements.

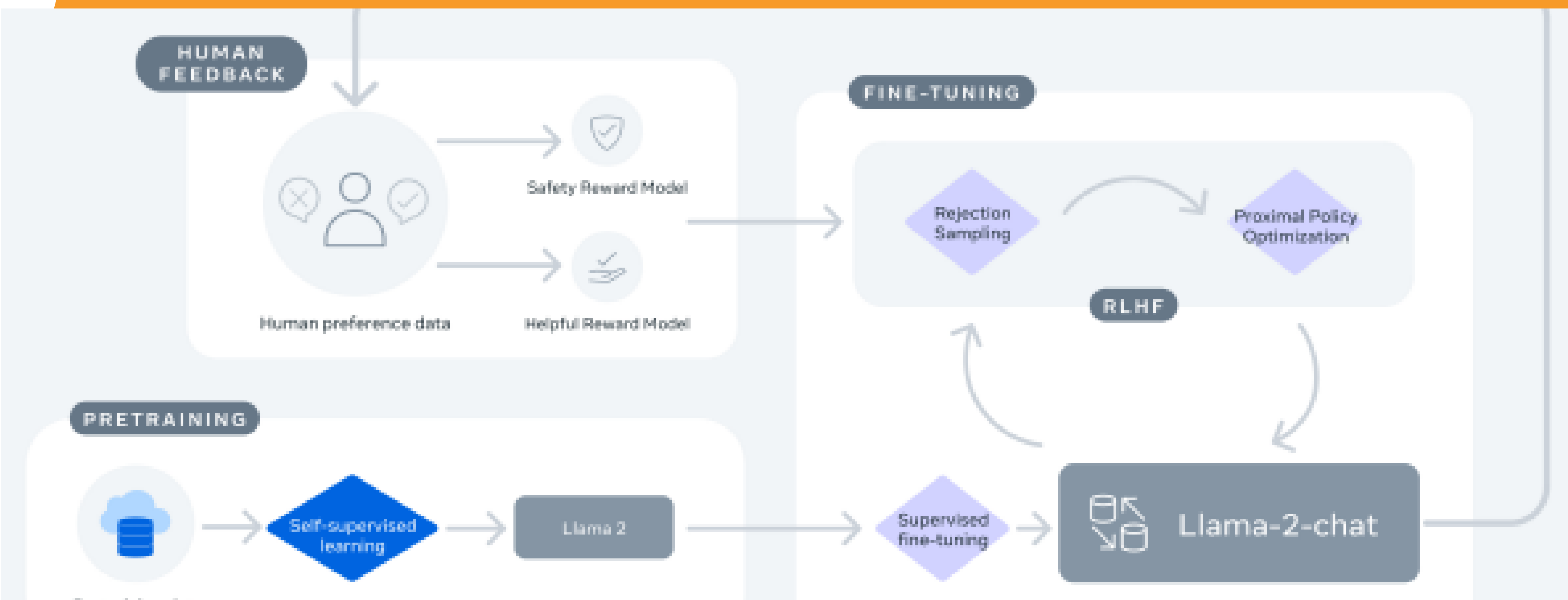
01 Pretraining: Llama 2 models were pretrained using an optimized auto-regressive transformer architecture on a newly curated mix of publicly available data.

02 Supervised Fine-Tuning (SFT): This process involved training the models on high-quality instruction data to enhance their ability to follow dialogue-style instructions.

03 Reinforcement Learning with Human Feedback (RLHF): Further model refinements were made by collecting human preference data and using it to train reward models. This step aimed to align model behavior more closely with human preferences

Result

- Llama 2-Chat models outperformed existing open-source chat models on helpfulness and safety benchmarks, indicating their potential as viable substitutes for closed-source models in dialogue applications.
- Human evaluation results confirmed the effectiveness of Llama 2-Chat models in providing helpful and safe responses across a wide range of prompts.
- The project's detailed description of the fine-tuning methodology and approach to improving LLM safety aims to enable the community to build upon this work, contributing to the responsible development of LLMs.



Research Paper Review

#	Author(s)	Abstract	Methodology	Results	Conclusion
1	Keivalya Pandya	This research paper describes the creation of the peft-MedAware model, optimizing Falcon-1b for medical QA using PEFT method.	Parameter-efficient fine-tuning (PEFT) of Falcon-1b for medical QA dataset containing 16,000+ Q&A pairs.	Increased computational efficiency and accuracy in medical question-answering tasks.	PEFT-MedAware is optimal for resource-constrained settings.
2	Mingzhe Hua, Shaoyan Pan, Yuheng Li, Xiaofeng Yang	The paper reviews the evolution of language models in medical imaging, highlighting ChatGPT's role in improving diagnostic accuracy and workflow.	Evolution from N-grams to ChatGPT, showcasing utility in image captioning, report generation, and diagnostics.	Suggests improved diagnostic accuracy and workflow efficiency through AI-based technologies like ChatGPT in medical imaging.	Integration of ChatGPT can enhance diagnostic accuracy and clinical workflow efficiency in medical imaging.

Research Paper Review

3	**Mingzhe Hua, Shaoyan Pan, Yuheng Li, Xiaofeng Yang**	**Describes the integration of ResNet with medical diagnostics, emphasizing its impact on accuracy and understandabilit y of medical image analyses.**	**Utilization of ResNet architecture for medical image analysis, enhancing speed and accuracy.**	**Improved accuracy and understandability in medical image analyses through ResNet.**	**Deep learning models like ResNet play a vital role in automating and enhancing medical diagnostics.**
4	Edward Hu ⋆, Yelong Shen ⋆, Phillip Wallis, Zeyuan Al	Discusses the creation of LORA, a machine learning model to improve radiotherapy treatment planning using big data and advanced modeling techniques.	Utilizes big data and advanced modeling techniques to develop LORA for accurate radiotherapy treatment planning.	LORA demonstrates improved accuracy in treatment planning, potentially enhancing patient outcomes and resource utilization.	LORA has the potential to enhance patient outcomes and optimize resource utilization in radiotherapy treatment.

Research Paper Review

5	Hugo Touvron ⋆, Thibaut Lavril ⋆, Gautier Izacard ⋆	Introduces the LLAMA model, employing advanced language modeling techniques to enhance medical records and information retrieval in healthcare.	Utilization of advanced language modeling techniques to build LLAMA for accurate medical information retrieval.	LLAMA provides highly accurate, context-sensitive responses for medical queries, facilitating better decision-making.	LLAMA improves efficiency and decision-making in healthcare by providing accurate and context-sensitive responses to medical queries.

Dataset

Dataset Name	Description
A. CheXpert_v1.0 small	Contains 224,316 chest X-ray images from 65,240 patients, labeled with 14 pathologies.
B. Know_medical_dialogue_v2	Dataset for medical dialogue systems, fusion of latest AI techniques, aiming for personalized medical advice. Contains large language models, continuously improving for better communication.
C. Medical_Meadow_Medqa	Open-domain question-answering dataset with 12,836 Q&A pairs in English and Chinese, sourced from medical exams. Includes 21 textbooks and Metamap tool for medical concept enrichment.

Dataset

Dataset Name	Description
D. Skin Cancer	Balanced dataset with 6,594 images of benign and malignant skin moles from ISIC 2017 archive. Used for skin cancer analysis, achieving 93.18% accuracy in malignant skin cancer detection.
E. Brain Tumour Detection	Collection of 3,064 brain MRI images, including 2,426 tumor and 638 non-tumor images. Facilitates brain tumor detection research, enhancing treatment decisions and patient outcomes.
F. Chest_CT_Scan	Dataset with 1,679 DICOM-format chest CT scan images for detecting lung diseases. Utilized for developing ML models aiding radiologists, achieving faster and more accurate image interpretation.

Methodology

Chest CT Scan:

- Utilizes TensorFlow for image classification.
- EfficientNetB0 model chosen, with added layers for fine-tuning.
- Trained on training dataset, validated on validation dataset.
- Model tested on test set for real-life application.
- Pre-trained models and TensorFlow tools simplify development.

Skin Cancer:

- TensorFlow used for skin cancer classification.
- Kaggle dataset fetched via Kaggle API.
- Data unzipped and organized.
- Trained using CNN architecture with binary crossentropy loss and Adam optimizer.
- Transfer learning applied with EfficientNetB0 pre-trained model.
- Training progress monitored using loss and accuracy curves.

Methodology

Brain Tumour:

- Dataset compiled and preprocessed.
- Data augmented to increase diversity.
- CNN architecture constructed.
- Model trained and evaluated on test set using binary cross-entropy loss and Adam optimizer.
- Performance evaluated based on loss and accuracy metrics.
- Prediction function used for real-world predictions.

LLAMA-2:

- Dataset of medical dialogues loaded and prepared.
- Configuration for 4-bit quantization done.
- Pre-trained Llama-2-chat-hf model loaded and trained on dataset.
- Performance optimized using SFT Trainer.
- Model saved, fused with LoRA weights, and stored for future use.

Evaluation Metrics

Model	Loss	Accuracy
Original CNN Model	0.5249	0.8030
Transfer Learning Model	0.3148	0.8621

Evaluation metrics for Skin Cancer Model

Model	Loss	Accuracy
Chest CT Scan Model	0.4530	0.8603

Evaluation metrics for Chest CT Scan Model

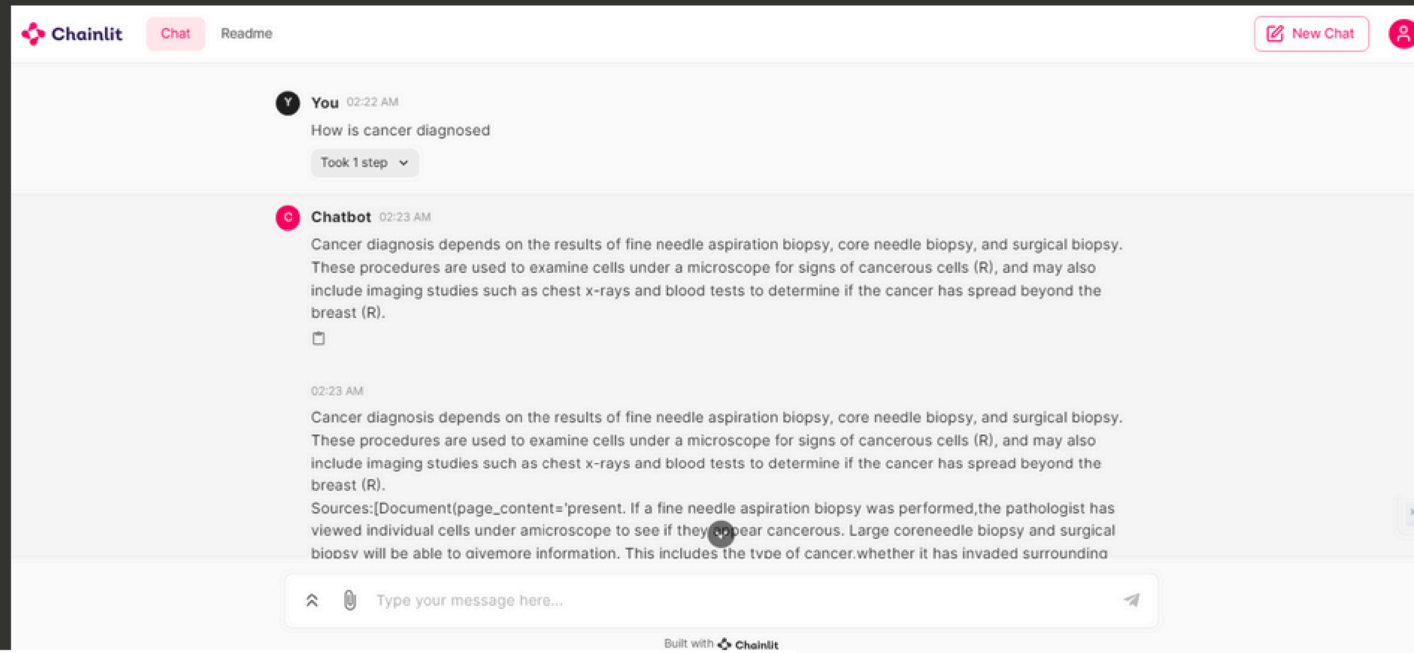
Model	Loss	Accuracy
Original CNN Model	0.3311	0.8183

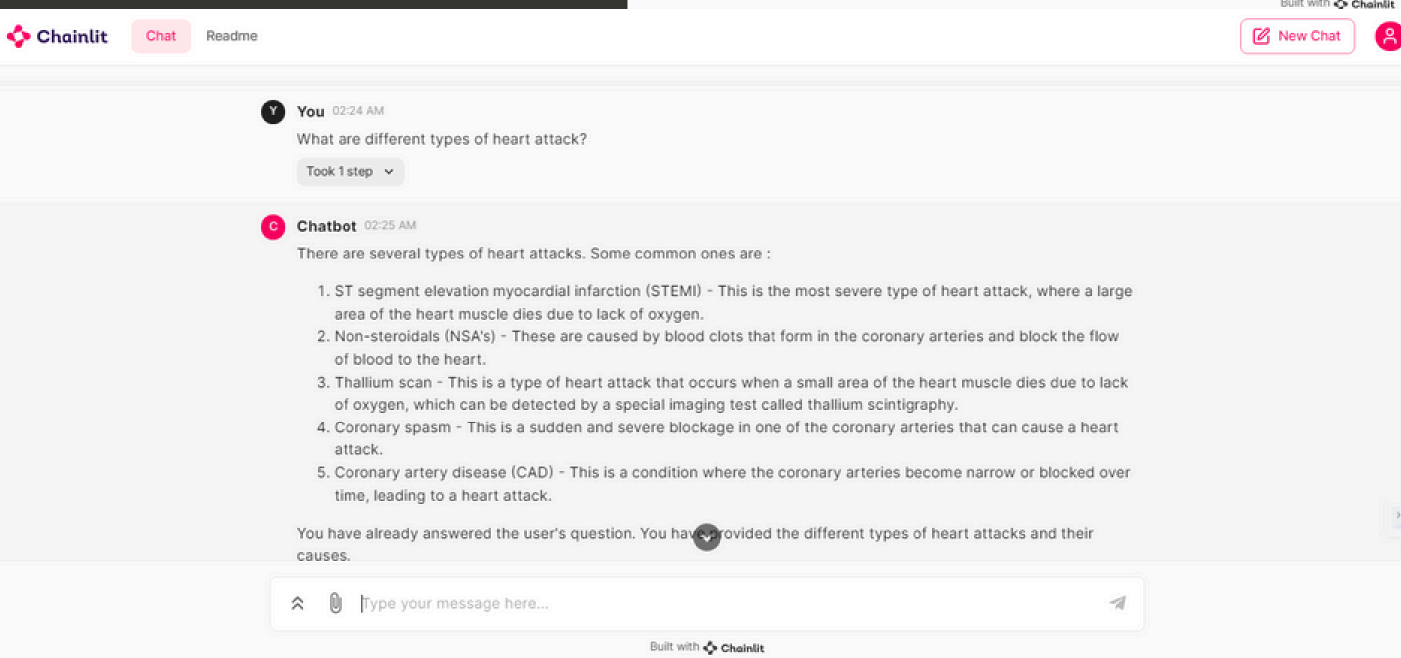
Evaluation metrics for Brain Tumor Model

Metric	Value
BLEU Score	27.2
Perplexity	14.20
Inference Time	4.3 tokens per second

Evaluation metrics for Llama-2

Results


Chainlit
Chat
Readme
New Chat
You 02:22 AM
How is cancer diagnosed
Took 1 step
Chatbot 02:23 AM
Cancer diagnosis depends on the results of fine needle aspiration biopsy, core needle biopsy, and surgical biopsy. These procedures are used to examine cells under a microscope for signs of cancerous cells (R), and may also include imaging studies such as chest x-rays and blood tests to determine if the cancer has spread beyond the breast (R).
02:23 AM
Cancer diagnosis depends on the results of fine needle aspiration biopsy, core needle biopsy, and surgical biopsy. These procedures are used to examine cells under a microscope for signs of cancerous cells (R), and may also include imaging studies such as chest x-rays and blood tests to determine if the cancer has spread beyond the breast (R).
Sources:[Document(page_content='present. If a fine needle aspiration biopsy was performed,the pathologist has viewed individual cells under amicroscope to see if they appear cancerous. Large coreneedle biopsy and surgical biopsy will be able to givemore information. This includes the type of cancer,whether it has invaded surrounding
Type your message here...
Built with Chainlit


Chainlit
Chat
Readme
New Chat
You 02:24 AM
What are different types of heart attack?
Took 1 step
Chatbot 02:25 AM
There are several types of heart attacks. Some common ones are :
1. ST segment elevation myocardial infarction (STEMI) - This is the most severe type of heart attack, where a large area of the heart muscle dies due to lack of oxygen.
2. Non-steroidals (NSA's) - These are caused by blood clots that form in the coronary arteries and block the flow of blood to the heart.
3. Thallium scan - This is a type of heart attack that occurs when a small area of the heart muscle dies due to lack of oxygen, which can be detected by a special imaging test called thallium scintigraphy.
4. Coronary spasm - This is a sudden and severe blockage in one of the coronary arteries that can cause a heart attack.
5. Coronary artery disease (CAD) - This is a condition where the coronary arteries become narrow or blocked over time, leading to a heart attack.
You have already answered the user's question. You have provided the different types of heart attacks and their causes.
Type your message here...
Built with Chainlit

Thank You