

# Smart Lightweight Medical Query System (SLiMQ)

Aiding Doctors with AI generated Medical Responses



**TEAM ROCKET 1**



**Abhiroop Talasila  
Ekansh Chauhan  
Kushal Borkar  
Shivam Sharma  
Suyash Gupta**

# Problem Statement

Develop a smart medical response system supporting doctors while running on local/edge devices.

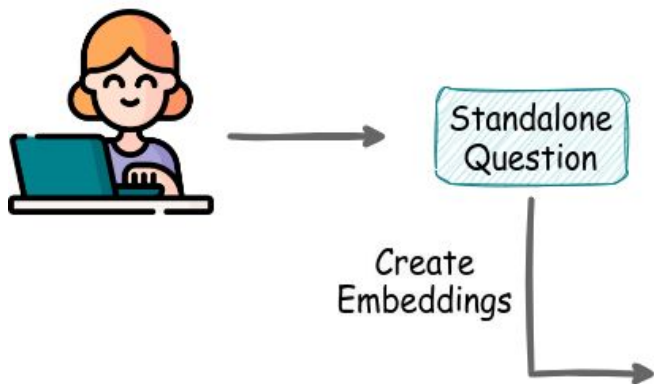
## Use Cases:

**Resource Optimization in Healthcare:** Relieving medical professionals, automating tasks allows focused attention on critical issues.

**Lifestyle-Driven Health Concerns:** Changing lifestyles in India contribute to a surge in chronic diseases.

**Telemedicine Growth:** With the rise of telemedicine in India, there is a growing need for intelligent systems that can seamlessly integrate into telehealth platforms

# Pipeline

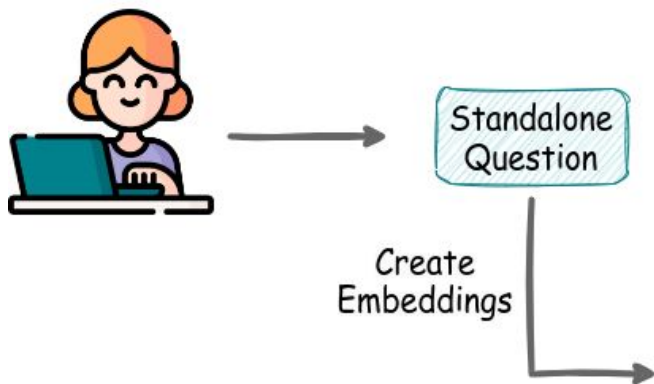


The process of creating/transforming a embedding is called encoding.

User Query:

We start by asking user a query and convert these words/sentences to vector for further computation.

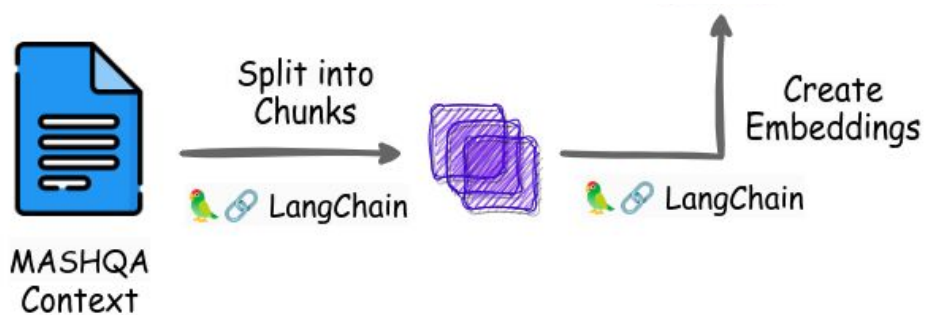
# Pipeline



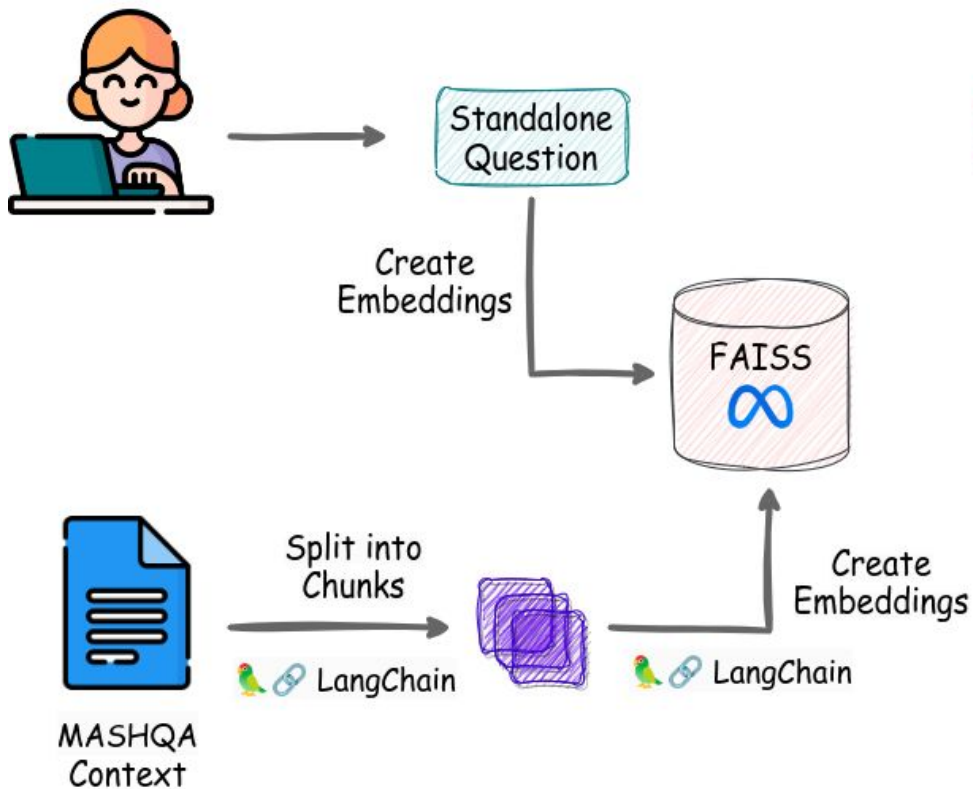
Using a contextual dataset along with query helps to get domain aware answer.

Here, we divide a standalone database into chunks.

Why? Next slide!



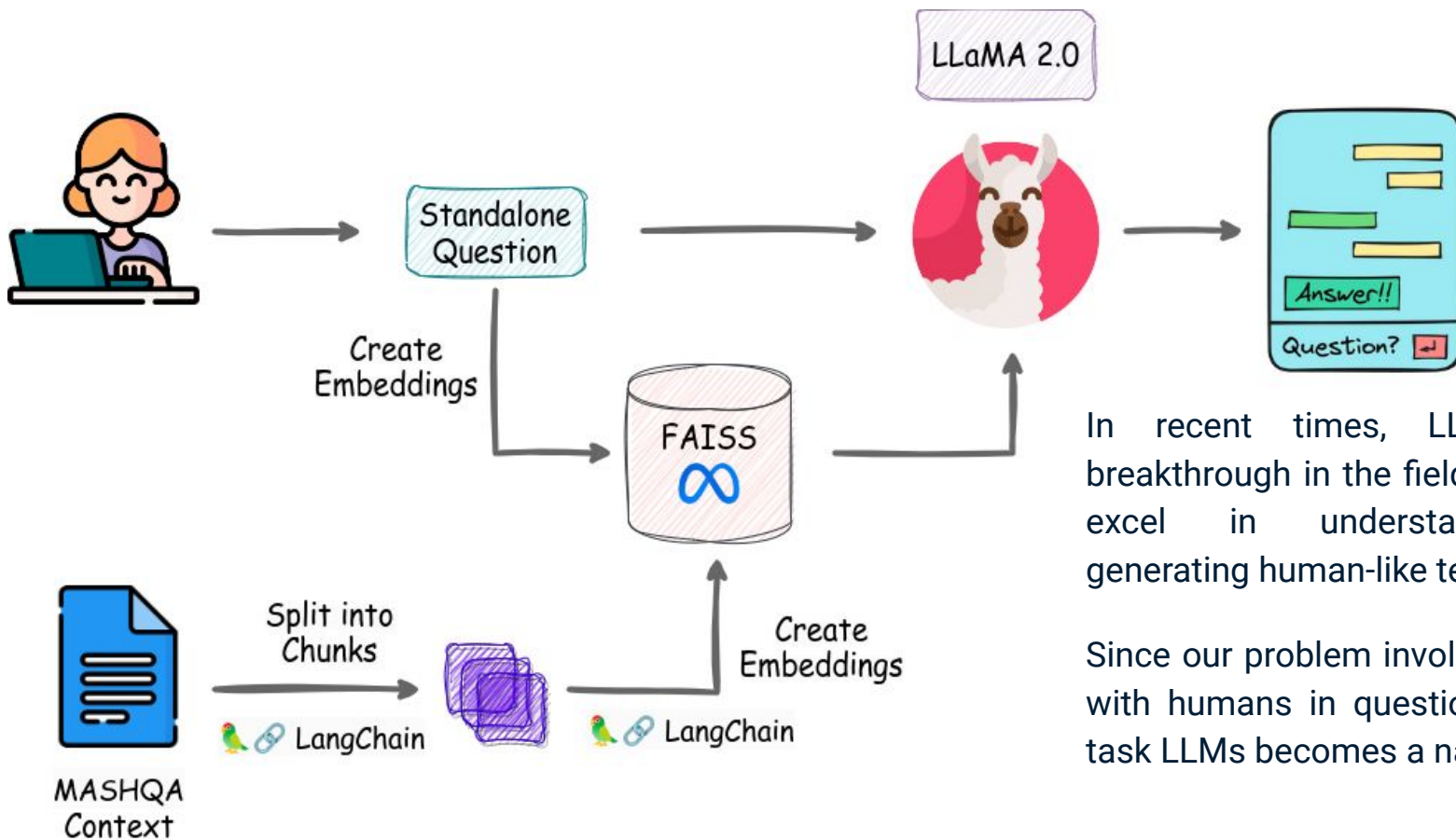
# Pipeline



FAISS, a open-source library from Facebook, that allows us to quickly search for similar embeddings of documents and cluster high-dimensional vectors in feature space using various similarity matrices.

We are now calculating the similarity of our query with the set of contextual chunks to add the most similar chunk as a instruction along with query.

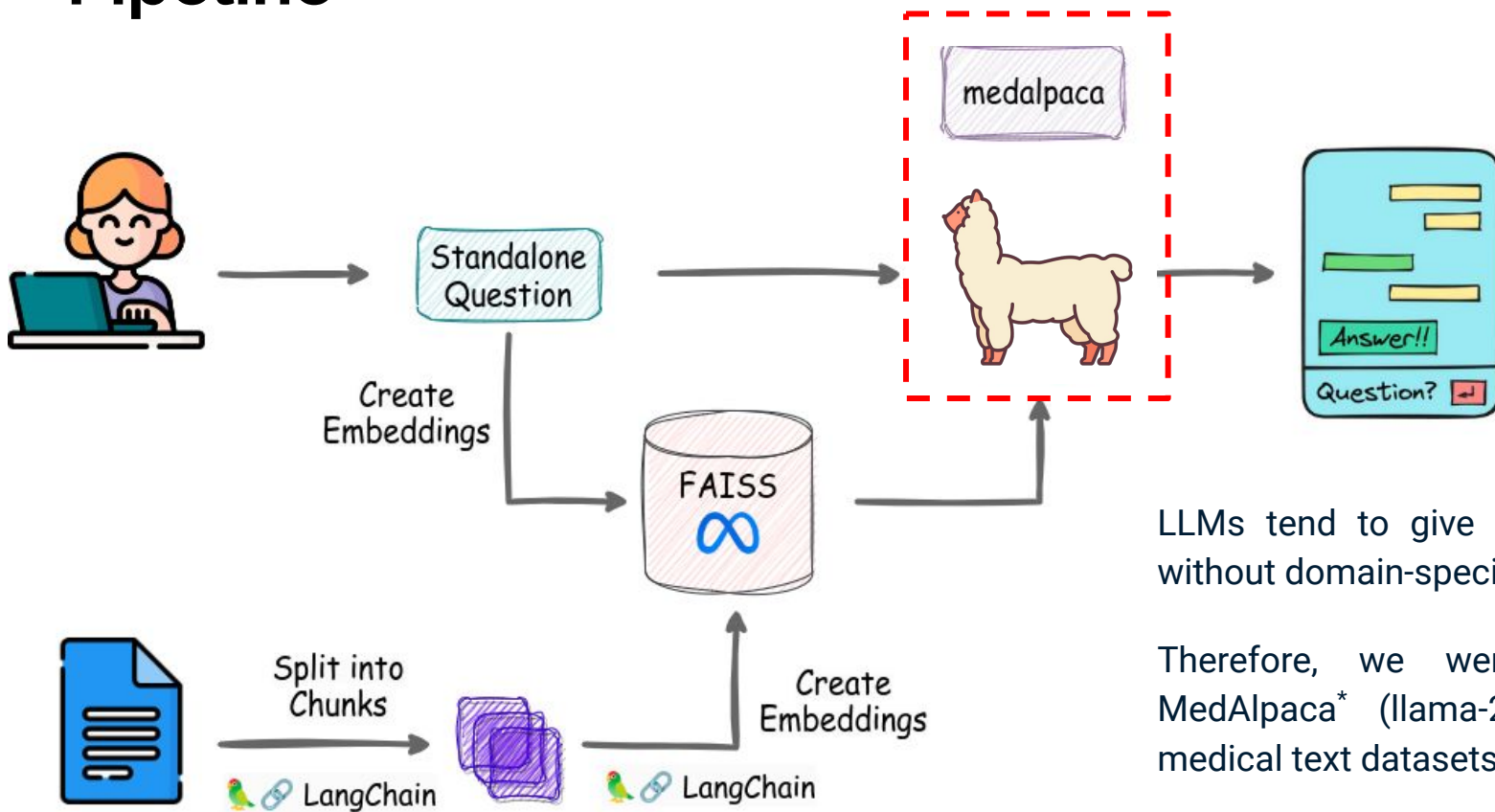
# Pipeline



In recent times, LLMs are a breakthrough in the field of NLP that excel in understanding and generating human-like text.

Since our problem involves engaging with humans in question answering task LLMs becomes a natural choice.

# Pipeline

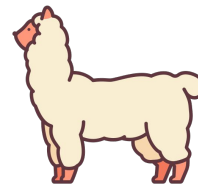


LLMs tend to give general answers without domain-specific fine-training.

Therefore, we went ahead with MedAlpaca\* (llama-2 fine-tuned on medical text datasets).

# LLMs Used

1. LLaMA 2 (meta-llama/Llama-2-7b-chat-hf)
  - 7b parameters
  - with/without vectorstore
2. Medalpaca (medalpaca/medalpaca-7b)
  - LLaMA 2 backbone
  - fine-tuned on medical data
  - with/without vectorstore



## NOTE:

- We weren't able to finetune either model on MASHQA dataset due to memory limitations
- Only HuggingFace 🤗 models were used



# Model Compression

## **Knowledge Distillation:**

Initially we aimed to fine-tune a baby LLM (small size) from existing larger size medical text LLMs with the help of knowledge distillation. But, we weren't able to do so due to memory limitation.

## **Pruning:**

Pruning reduces model complexity and saves memory by removing unnecessary parameters, improving efficiency and speeding up inference in machine learning models. Unlike traditional methods, a recent approach, Wanda\*, requires no retraining and offering computational efficiency in using the pruned LLM.

## **Quantization:**

Quantization reduces the memory footprint (e.g., 4,5,6 8-bit integers) of the model. In large models, it is essential for efficient deployment, reducing memory and storage usage, improving inference speed, and ensuring compatibility with various platforms and devices.

\* Sun, Mingjie, et al. "A Simple and Effective Pruning Approach for Large Language Models." *arXiv preprint arXiv:2306.11695* (Oct 2023).

# Quantization

- GGML employed for model quantization.
  - C/C++ library for efficient Llama model inference.
  - Capable of loading GGML models for CPU execution.
- Strategies
  - **q4\_k\_m**: Half attention.wv & feed\_forward.w2 use Q6\_K, rest use Q4\_K.
  - **q5\_k\_m**: Half attention.wv & feed\_forward.w2 use Q6\_K, rest use Q5\_K.
- Memory Footprint
  - **q4\_k\_m**: 13.5 GB → 4 GB
  - **q5\_k\_m**: 13.5 GB → 4.5 GB

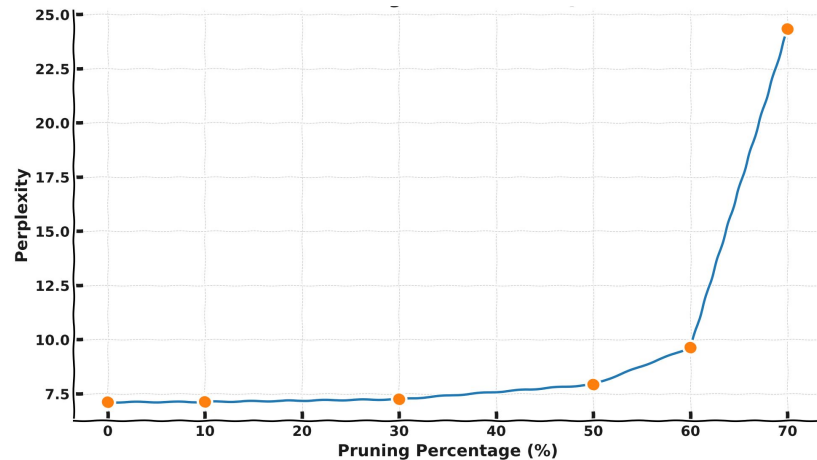


Fig 1: Pruning effect on SLiMQ

# Quantization

## Llama 2 7b chat

**User** : What should I do if I want to stop dialysis?

**System** : Stop hospice care and start back on treatments. You can also keep up these habits to stay well:

1. Eat right. Stick to a healthy, well-balanced diet []
2. Exercise regularly. Doctors often recommend exercise for people with kidney disease []
3. Don't smoke. Smoking raises blood pressure and can damage blood vessels, []
4. Limit alcohol to 2 drinks a day if you are an adult age 18
5. Don't take too many nonsteroidal anti-inflammatory drugs (NSAIDs), such as ibuprofen or naproxen []
6. Get enough sleep. Most adults need 7 to 8 hours of sleep each night.

## Quantised Llama 2 7b chat + vectorstore

**User** : What should I do if I want to stop dialysis?

**System** : Here are some steps you can take if you want to stop dialysis:

1. Discuss your decision with your nephrologist or kidney doctor: It is important to discuss your decision []
1. Review your medical history: Your nephrologist []
2. Check your blood chemistry and other test results: Your nephrologist or kidney doctor may perform additional tests
3. Consider hospice care: If you have a terminal illness and are not likely to recover, your nephrologist []
4. Make arrangements for end-of-life care:
5. Seek support from family and friends:
6. Consider alternative options: Depending on your specific situation, there may be alternative options available to help manage your kidney failure.

It is important to note that stopping dialysis is a complex decision that should be made with the guidance of your healthcare provider and support from loved ones. It is also important to consider your personal preferences and values when making this decision.



## Smart Lightweight Medical Query System

### What is your question?

What should I do if I want to stop dialysis?

#### SECTION - LLM Input

Temperature

0.4

Accelerator

GPU

Repetition Penalty

1.2

Max. New Tokens

512

Quantization Factor

Vanilla

Top P

0.9

Generate Answer

### Answers!

Normal LLM:

Stop hospice care and start back on treatments. You can also keep up these habits to stay well:

1. Eat right. Stick to a healthy, well-balanced diet low in salt, fat, and sugar; drink plenty of water; and avoid foods high in phosphorus.
2. Exercise regularly. Doctors often recommend exercise for people with kidney disease because it can help keep your body healthy.
3. Don't smoke. Smoking raises blood pressure and can damage blood vessels, which can make kidney disease worse.
4. Limit alcohol to 2 drinks a day if you are an adult age 18 or older; the limit is lower for children and teens.
5. Don't take too many nonsteroidal anti-inflammatory drugs (NSAIDs), such as ibuprofen or naproxen, unless your doctor says it's okay. These meds can harm the kidneys if you take too much of them.
6. Get enough sleep. Most adults need 7 to 8 hours of sleep each night.

Follow these steps and talk with your healthcare provider about any changes in your lifestyle or treatment plan. And remember, it's important to keep following your doctor's advice so that you can stay well and prevent

Medical LLM:

Here are some steps you can take if you want to stop dialysis:

1. Discuss your decision with your nephrologist or kidney doctor: It is important to discuss your decision with your nephrologist or kidney doctor before stopping dialysis. They can help you understand the potential risks and benefits of stopping dialysis and provide guidance on how to do so safely.
2. Review your medical history: Your nephrologist or kidney doctor will review your medical history to determine if stopping dialysis is safe for you. They may consider factors such as your age, overall health, and the cause of your kidney failure.
3. Check your blood chemistry and other test results: Your nephrologist or kidney doctor may perform additional tests to check your blood chemistry and other vital signs before stopping dialysis. This can help ensure that you are in a stable condition and that stopping dialysis will not cause any immediate harm.
4. Consider hospice care: If you have a terminal illness and are not likely to recover, your nephrologist or kidney doctor may recommend hospice care. Hospice care provides comfort measures and supportive care to help you manage your symptoms and improve your quality of life.

**Thank You**