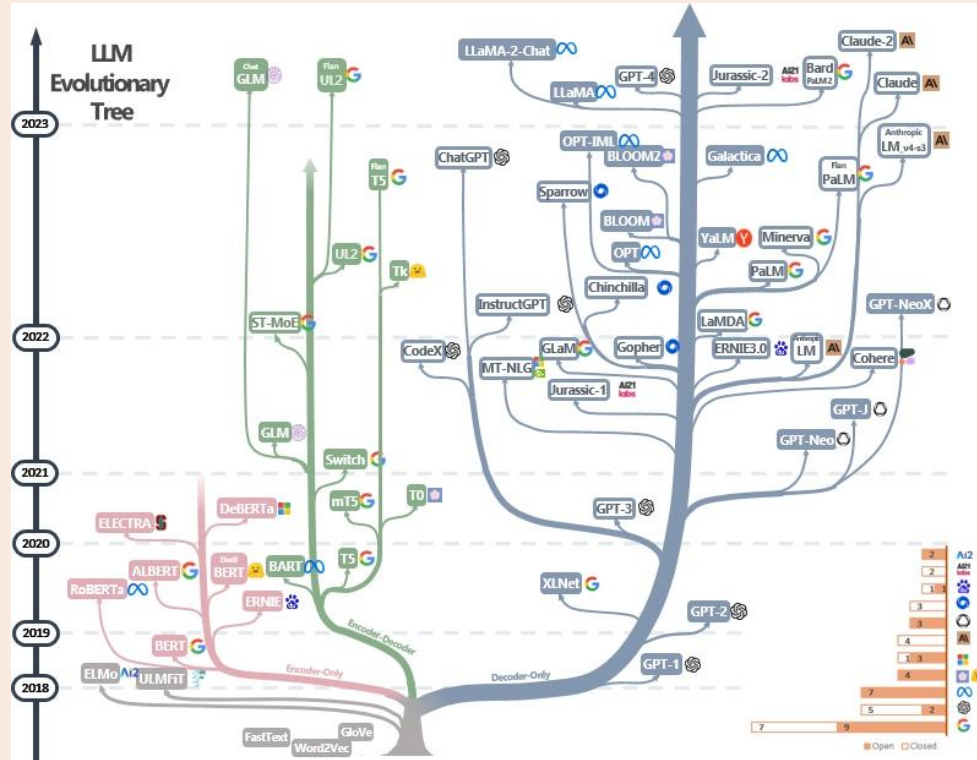


The Next Million AI Systems

Mark Huang
Co-founder, Chief Architect



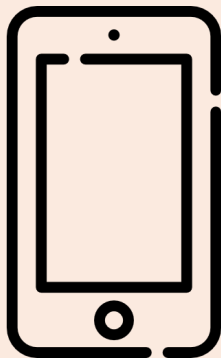
Growth in # of Models



There are more cellular connections than humans

11.9bn

cellular connections

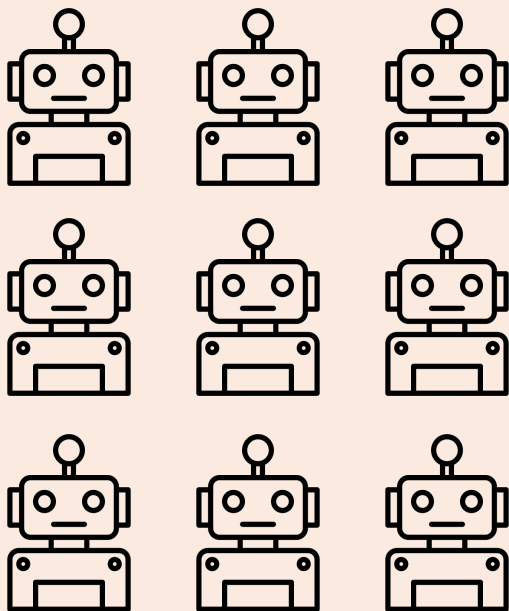


8.1bn

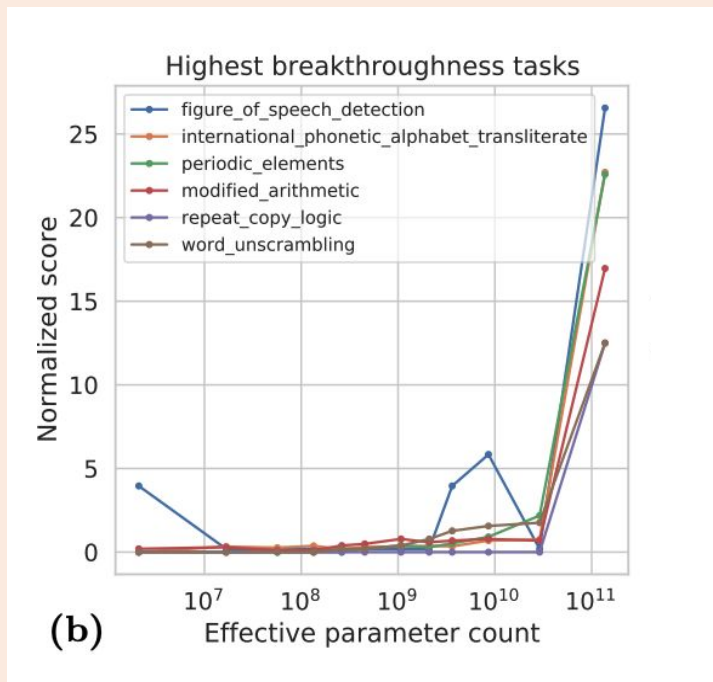
world population



There will be more AI models than humans

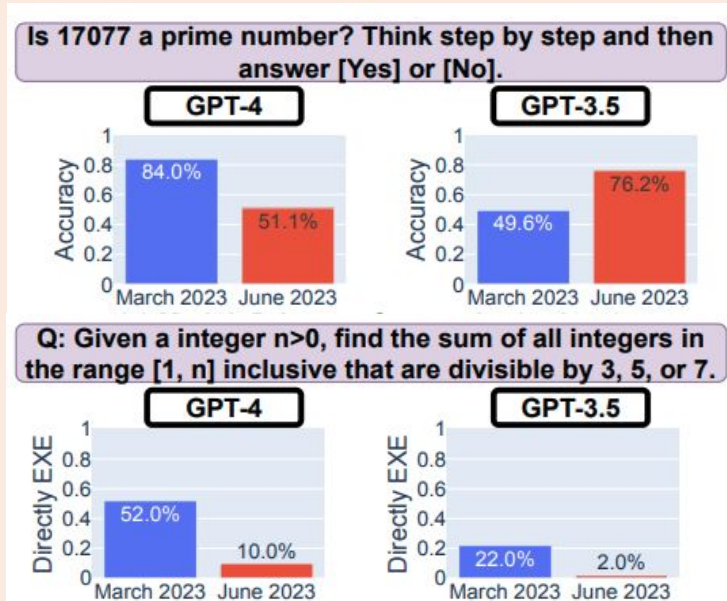
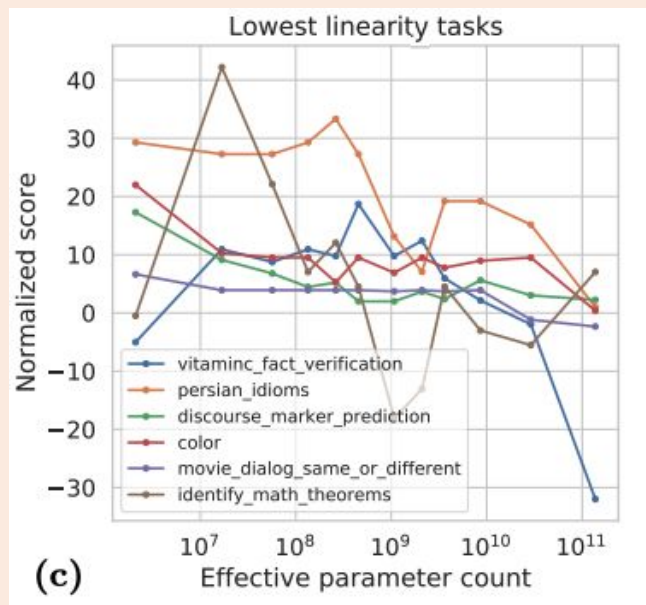


LLMs have emergent capabilities



Srivastava, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

No Free Lunch: Some tasks degrade



Cost as a function of number of parameters

	Training	Inference
Model Weights	$O(2)$	$O(2)$
Optimizer States	$O(4)$	
Activations	$O(2)$	
Context Length	$O(n \log n)$	$O(1)$

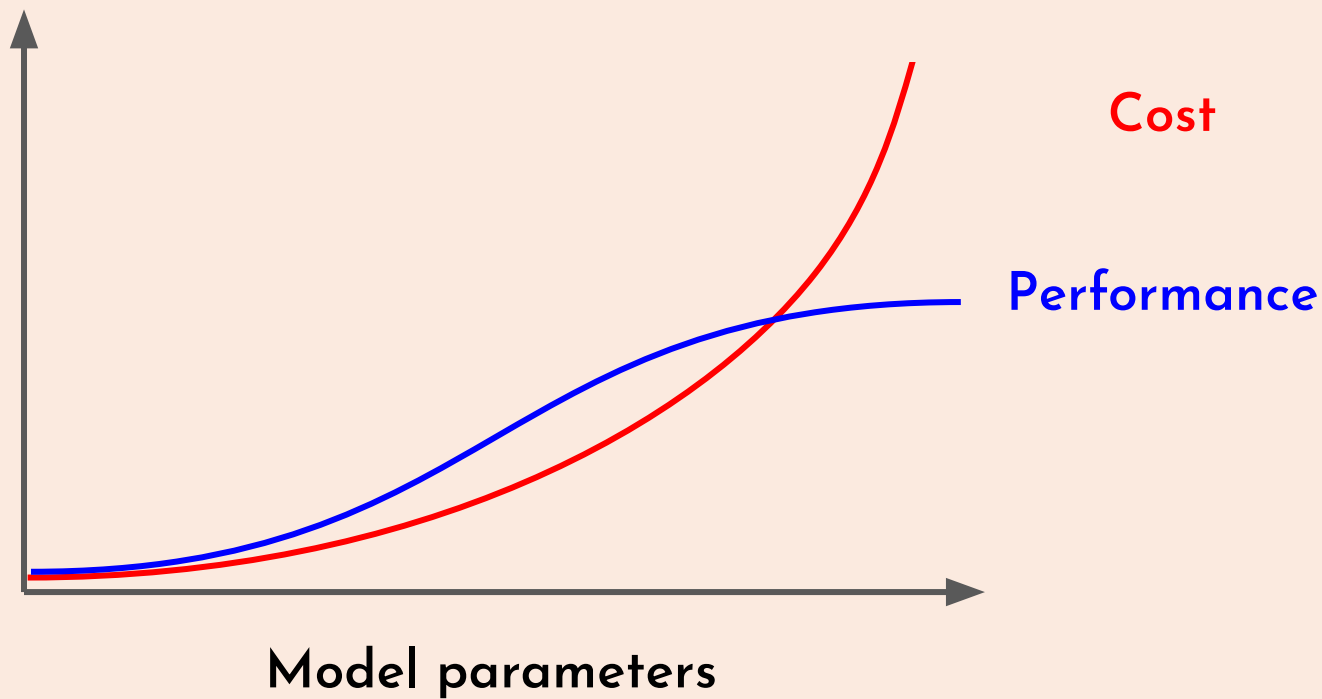
\$3.50/hr per H100 GPU

Assuming a 70B parameter model

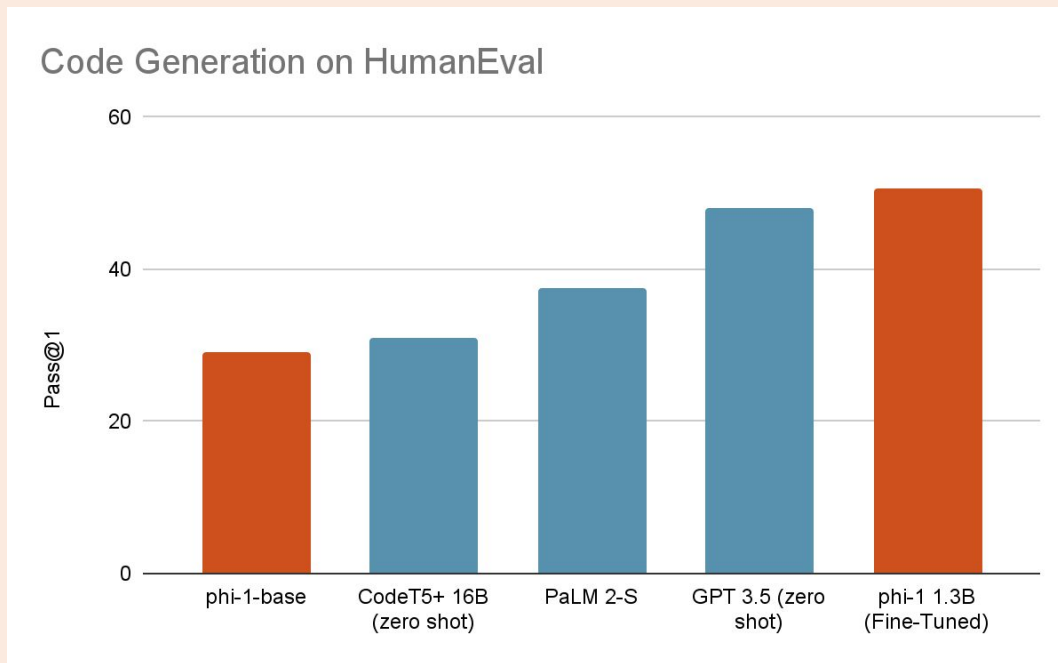
100 Million Tokens (Approximately 24,000 full samples)

= \$3,500 per experiment

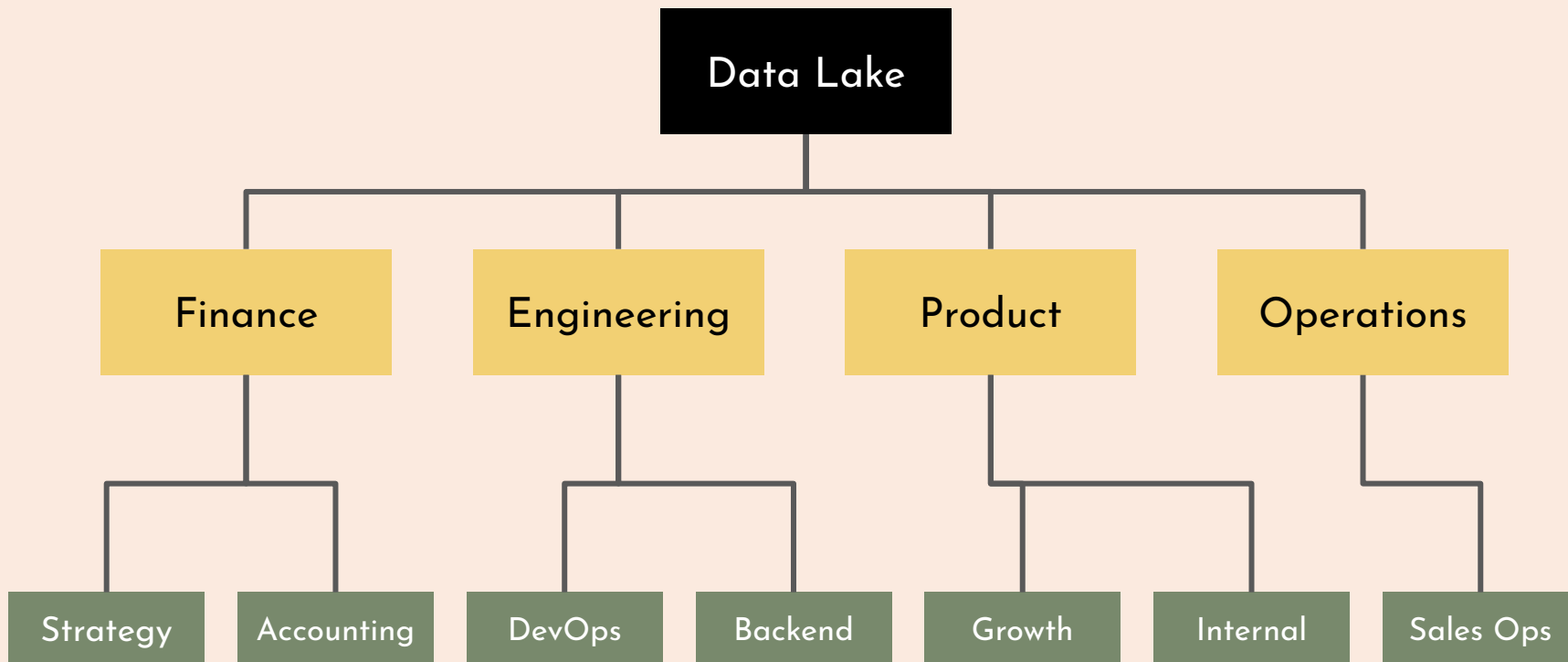
There are rapidly increasing costs AND diminishing marginal returns in single model development



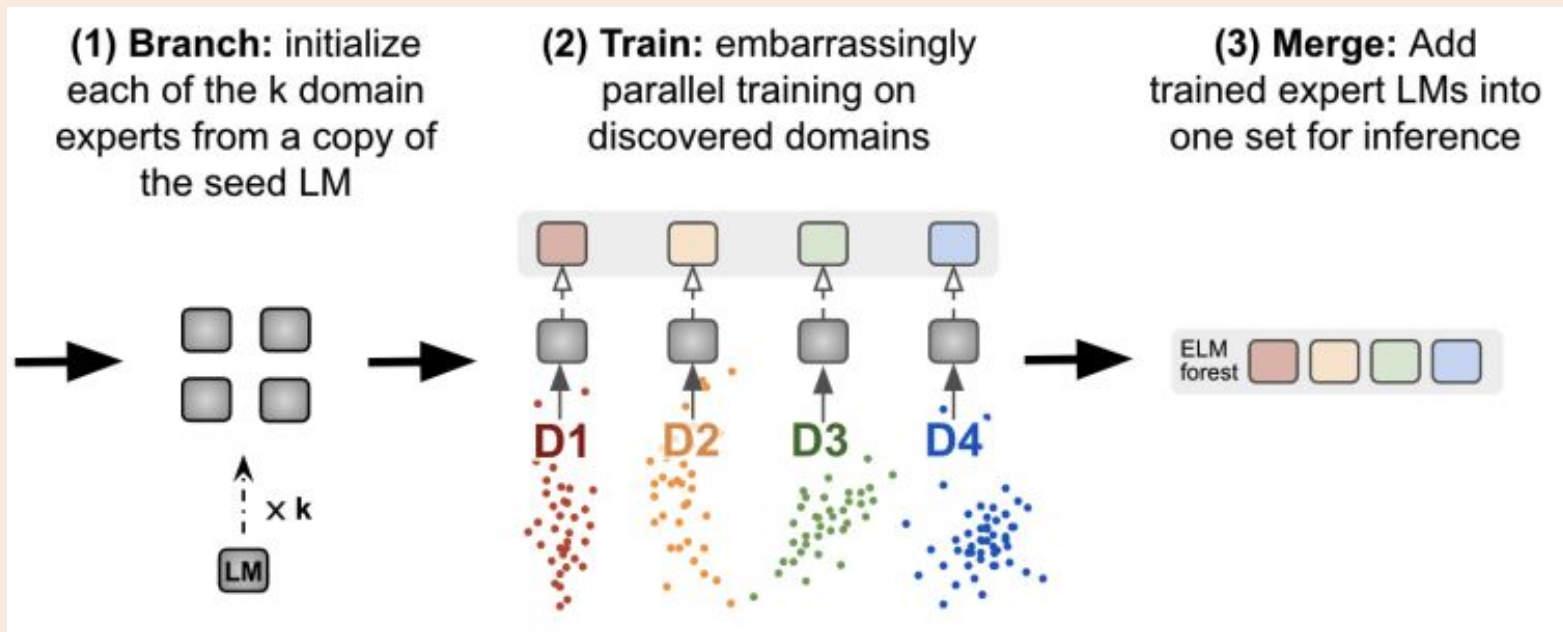
Fine-tuned small models outperform large models



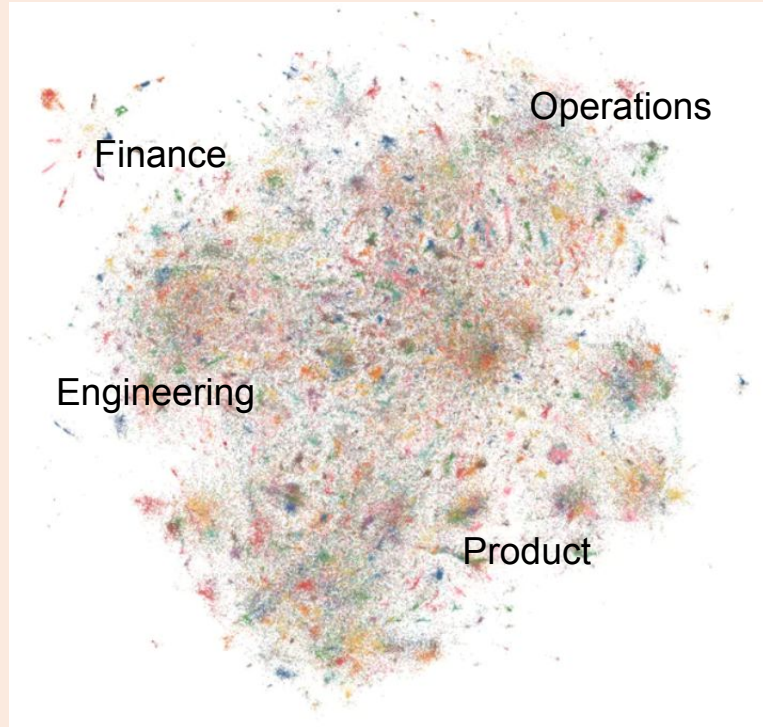
The Enterprise Data Corpus



Mixture of Experts (MoE): Cluster-Branch-Train-Merge



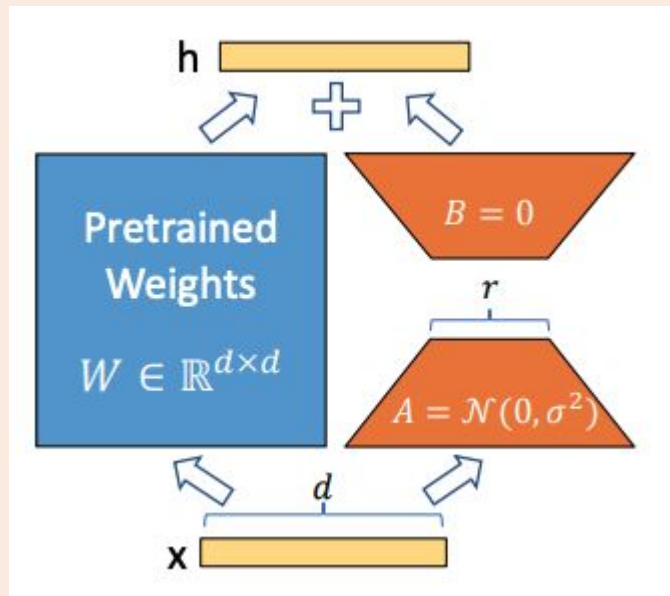
Embed Enterprise Corpus



Experts as Adapters

- Sparse Networks can be discovered without labels
- Performance increases with more data and more experts
- **Low Rank Adaptation (LoRA)** is critical
 - Lower training costs (4x less)
 - Faster expert switching

Parameter-Efficient Fine-Tuning (PEFT)



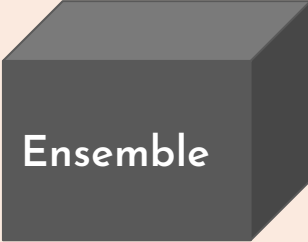
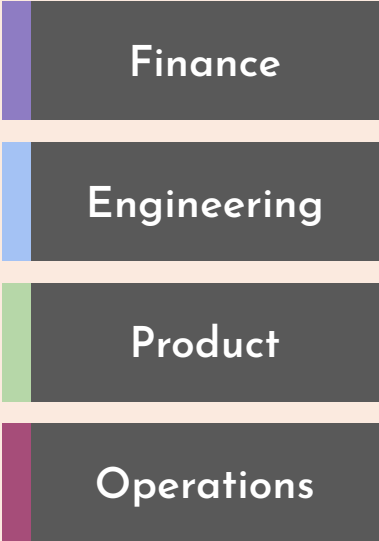
Mixture of Experts (MoE) + Low Rank Adaptation (LoRA)

Seed LLM

LoRA Adapter

Router

Merge

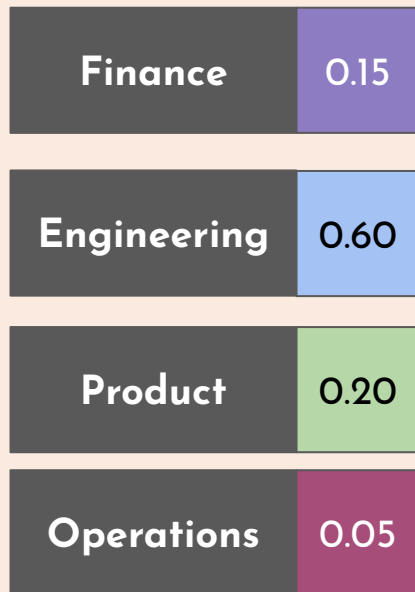


Inference Pipeline

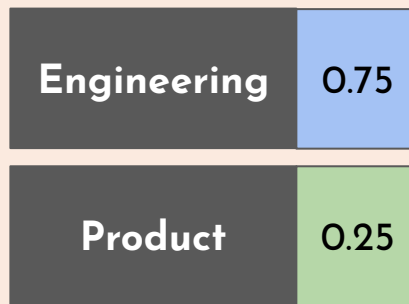
Prompt

Give me an estimate for the number of GPUs I need to support the personalization chatbot project based on the monthly active users from the current search product.

Embedding Probabilities



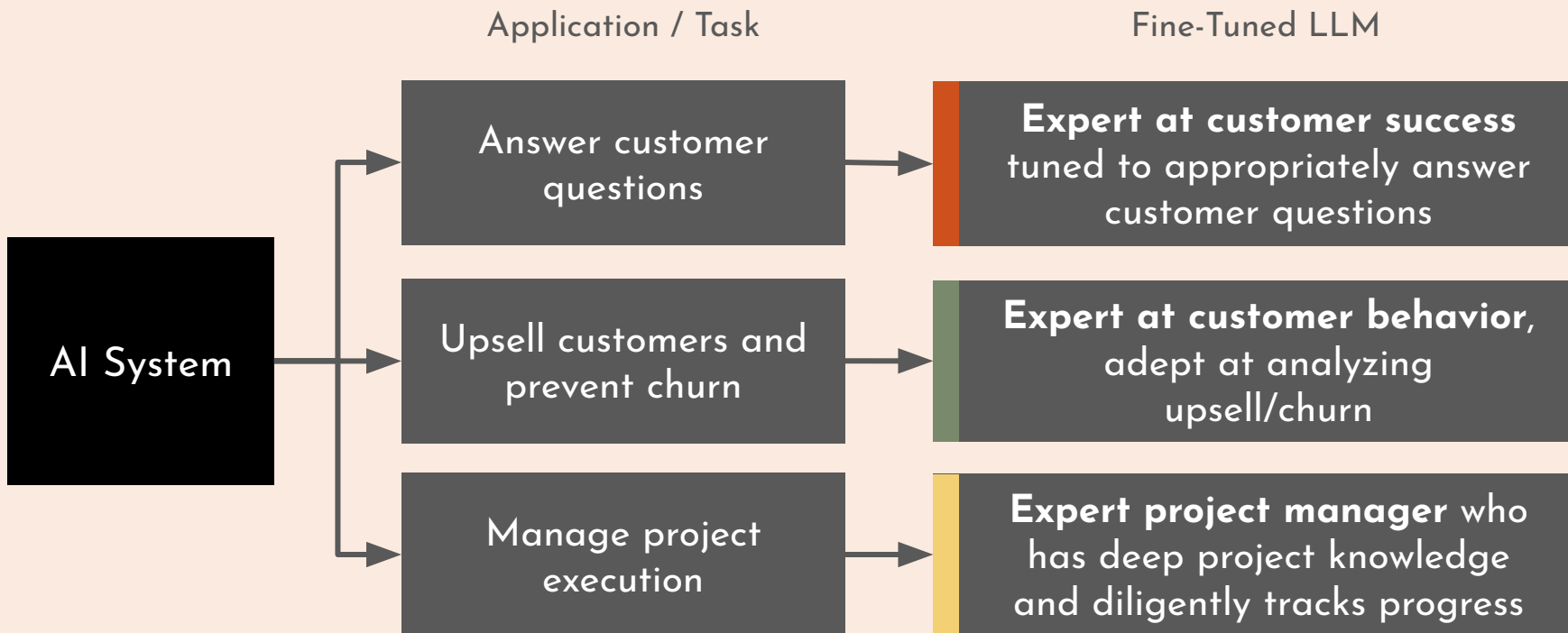
Top-K



Merge

There were approximately 1 million MAU over the last 6 months in search and the chatbot requires a 70b parameter model, so it will require at least 20 H100 GPUs.

Leveraging Mixture of Experts (MoE) ensures you have best-in-class performance on all tasks

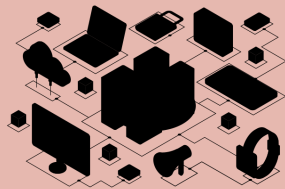


What is blocking this future?

1

Complex Infrastructure

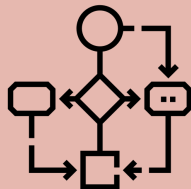
Each model requires dedicated infrastructure



2

Complex Development

Each model requires significant pre-training, etc.



3

High Cost

Development costs for many models



How do we solve this?

1

Complex Infrastructure

Each model requires dedicated infrastructure



One AI Cloud to host many models

2

Complex Development

Each model requires significant pre-training, etc.



Simplify model development

3

High Cost

Development costs for many models

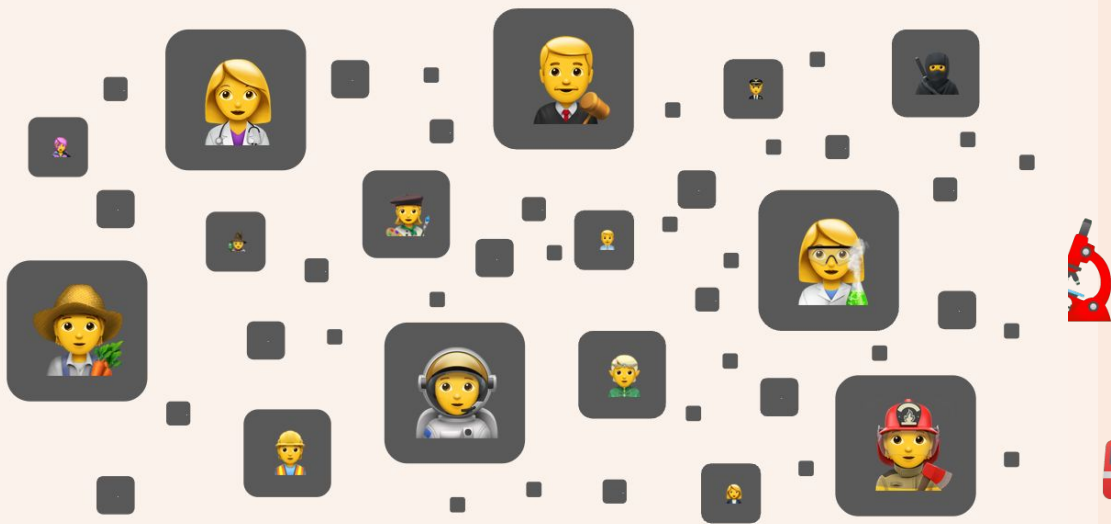



Efficient fine-tuning

Gradient is bringing the AI Cloud to reality: a single platform that can power millions of AI models

 **gradient AI Cloud**

APIs for Rapid,
Scalable LLM
Fine-Tuning and
Inference





Learn more at gradient.ai