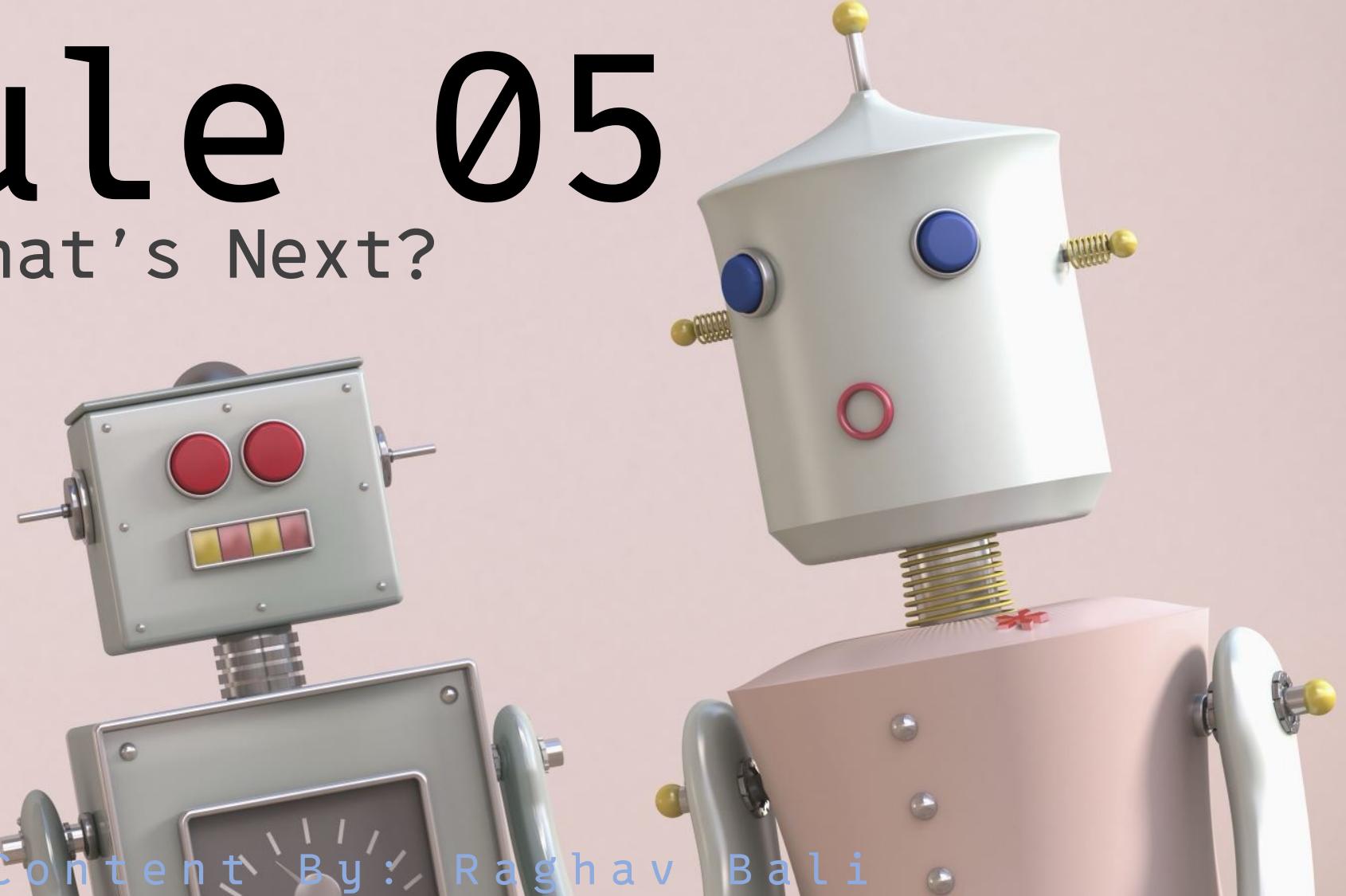


Module 05

What's Next?



Content By : Raghav Bali

What's Next?

Optimizations



Perspectives



Maturity



Content By: Raghav Ballal

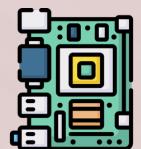
Optimizations



Optimizations



Architectural

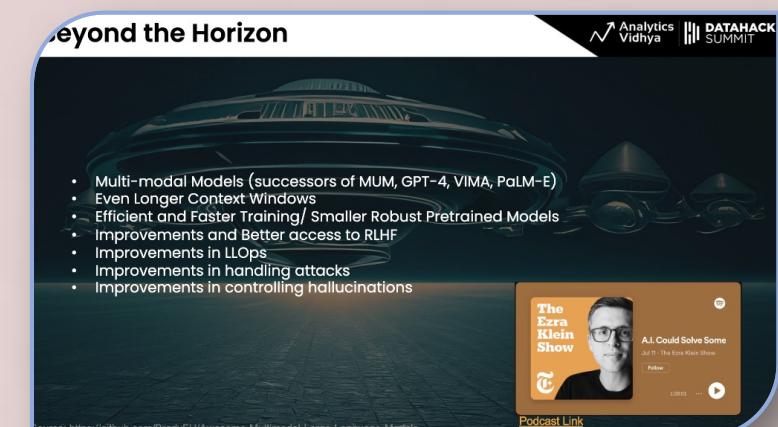


Hardware / Software

Optimizations

Architectural

- GeLU, Flash Attention, Mamba, BASED, MiniSequence ...



Optimizations

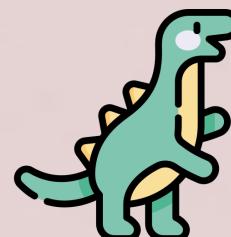


Architectural

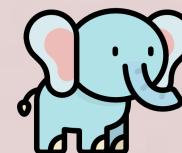
- Model Size V/s Performance Efficiency
 - SLMs have entered the arena



GPT



GPT-4

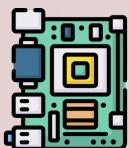


Phi-2



On-Device?

Optimizations



Hardware / Software



[tinygrad](#)

We write and maintain [tinygrad](#), the fastest growing neural network framework (over 23,000 GitHub stars)

It's extremely simple, and breaks down the most [complex networks](#) into 3 [OpTypes](#)

ElementwiseOps are UnaryOps, BinaryOps, and TernaryOps.
They operate on 1-3 tensors and run elementwise.

example: SQRT, LOG2, ADD, MUL, WHERE, etc...

ReduceOps operate on one tensor and return a smaller tensor.
example: SUM, MAX

MovementOps are virtual ops that operate on one tensor and move the data around
Copy-free with [ShapeTracker](#).
example: RESHAPE, PERMUTE, EXPAND, etc...

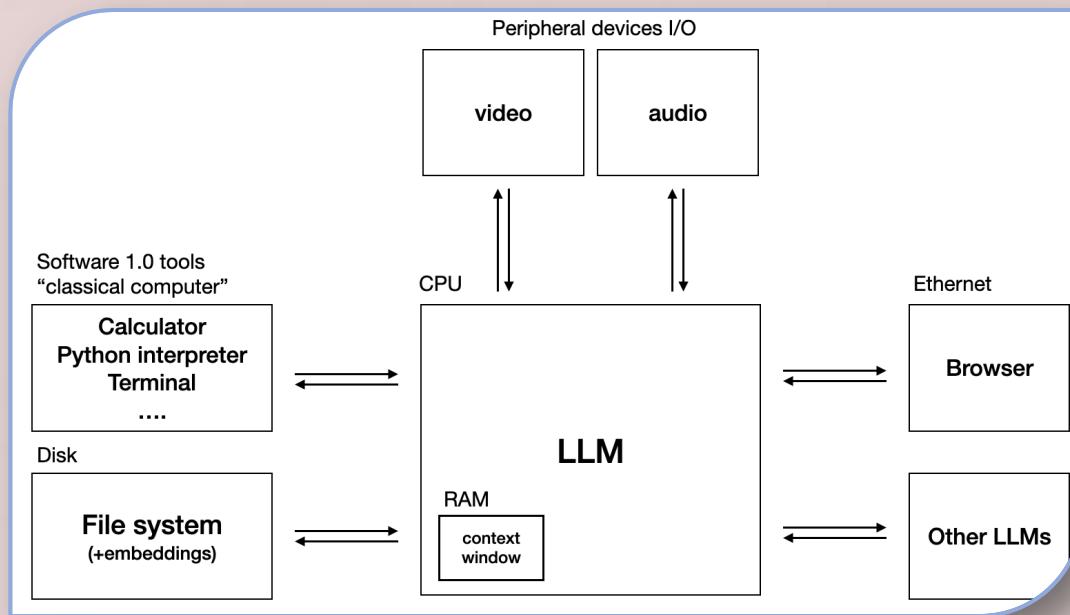
Want to know...where are your CONVs and MATMULS? Read the code to solve this mystery.



Perspectives

Content By: Raghav Bali

LLM-OS



Source: Andrey Karpathy/[tweet](#)

AI Products



Rabbit R1



AI Pin

Search-GPT?

The image displays three search engine interfaces side-by-side, highlighting how AI is changing search results.

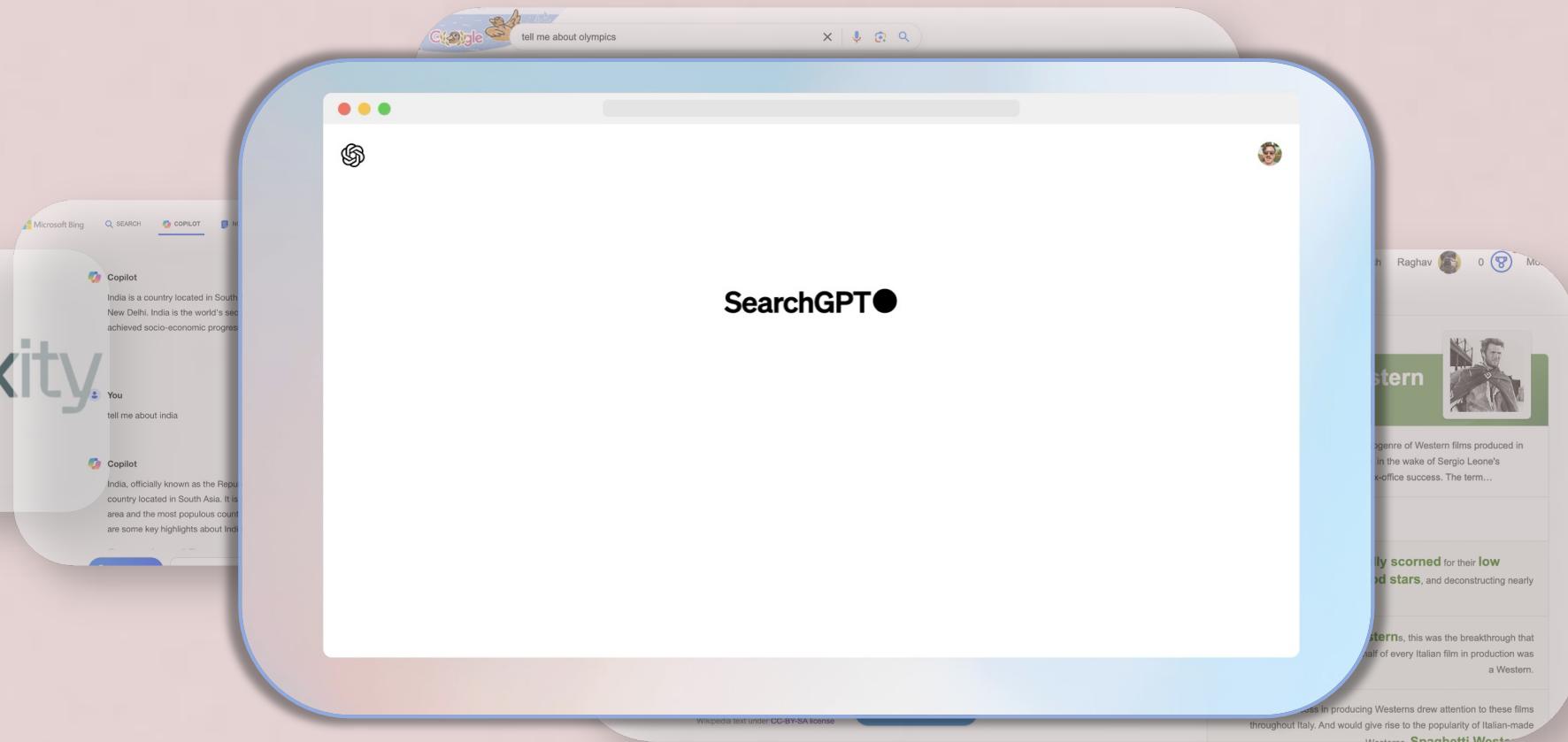
- Google Search:** A search for "tell me about olympics" shows a summary card from "Search Labs | AI Overview" detailing the history of the Olympic Games. It includes a link to the official site and a "Learn more" button.
- Microsoft Bing:** A search for "what is a spaghetti western" uses GPT-powered results. It shows a summary card for "Spaghetti Western - Wikipedia" with a photo of Clint Eastwood and links to "Terminology", "Production", and "People mentioned".
- Perplex:** A search for "tell me about india" shows results from "Copilot". It includes a summary card for India with a photo of the Taj Mahal and links to "Content" (Terminology, Production, Context and so...), "Further develo...", "Other notable t...", "Reception", and "Legacy".

Content By: Raghav Bali

Search-GPT ?



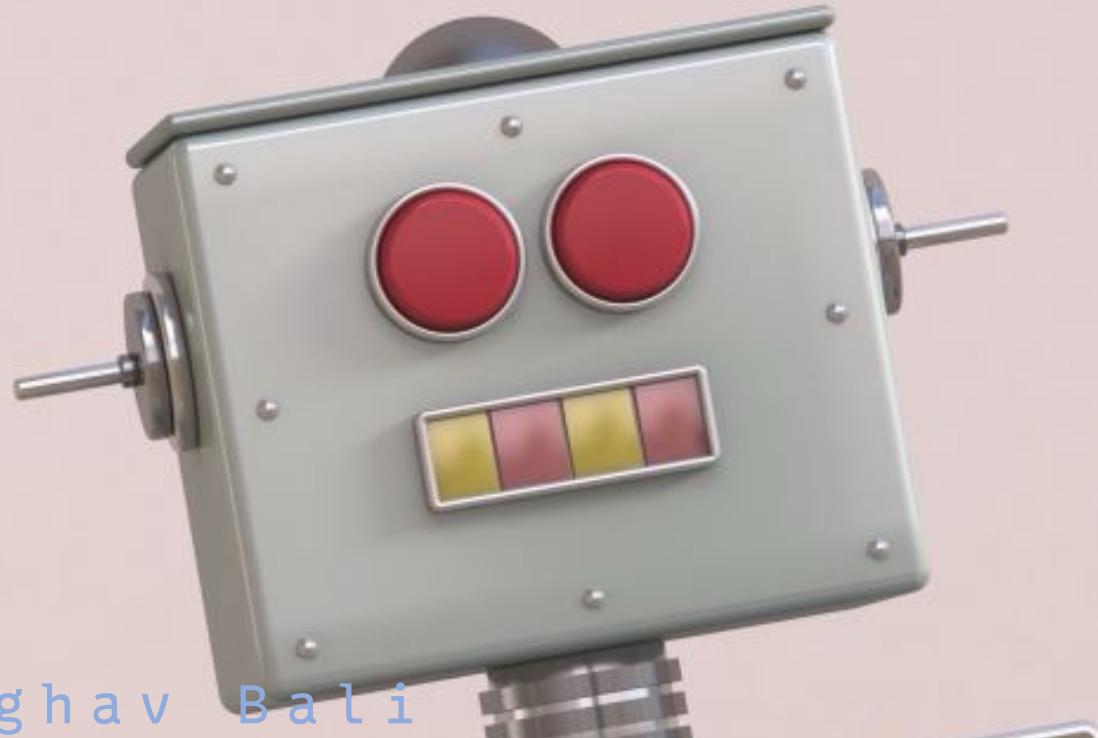
perplexity



Source: <https://openai.com/index/searchgpt-prototype/>

Content By: Raghav Bali

Maturity



Content By: Raghav Bali

Maturity

LLMOps



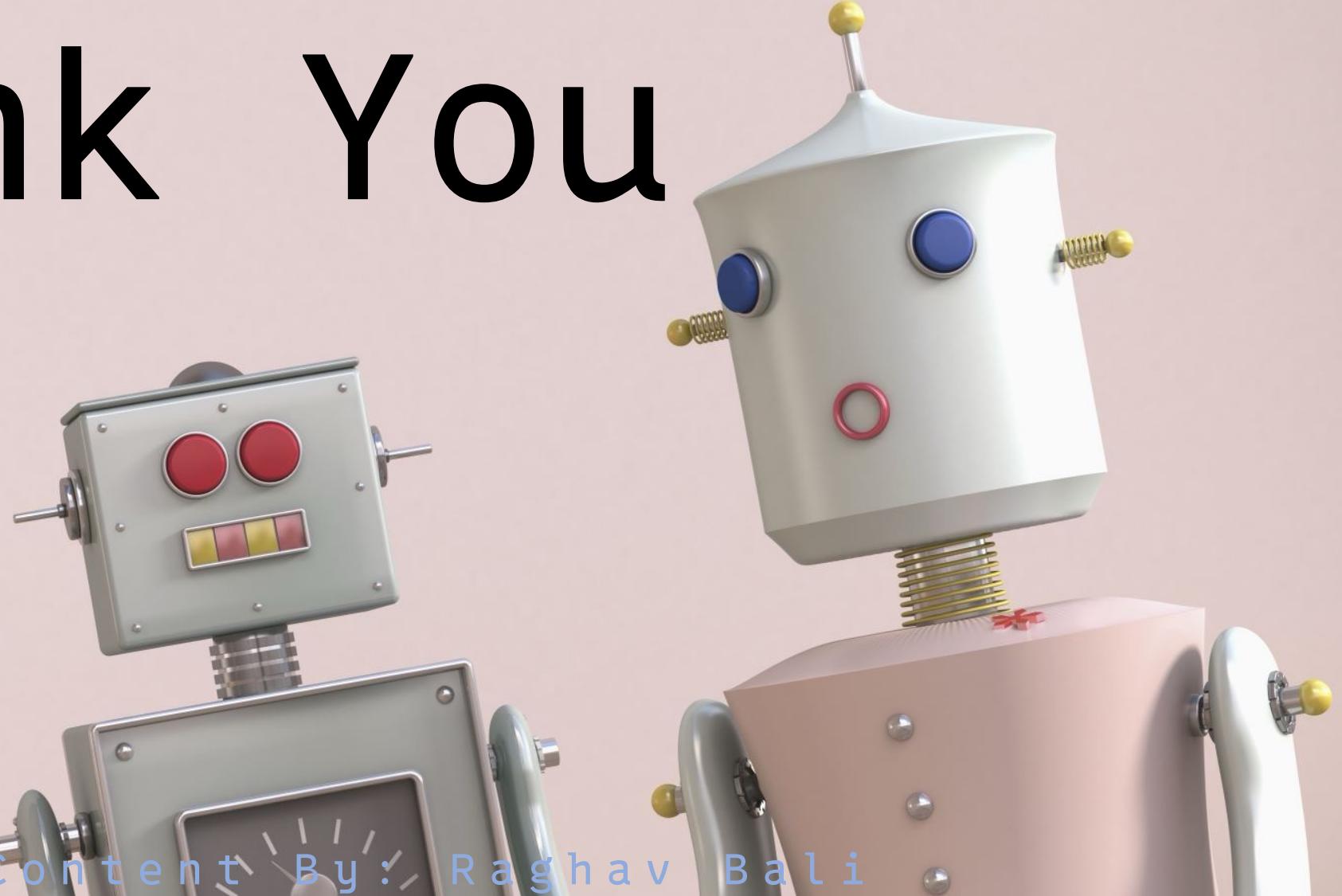
Llama Guard 3

SynthID

Hey there! I'm Devin and I'm a software engineer.

Enter a coding task below to get started.

Thank You



Content By : Raghav Bali