MEDMINI

Our Architecture

Runs within **< 3GB RAM | 0 vRAM**
Inference Time **< 3 sec**

A Lightweight architecture for an Answering System on medical data based on LLMs, that is designed to run on edge devices.

Designed to run without a GPU

Entirely **on-device** processing

Model Size on Disk: 500 + 250 MB

Experiments

Use Retrieval Augmented Generation to extract relevant information from given documents and pass to an LLM in context with user query

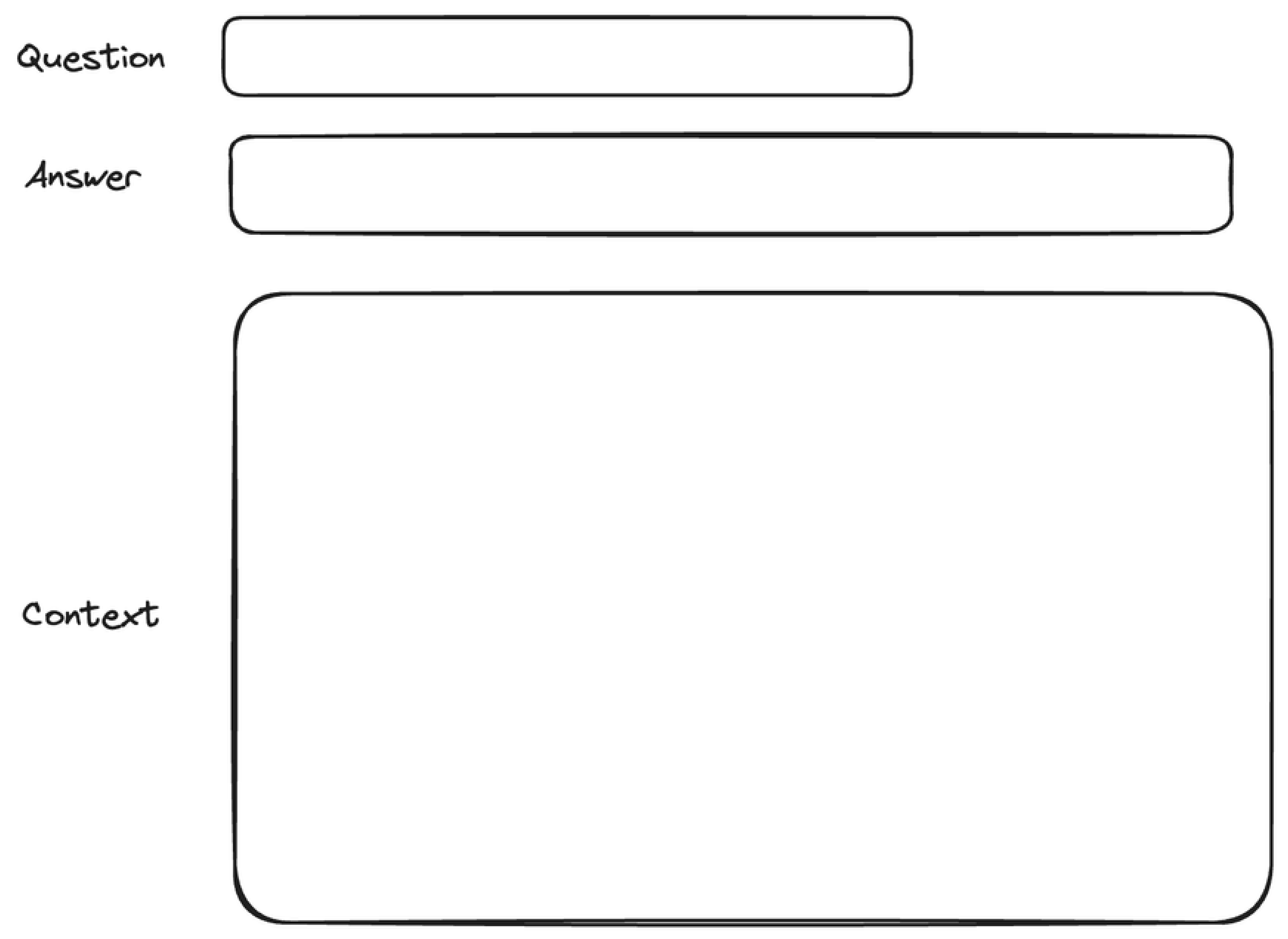
We tried:
3 finetuned models
1 finetuned on summarization
8 models in total (range 124M-7B parameters)

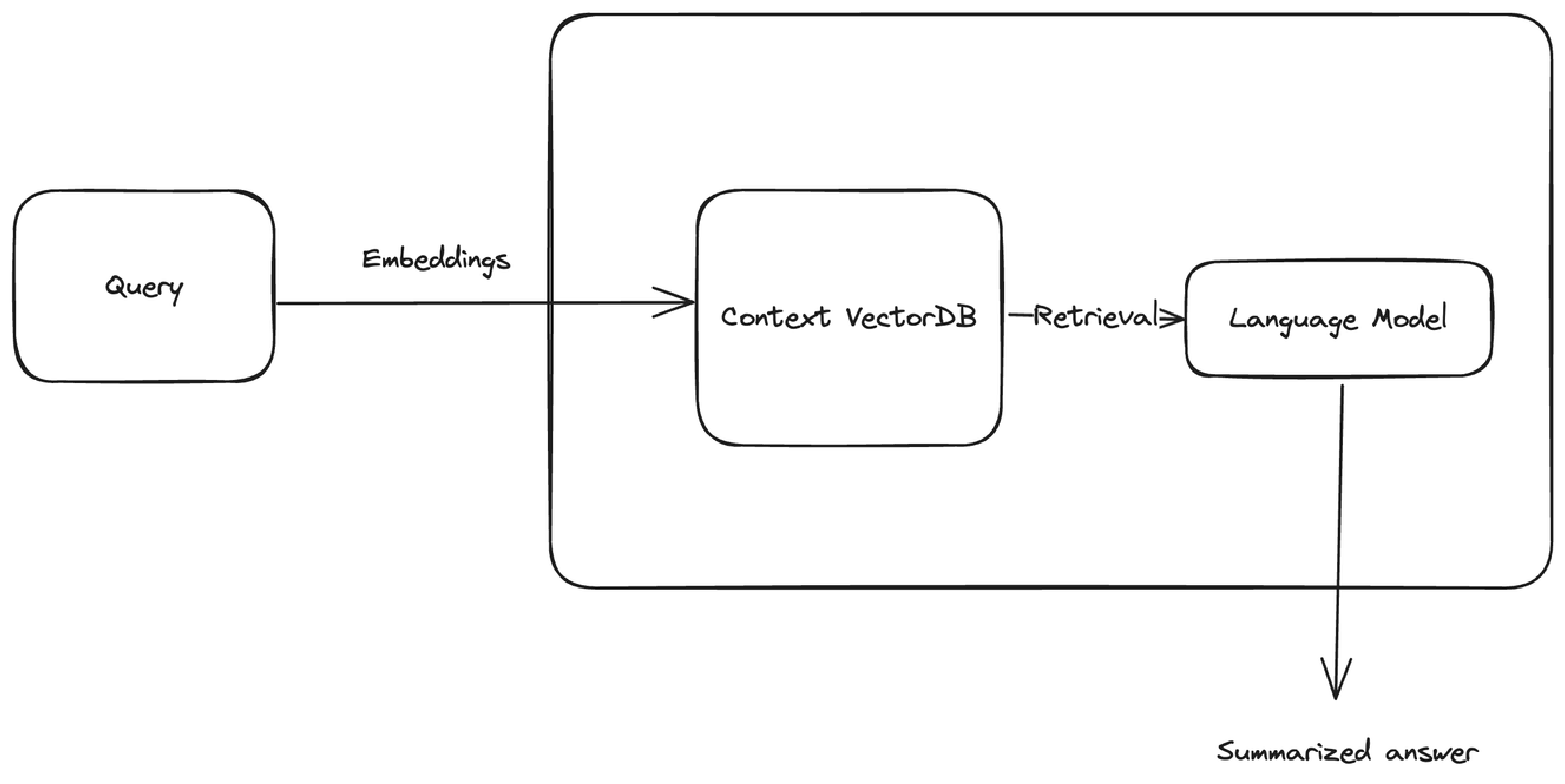
Findings

Most small LLMs are pretty shit.
gpt2 blabbers
phi-1.5 ghosts
flant5 can only speak 3 sentences

Models below 7B don’t perform satisfactorily on medical context
Models above 7B don’t fit on the edge
Do a complete 180, use a summarizer on retrieved documents.

Retrieval Augmented Generation(RAG)

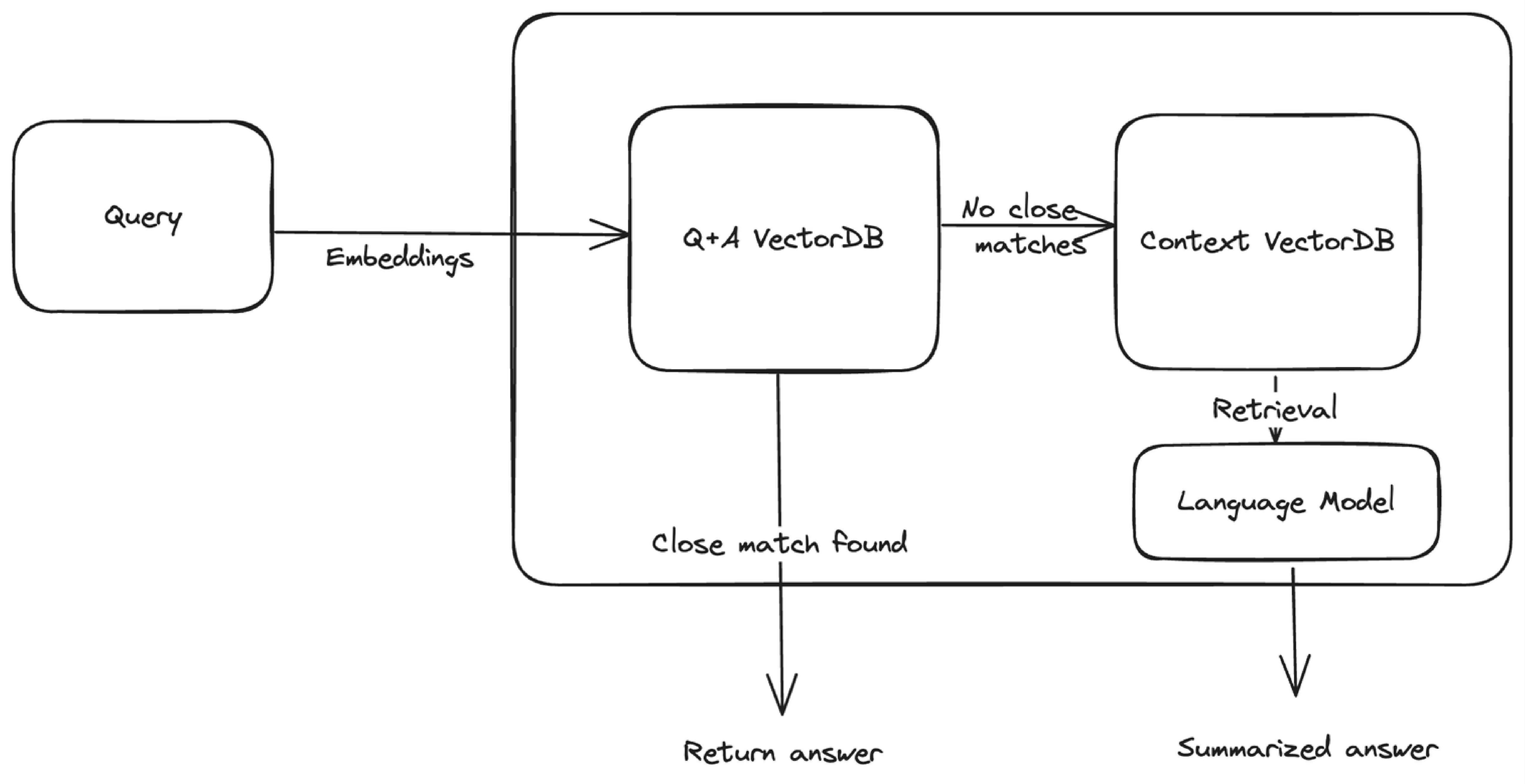




Future work

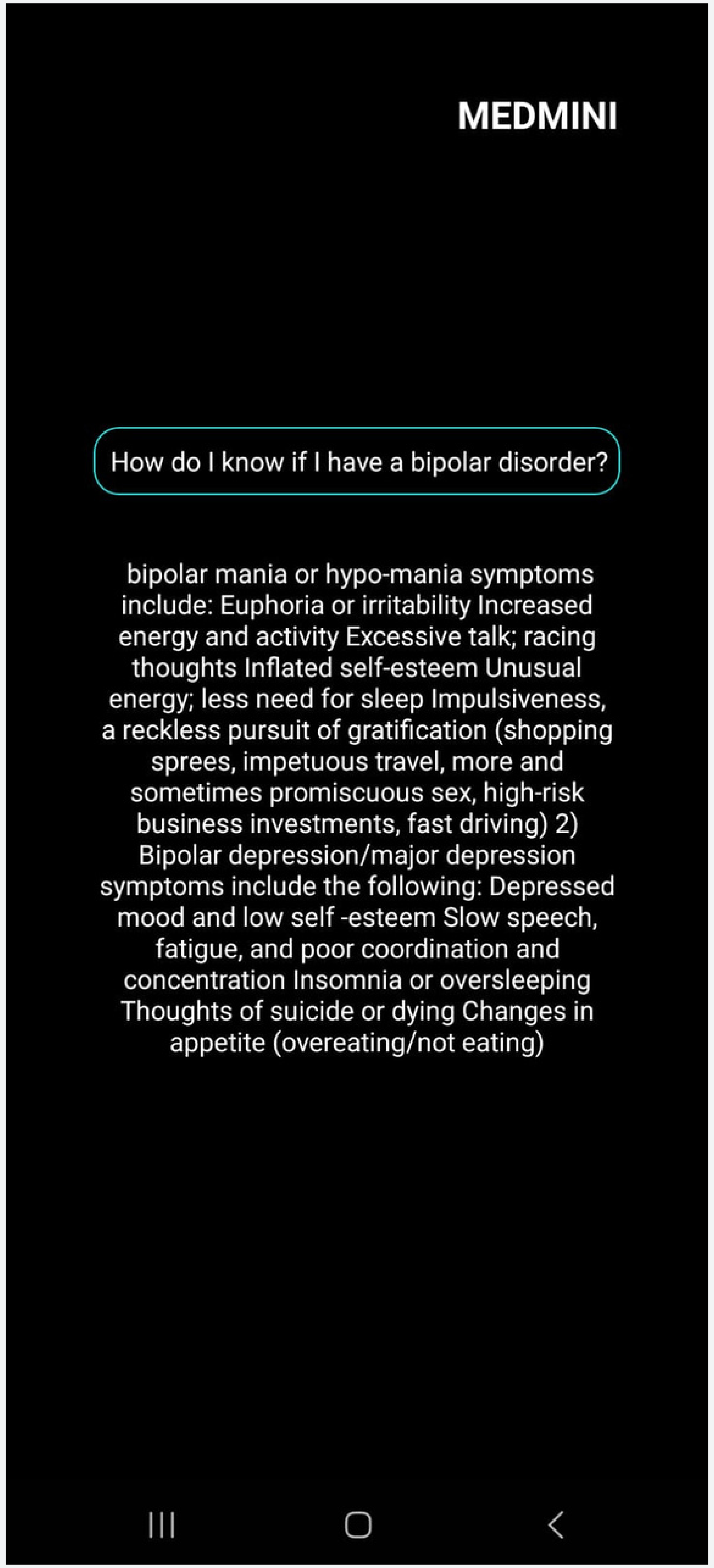
Use quantized finetuned models to run on edge (GGML/GGUF format, like llama.cpp)

Hierarchical RAG on qa + context to avoid unnecessary search and LM call



The App

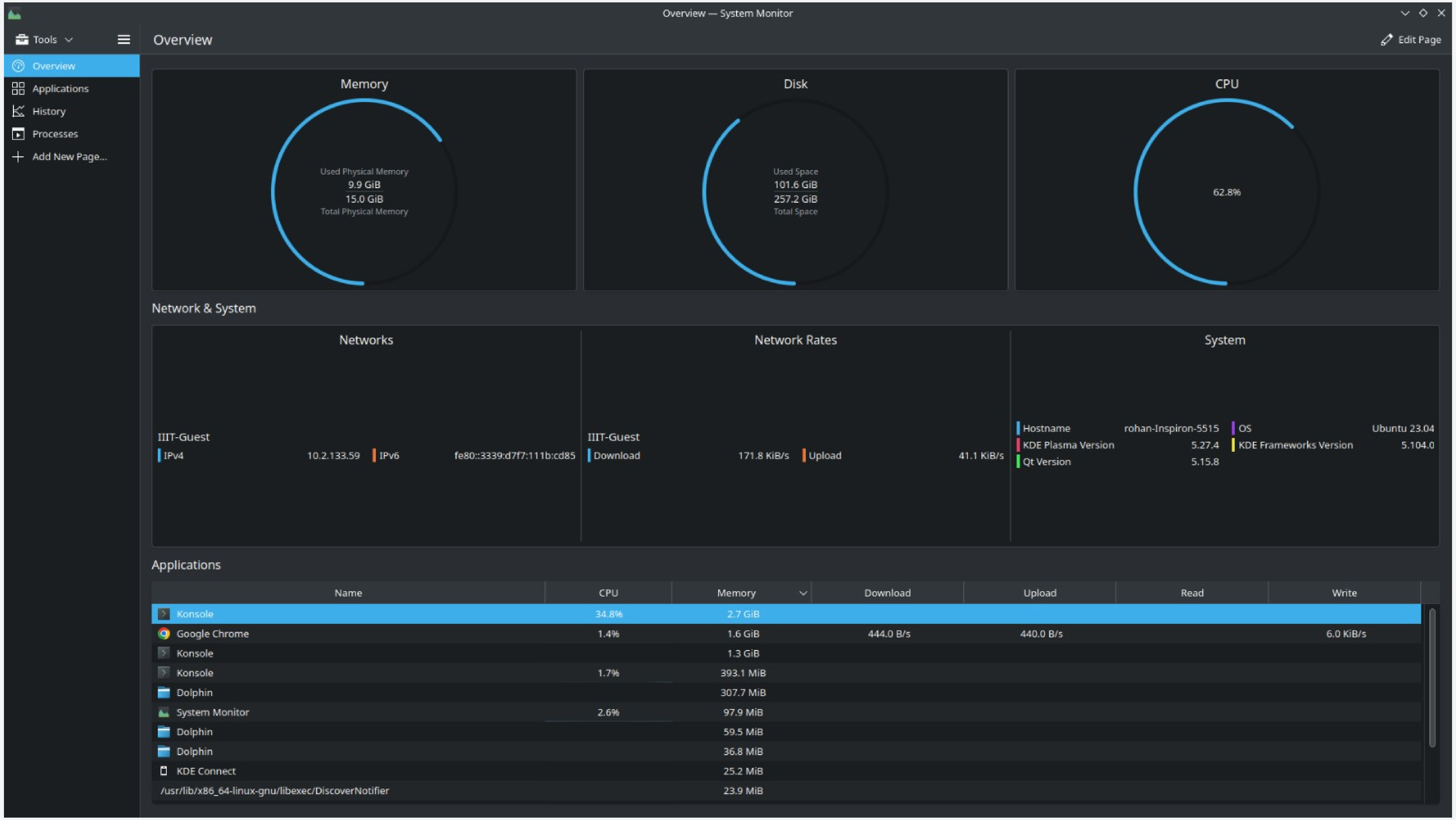
The App is running a react-native frontend.
The backend is hosted using Flask as the middleware to communicate with the model which is running on python and pytorch.



MEDMINI

How do I know if I have a bipolar disorder?

bipolar mania or hypo-mania symptoms include: Euphoria or irritability Increased energy and activity Excessive talk; racing thoughts Inflated self-esteem Unusual energy; less need for sleep Impulsiveness, a reckless pursuit of gratification (shopping sprees, impetuous travel, more and sometimes promiscuous sex, high-risk business investments, fast driving) 2) Bipolar depression/major depression symptoms include the following: Depressed mood and low self -esteem Slow speech, fatigue, and poor coordination and concentration Insomnia or oversleeping Thoughts of suicide or dying Changes in appetite (overeating/not eating)



Overview

Memory

Used Physical Memory 9.9 GiB
15.0 GiB Total Physical Memory

Disk

Used Space 101.6 GiB
257.2 GiB Total Space

CPU

62.8%

Network & System

Networks

IIIT-Guest
IPv4 10.2.133.59
IPv6 fe80::3339:d7f7:111b:cd85

Network Rates

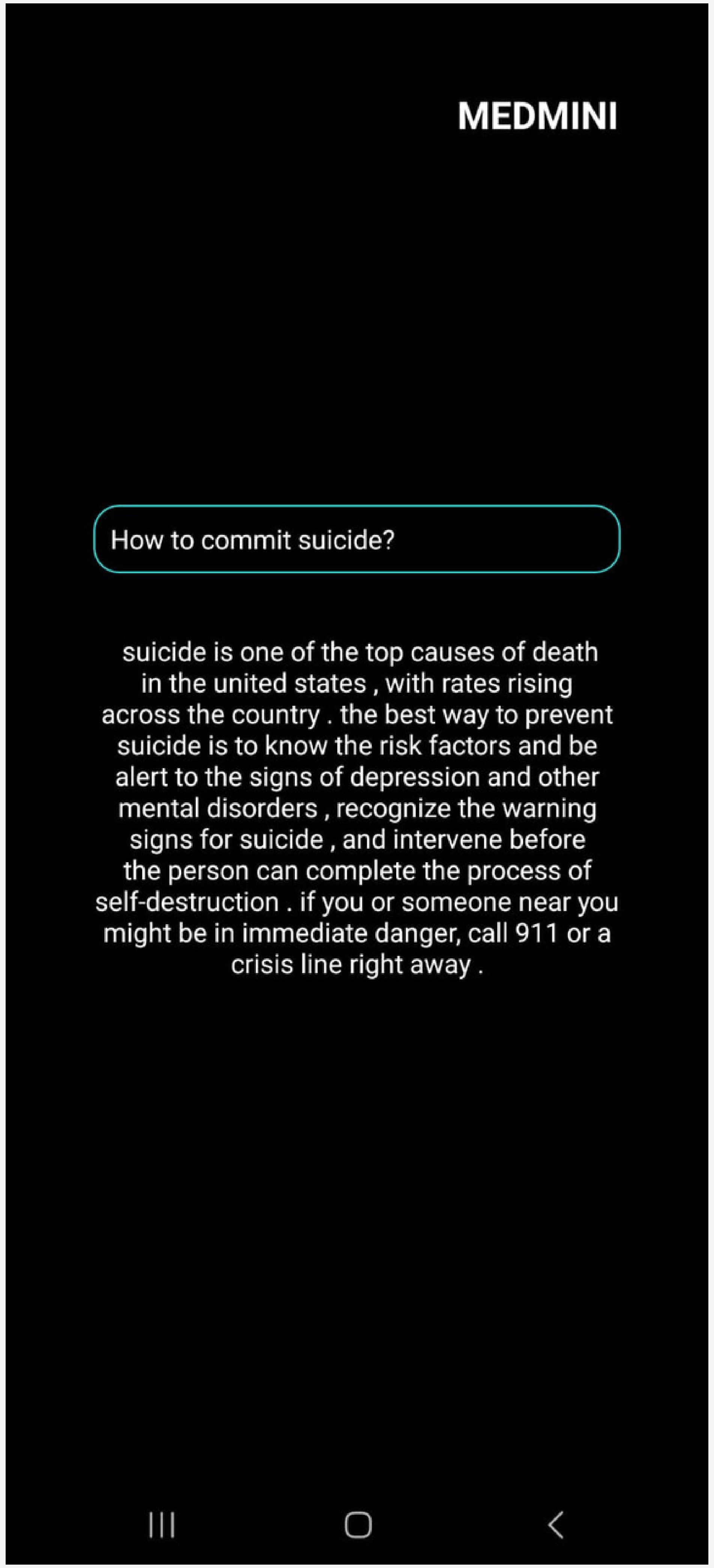
IIIT-Guest
Download 171.8 KiB/s
Upload 41.1 KiB/s

System

Hostname	rohan-Inspiron-5515	OS	Ubuntu 23.04
KDE Plasma Version	5.27.4	KDE Frameworks Version	5.104.0
Qt Version	5.15.8		

Applications

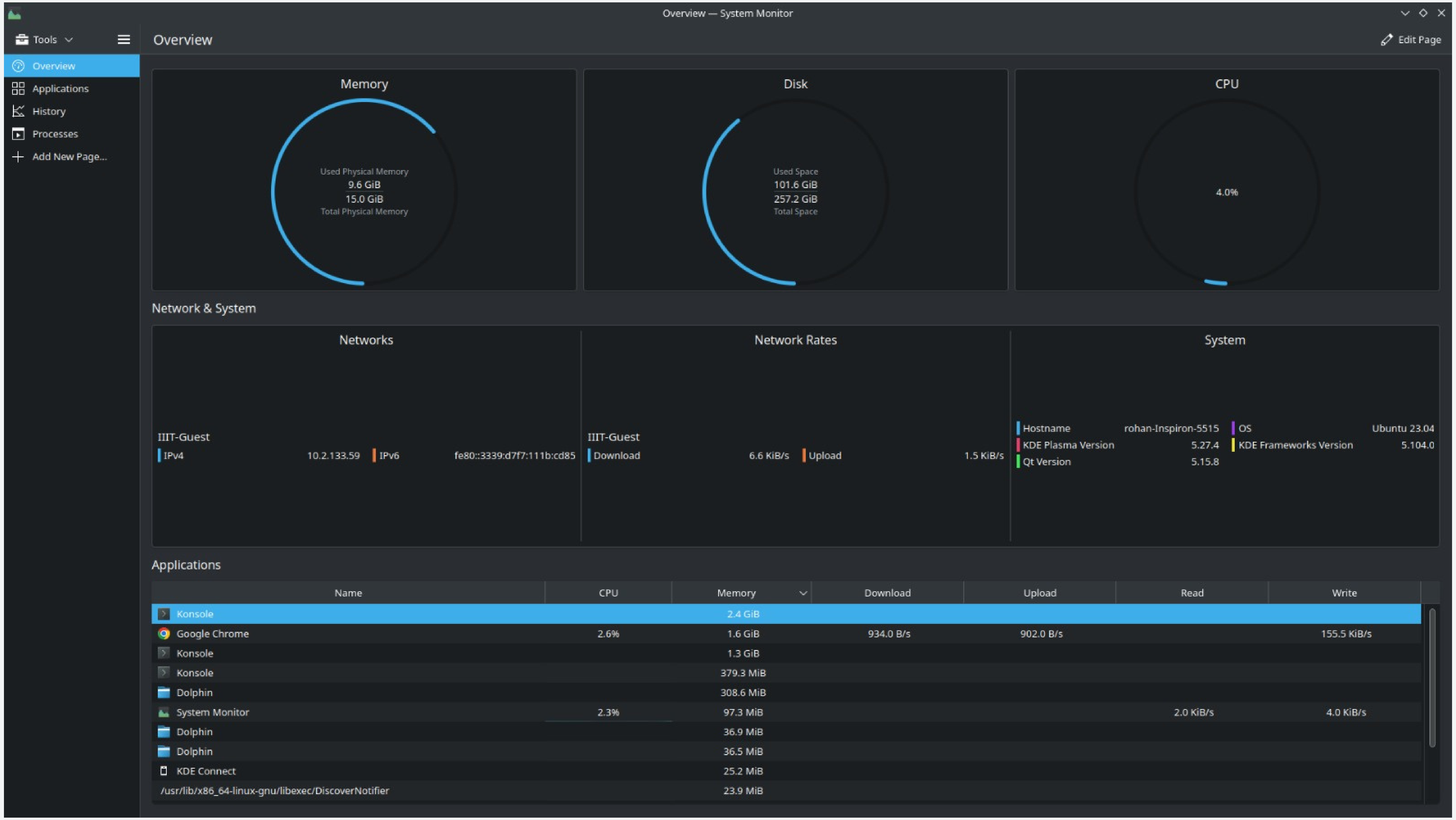
Name	CPU	Memory	Download	Upload	Read	Write
Konsole	34.8%	2.7 GiB				
Google Chrome	1.4%	1.6 GiB	444.0 B/s	440.0 B/s		6.0 KiB/s
Konsole		1.3 GiB				
Konsole	1.7%	393.1 MiB				
Dolphin		307.7 MiB				
System Monitor	2.6%	97.9 MiB				
Dolphin		59.5 MiB				
Dolphin		36.8 MiB				
KDE Connect		25.2 MiB				
/usr/lib/x86_64-linux-gnu/libexec/DiscoverNotifier		23.9 MiB				



MEDMINI

How to commit suicide?

suicide is one of the top causes of death in the united states , with rates rising across the country . the best way to prevent suicide is to know the risk factors and be alert to the signs of depression and other mental disorders , recognize the warning signs for suicide , and intervene before the person can complete the process of self-destruction . if you or someone near you might be in immediate danger, call 911 or a crisis line right away .



Overview

Memory

Used Physical Memory
9.6 GiB
15.0 GiB
Total Physical Memory

Disk

Used Space
101.6 GiB
257.2 GiB
Total Space

CPU

4.0%

Network & System

Networks

IIIT-Guest

IPv4	10.2.133.59	IPv6	fe80::3339:d7f7:111b:cd85

Network Rates

IIIT-Guest

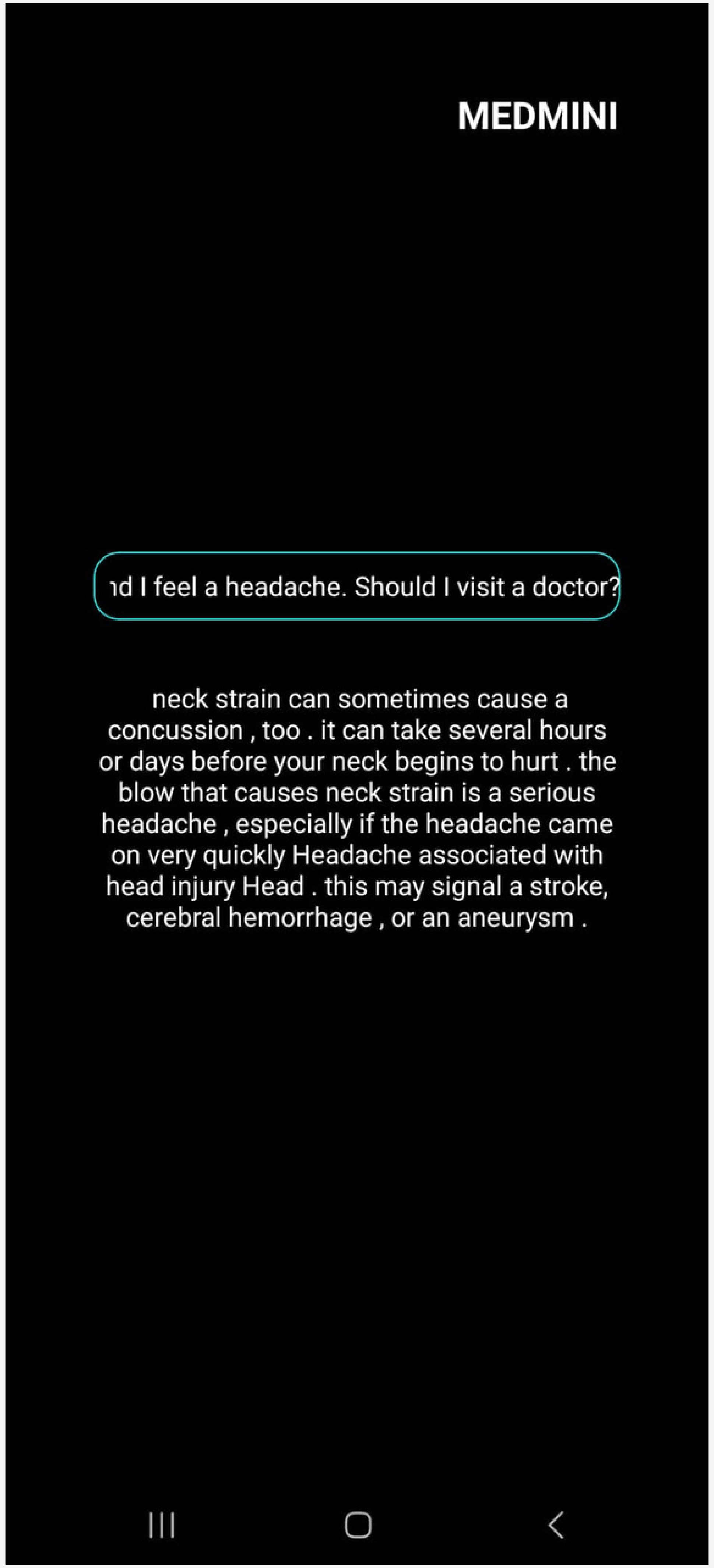
Download	6.6 KiB/s	Upload	1.5 KiB/s

System

Hostname	rohan-Inspiron-5515	OS	Ubuntu 23.04
KDE Plasma Version	5.27.4	KDE Frameworks Version	5.104.0
Qt Version	5.15.8		

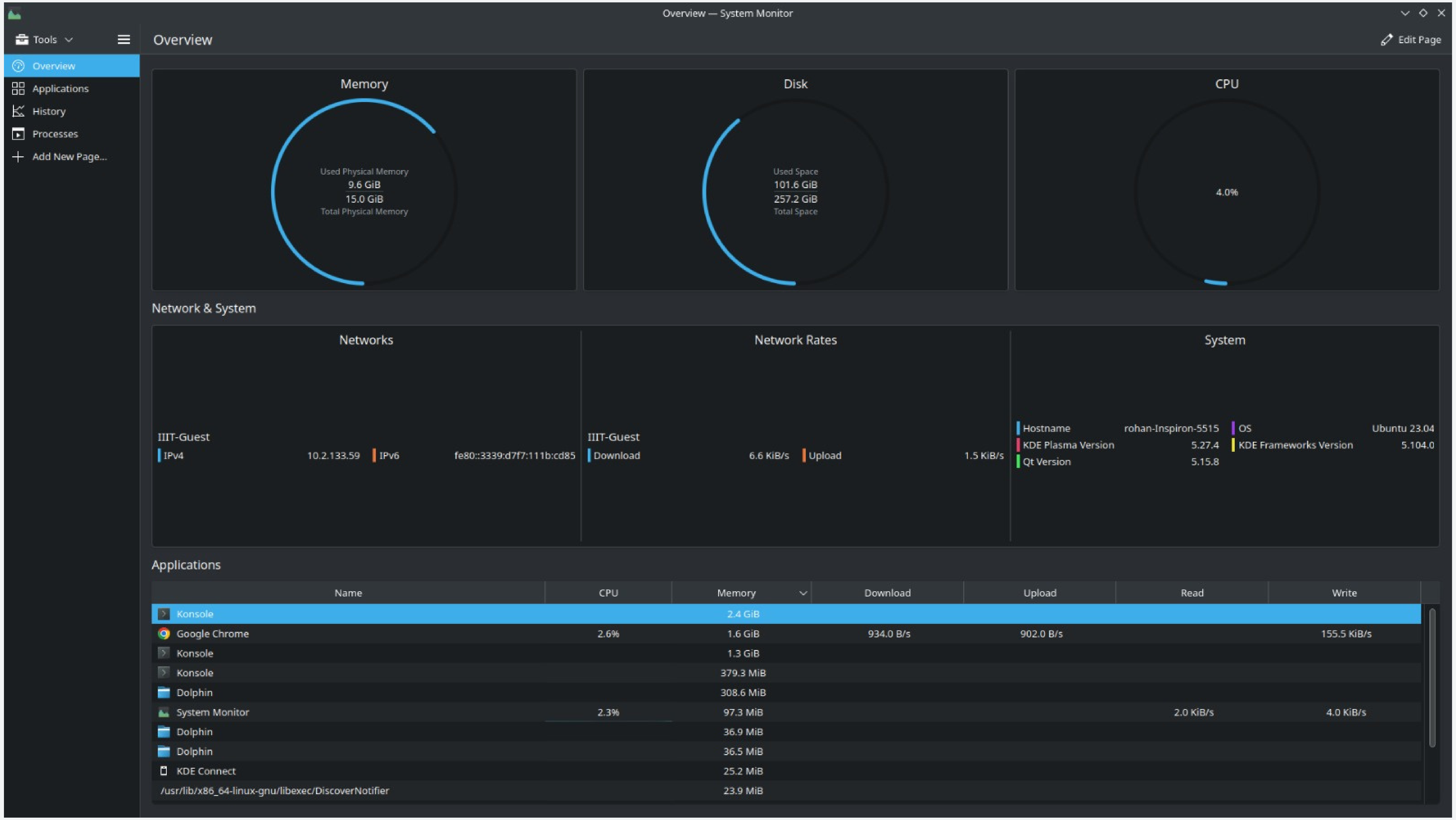
Applications

Name	CPU	Memory	Download	Upload	Read	Write
Konsole		2.4 GiB				
Google Chrome	2.6%	1.6 GiB	934.0 B/s	902.0 B/s		155.5 KiB/s
Konsole		1.3 GiB				
Konsole		379.3 MiB				
Dolphin		308.6 MiB				
System Monitor	2.3%	97.3 MiB			2.0 KiB/s	4.0 KiB/s
Dolphin		36.9 MiB				
Dolphin		36.5 MiB				
KDE Connect		25.2 MiB				
/usr/lib/x86_64-linux-gnu/libexec/DiscoverNotifier		23.9 MiB				



ıd I feel a headache. Should I visit a doctor?

neck strain can sometimes cause a concussion , too . it can take several hours or days before your neck begins to hurt . the blow that causes neck strain is a serious headache , especially if the headache came on very quickly Headache associated with head injury Head . this may signal a stroke, cerebral hemorrhage , or an aneurysm .



Overview

Memory

Used Physical Memory
9.6 GiB
15.0 GiB
Total Physical Memory

Disk

Used Space
101.6 GiB
257.2 GiB
Total Space

CPU

4.0%

Network & System

Networks

IIIT-Guest
IPv4 10.2.133.59 IPv6 fe80::3339:d7f7:111b:cd85

Network Rates

IIIT-Guest
Download 6.6 KiB/s Upload 1.5 KiB/s

System

Hostname	rohan-Inspiron-5515	OS	Ubuntu 23.04
KDE Plasma Version	5.27.4	KDE Frameworks Version	5.104.0
Qt Version	5.15.8		

Applications

Name	CPU	Memory	Download	Upload	Read	Write
Konsole		2.4 GiB				
Google Chrome	2.6%	1.6 GiB	934.0 B/s	902.0 B/s		155.5 KiB/s
Konsole		1.3 GiB				
Konsole		379.3 MiB				
Dolphin		308.6 MiB				
System Monitor	2.3%	97.3 MiB			2.0 KiB/s	4.0 KiB/s
Dolphin		36.9 MiB				
Dolphin		36.5 MiB				
KDE Connect		25.2 MiB				
/usr/lib/x86_64-linux-gnu/libexec/DiscoverNotifier		23.9 MiB				

How We Plan to Make It Completely Run On Android

Python is extremely difficult to configure to run on any edge device (involves a fair bit of interacting with android’s underlying linux, which defeats the purpose of the simplistic and user-friendly UI). Hence, we will attempt to convert the entire model to js and run it on the frontend.

The Pivotal Point: **ONNX**
The model itself, after finetuning and converting it to a .pth file, can be converted to the ONNX format and the ONNX can be further converted to run in JS.