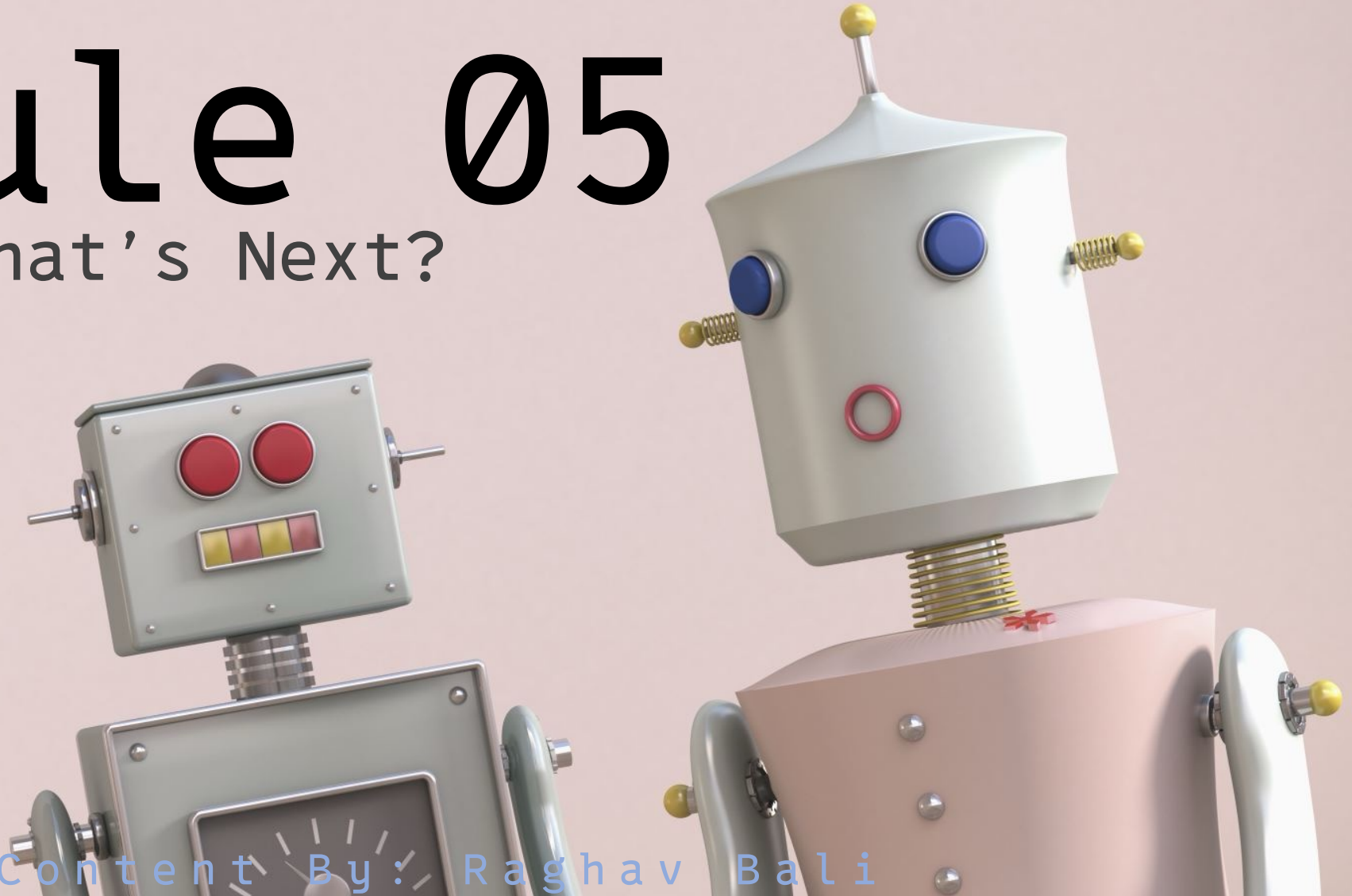
Module 05

What's Next?



Content By: Raghav Bali

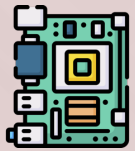
What's Next?

Optimizations

Content By: Raghav Bali

Optimizations

Architectural



Hardware / Software

Content By: Raghav Bali

Optimizations

Architectural

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

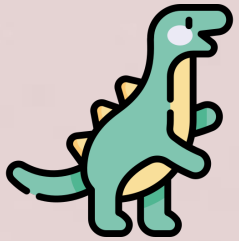
Content By: Raghav Bali

Optimizations

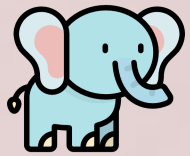
Architectural

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

GPT



GPT-4

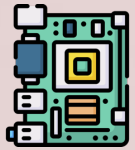


Phi-2

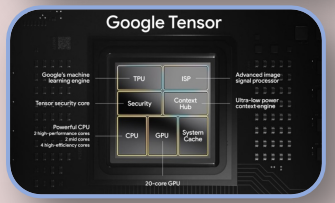
On-Device?

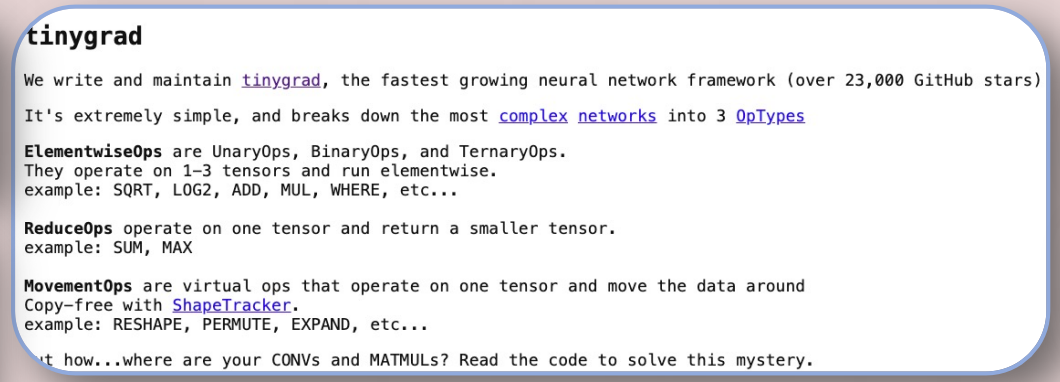
Content By: Raghav Bali

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...

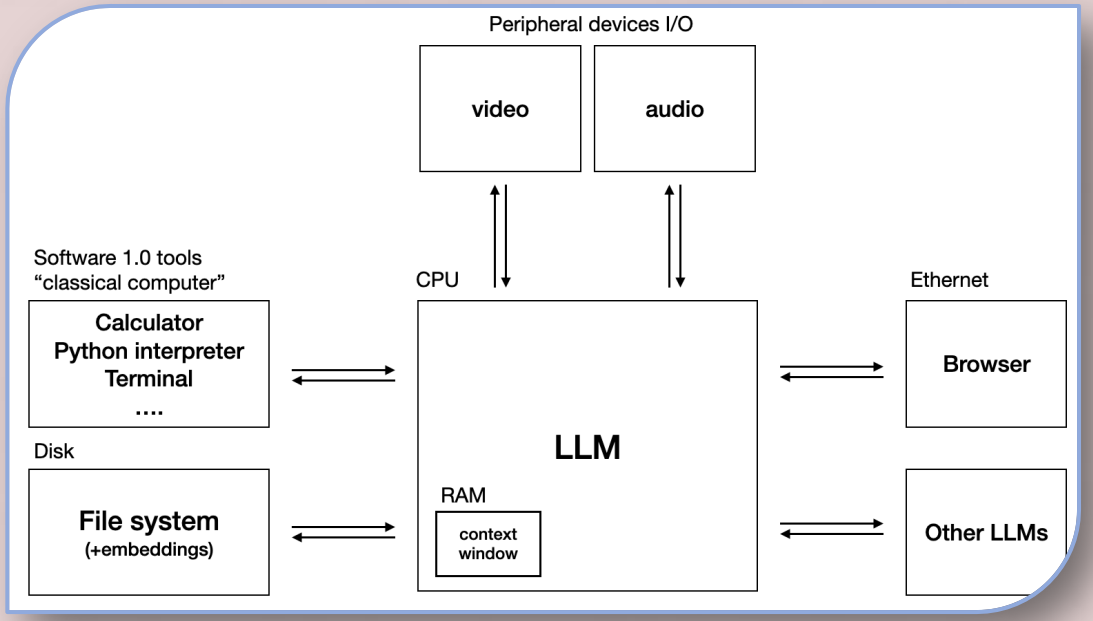
...t how...where are your CONVs and MATMULs? Read the code to solve this mystery.

Content By: Raghav Bali

Perspectives

Content By: Raghav Bali

LLM-OS



Source: Andrej Karpathy/tweet

Content By: Raghav Bali

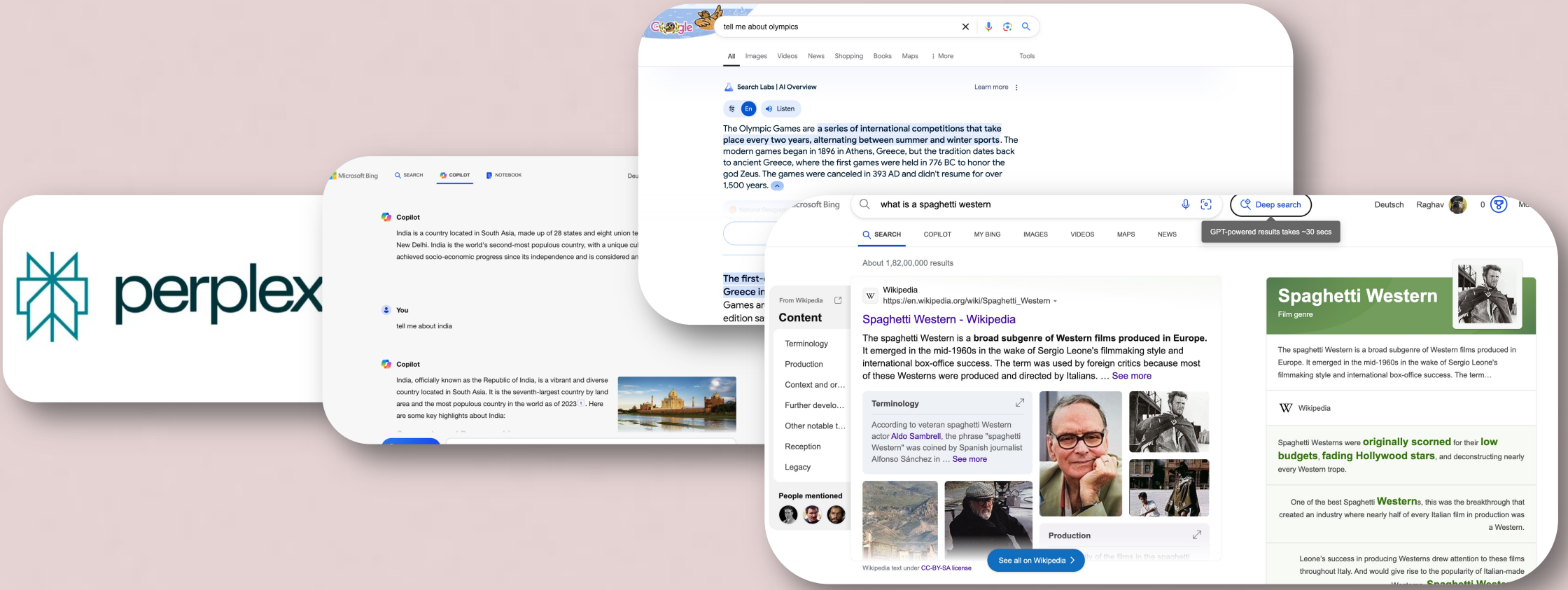
AI Products

Rabbit R1

AI Pin

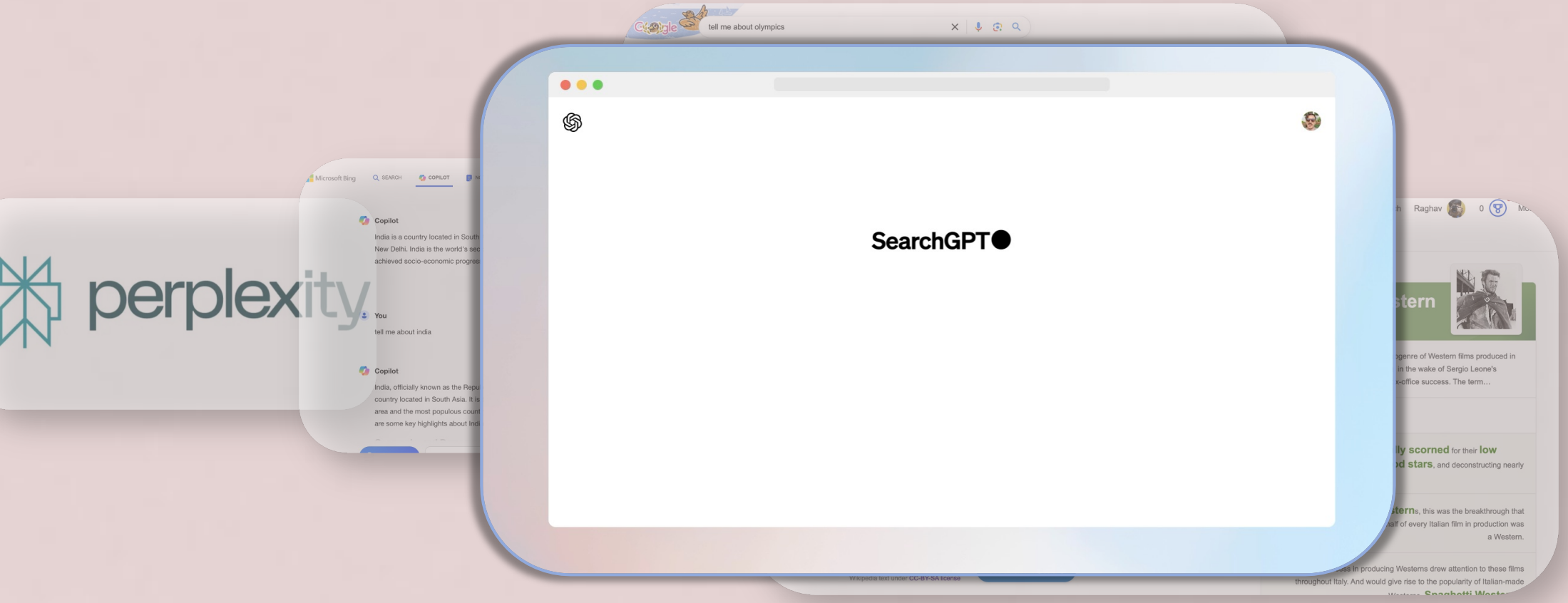
Content By: Raghav Bali

Search-GPT ?



Content By: Raghav Bali

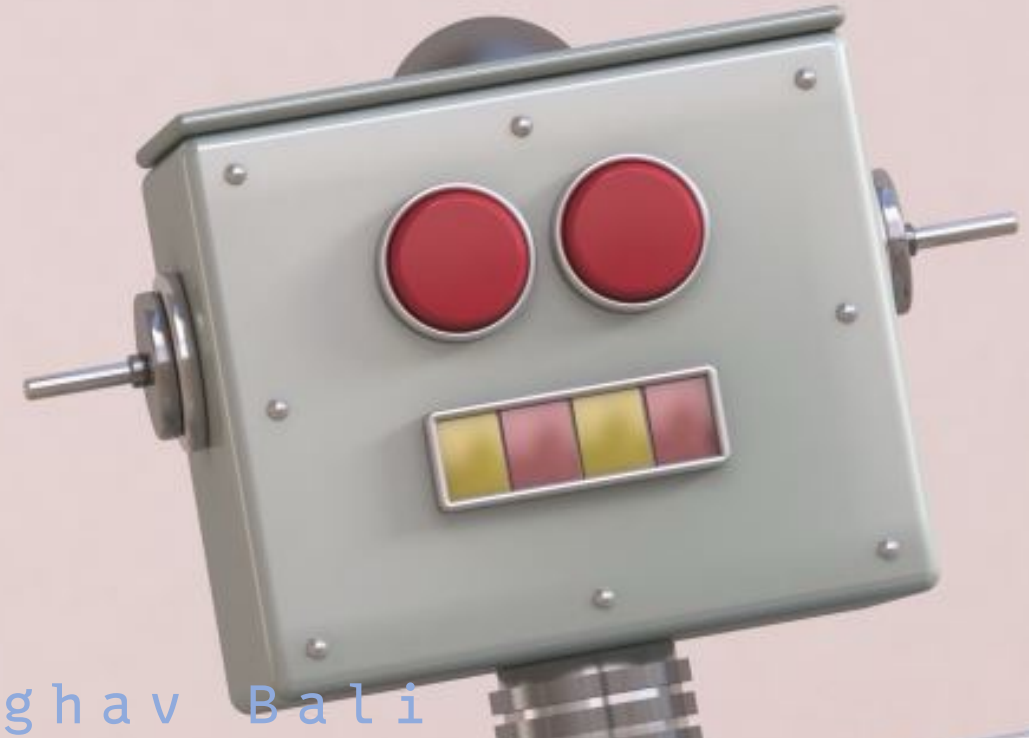
Search-GPT ?



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

Content By: Raghav Bali

Maturity



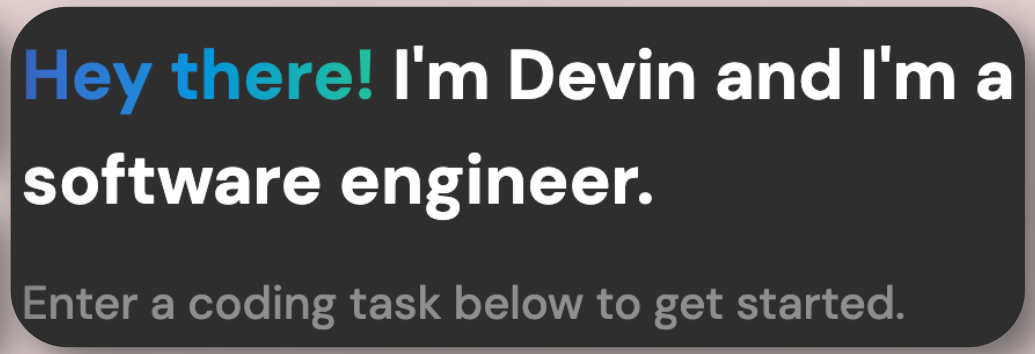
Content By: Raghav Bali

Maturity

LLMOps

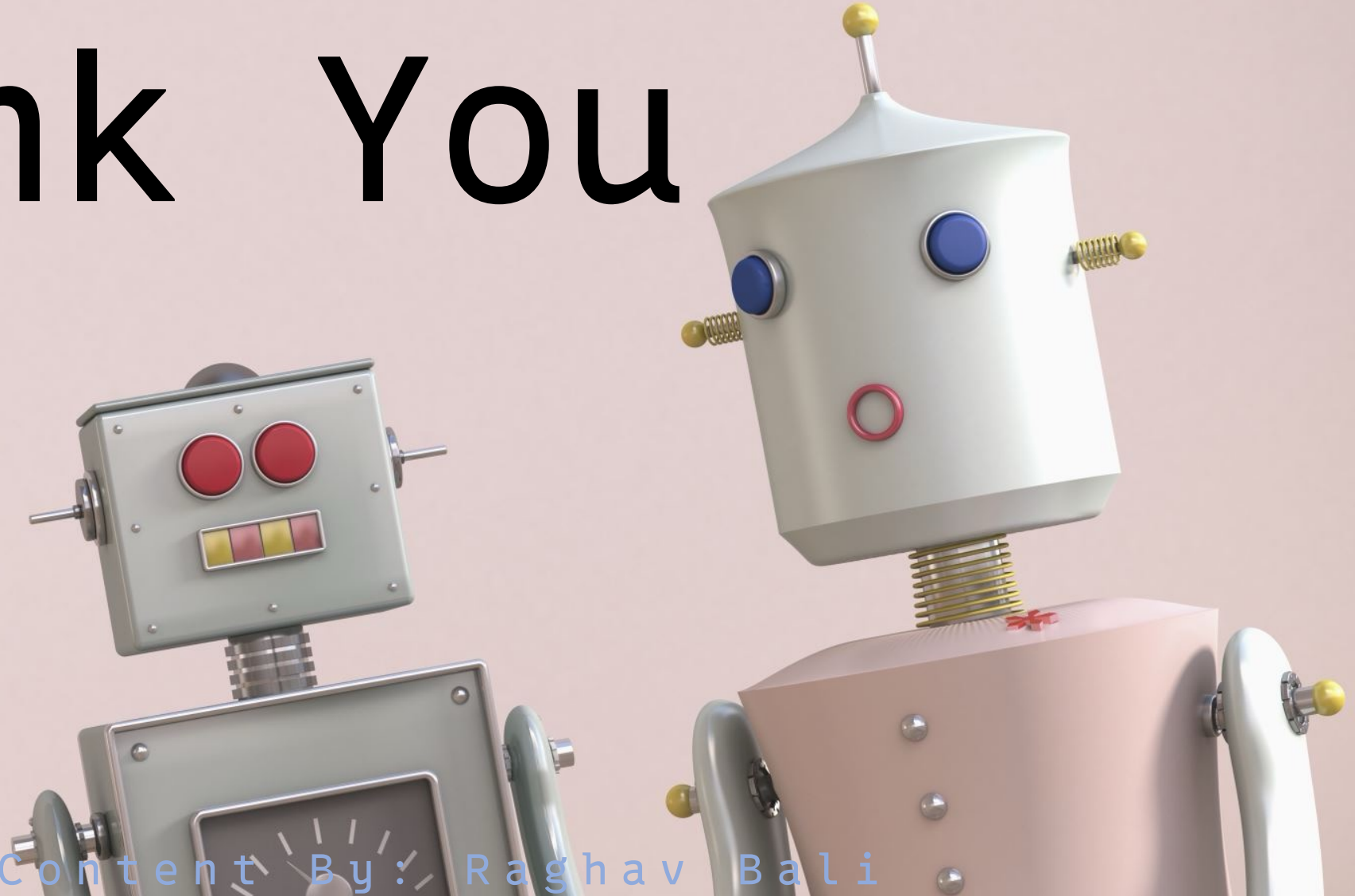
Llama Guard 3

SynthID



Content By: Raghav Bali

Thank You



Content By: Raghav Bali