



Improve your Llama RAG with AI Studio *and RAFT*

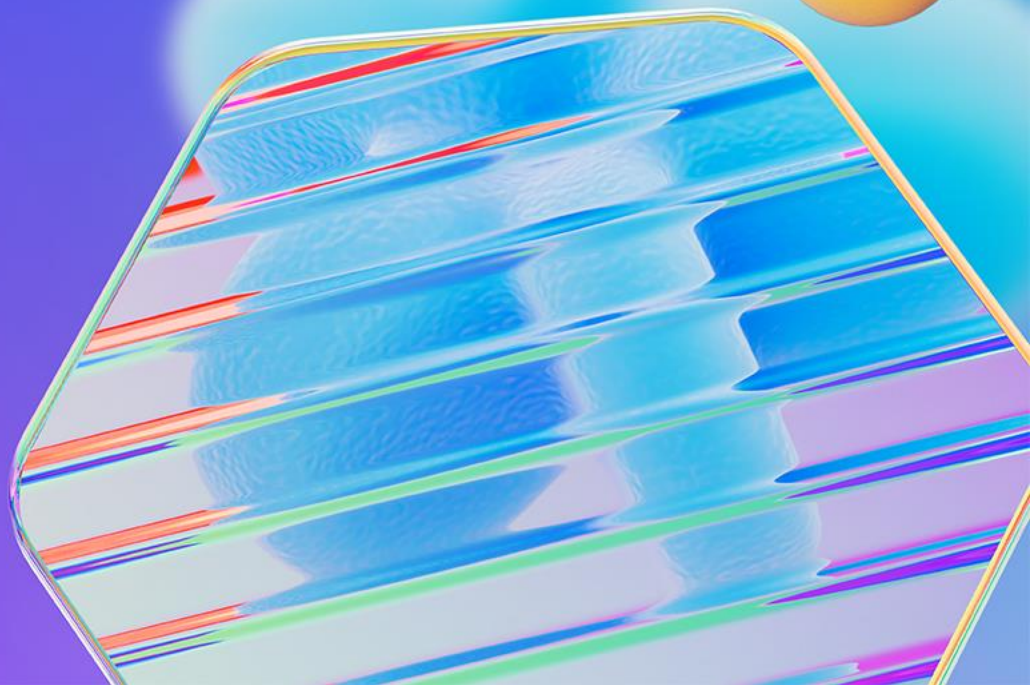
GenAI Summit SF 2024



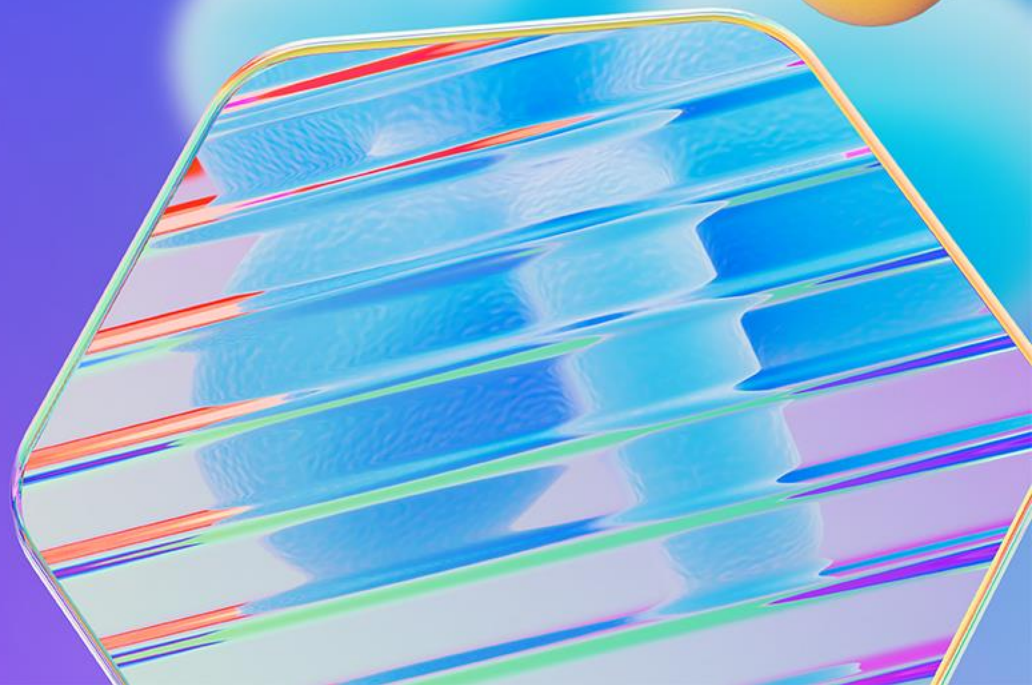
Cedric Vidal

Principal AI Advocate
Microsoft

Large Language Models



Current domain adaptation technics and their challenges



In context-learning and challenges

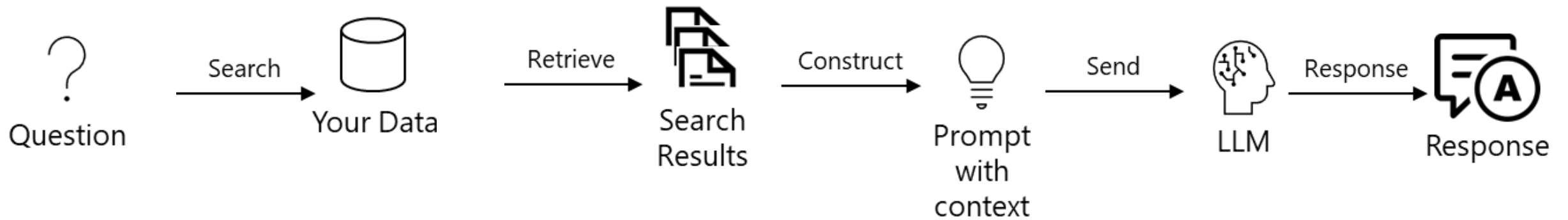


RAG and its challenges



Anatomy of RAG

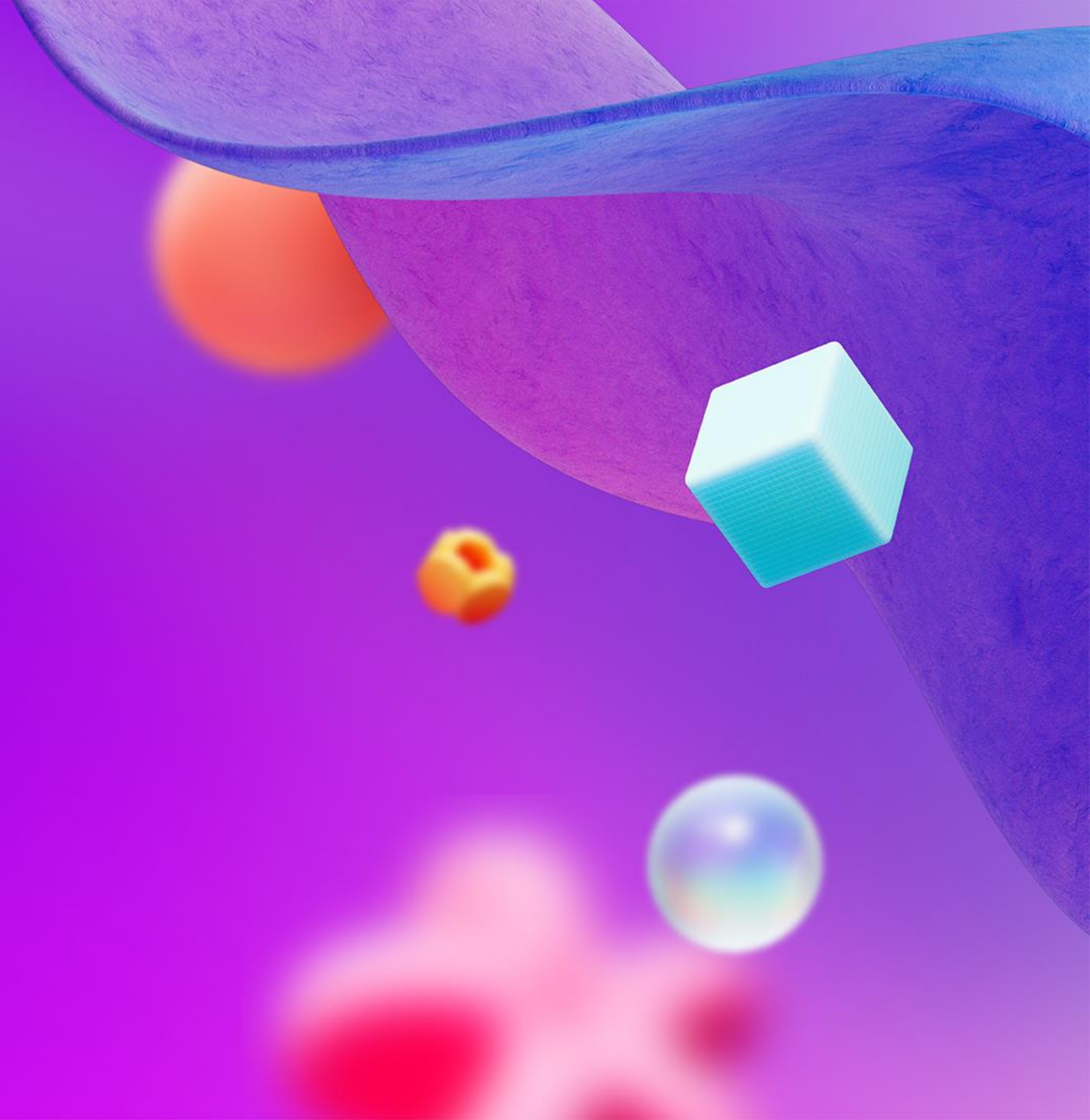
Augment



Retrieve

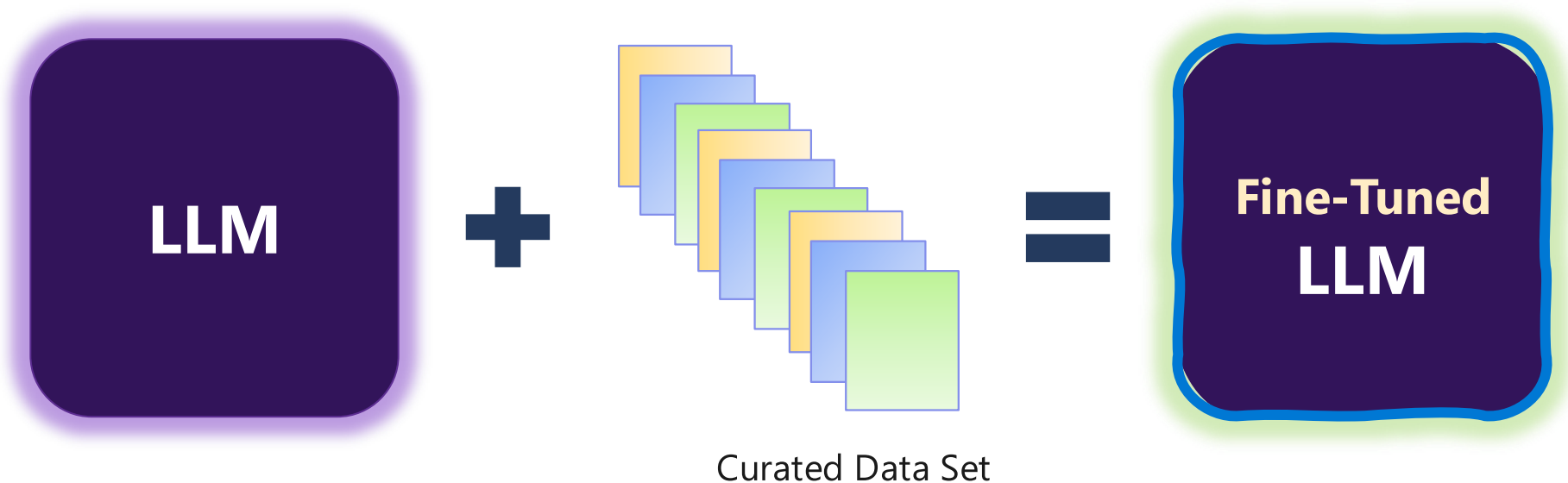
Generate

Fine-tuning and its challenges



What is fine tuning?

Fine-tuning refers to **customizing a pre-trained LLM** with additional training on a specific task or new dataset for enhanced performance and accuracy



Why fine tune?

Better performance

Developers hope that by fine tuning models with their own data and instructions, they'll get better results for their tasks

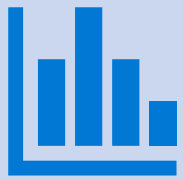
Cheaper or faster models

You may want to fine tune a smaller model for a specific task, instead of using an expensive general-purpose model

Differentiation

Most people won't train a foundation model; fine tuning with proprietary data provides a competitive advantage

Challenges of fine tuning



DATA



COMPUTATIONAL
HORSEPOWER



EXPERTISE

Recap of current domain adaptation technics

FSL

Few Shot learning

RAG

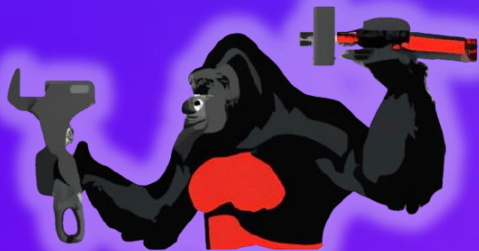
Retrieval Augmented
Generation

Fine-tuning

Fine-tuning

Research | UC Berkeley's RAFT

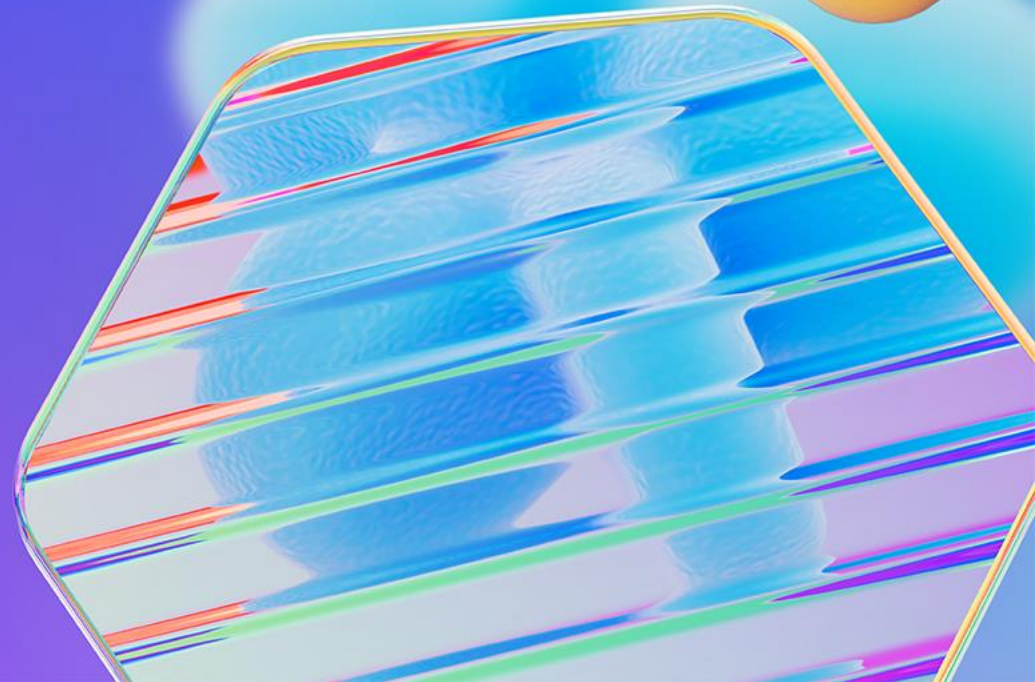
Retrieval Augmented Fine Tuning



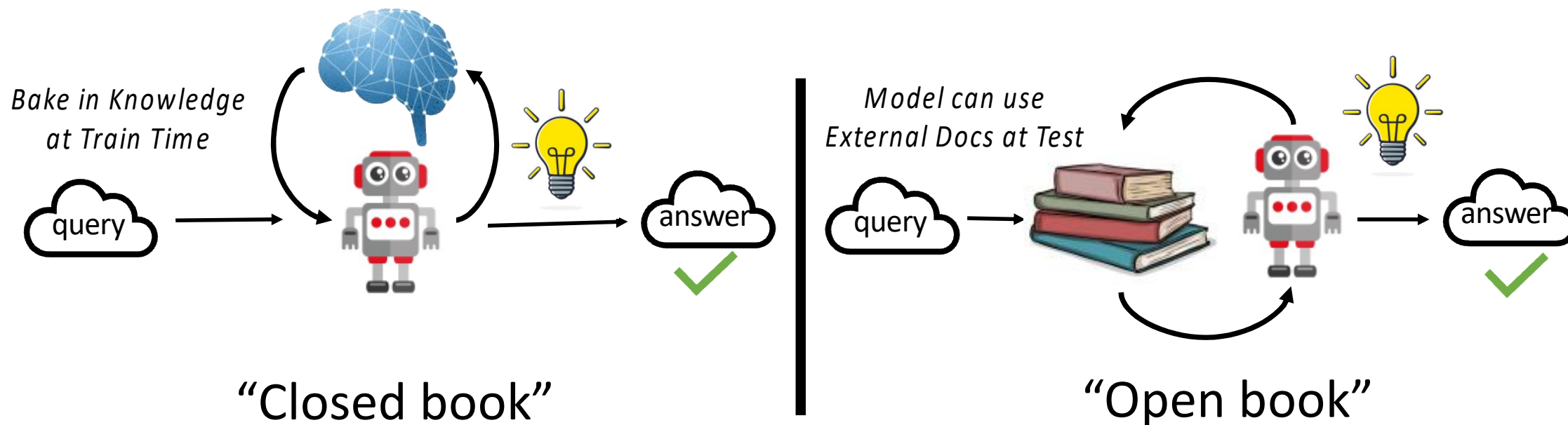
GORILLA



Berkeley
UNIVERSITY OF CALIFORNIA



Understanding RAFT with an Exam Analogy



A detailed illustration of a llama standing in a grand, ornate library. The llama is wearing round glasses, a white shirt, a patterned tie, and a dark suit jacket. It is holding a stack of books in its front paws. The library has high ceilings, arched windows, and floor-to-ceiling bookshelves filled with books. The lighting is warm and golden, creating a cozy atmosphere.

RAFT: Build your own librarian

Generated using Dalle-3 (Notebook <https://aka.ms/raft-genai-summit-2024-vis>)

RAFT: Inference

top-k retrieved documents

WIKIPEDIA
The Free Encyclopedia


Count Dracula

Count Dracula (/ˈdrækjulə, -jəˈ/) is the title character of Bram Stoker's 1897 gothic horror novel *Dracula*. He is considered the prototypical and archetypal vampire in subsequent works of fiction. Aspects of the character are believed by some to have been inspired by the 15th-century Wallachian prince Vlad the Impaler, who was also known as Vlad Dracula, and by Sir Henry Irving, an actor for whom Stoker was a personal assistant.^[12]

One of Dracula's most iconic powers is his ability to turn others into vampires by biting them and infecting them with the vampiric disease. Other characteristics have been added or altered in subsequent popular fictional works, including films, cartoons and breakfast cereals.

Stoker's creation

Bram Stoker's novel takes the form of an epistolary tale, in which Count Dracula's characteristics, powers, abilities, and weaknesses




Bela Lugosi as Count Dracula in the

WIKIPEDIA
The Free Encyclopedia

Bodyboarding

(Redirected from Body boarding)

Bodyboarding is a water sport in which the surfer rides a bodyboard on the crest, face, and curl of a wave which is carrying the surfer towards the shore. Bodyboarding is also referred to as *Boogieboarding* due to the invention of the "Boogie Board" by Tom Morey in 1971. The average bodyboard consists of a short, rectangular piece of hydrodynamic foam. Bodyboarders typically use swim fins for additional propulsion and control while riding a breaking wave.



Origin

Bodyboarding originates from an ancient form of water sports


WIKIPEDIA
The Free Encyclopedia

Bat

(Redirected from Bats)

Bats are flying mammals of the order **Chiroptera** (/kaɪˈroʊptərə/).^[1] With their forelimbs adapted as wings, they are the only mammals capable of true and sustained flight. Bats are more agile in flight than most birds, flying with their very long spread-out digits covered with a thin membrane or patagium. The smallest bat, and arguably the smallest extant mammal, is Kitt's hog-nosed bat, which is 29–34 millimetres (1¹/₈–1³/₈ inches) in length, 150 mm (6 in) across the wings and 2–2.6 g (1⁄64–1⁄20 oz) in mass. The largest bats are the flying foxes, with the giant golden-crowned flying fox (*Acerodon jubatus*) reaching a weight of 1.6 kg (3¹/₂ lb) and having a wingspan of 1.7 m (5 ft 7 in).

The second largest order of mammals after rodents, bats comprise about 20% of all classified mammal species worldwide, with over 1,400 species. These were traditionally divided into two suborders: the largely fruit-eating megabats, and the



Bat
Temporal range:



Vampire

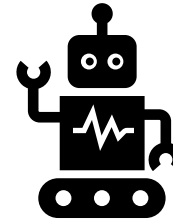


Bat tail



Microbat

What kind of vision do most microbats have?



The RAFT-trained LLM synthesizes the answer from relevant documents

Mesopic vision

Demo - RAFT - 03 - Promptflow - RAG



Demo - RAFT - 04 - Promptflow - RAFT



RAFT: Chain-of-thought ++

Oracle Context

Its skull was long and narrow, and its face had an upturned snout.[6]Based on its skull dimensions, it may have had a wingspan of approximately 50 cm (20 in) and a body mass of 60 g (2.1 oz). The proportions are equivalent to a smaller megabat or larger microbat of modern chiropteran fauna.[7] Taxonomy and etymology DescriptionIts braincase was 14.5–14.8 mm (0.57–0.58 in) wide and 13.4–14.8 mm (0.53–0.58 in) high.[8]Some researchers believe that D. draculae would have preyed on megafauna,[9] while others believe that they would have preyed on Caviomorpha.[10] Other potential prey items that would have been available to D.

Question

What was the estimated wingspan of D. draculae?

CoT Answer

To answer the question, we need to identify the estimated wingspan of D. draculae.

The context provided gives us the necessary information to determine this.

First, we look for the part of the context that directly mentions the wingspan of D. draculae.

We find it in the sentence: `##begin_quote##` Based on its skull dimensions, it may have had a wingspan of approximately 50 cm (20 in) and a body mass of 60 g (2.1 oz). `##end_quote##`.

This sentence explicitly mentions the estimated wingspan of D. draculae as approximately 50 cm (20 in).

Therefore, based on the context provided, we can conclude that the estimated wingspan of D. draculae is approximately 50 cm (20 in).

`<ANSWER>` : 50 cm (20 in)

RAFT improves RAG performance for all specialized domains

	PubMed	HotpotQA	HuggingFace	Torch Hub	TensorFlow Hub
GPT-3.5 + RAG	71.60	41.5	29.08	60.21	65.59
LLaMA2-7B	56.5	0.54	0.22	0	0
LLaMA2-7B + RAG	58.8	0.03	26.43	08.60	43.06
DSF	59.7	6.38	61.06	84.94	86.56
DSF + RAG	71.6	4.41	42.59	82.80	60.29
RAFT (LLaMA2-7B)	73.30	35.28	74.00	84.95	86.86

Dataset generation using RAFT

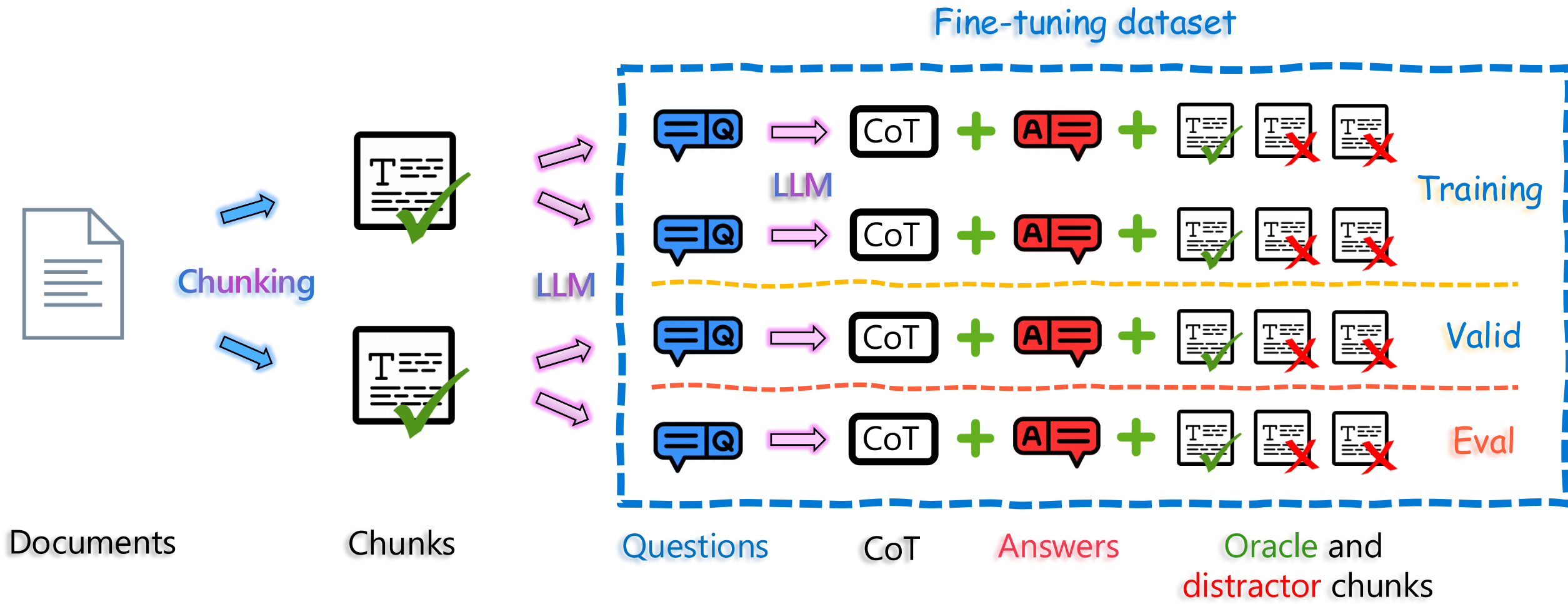
DEMO



Demo - RAFT - 01 – Gen - 00:24



RAFT Dataset generation



Demo - RAFT - 01 – Gen - 00:24



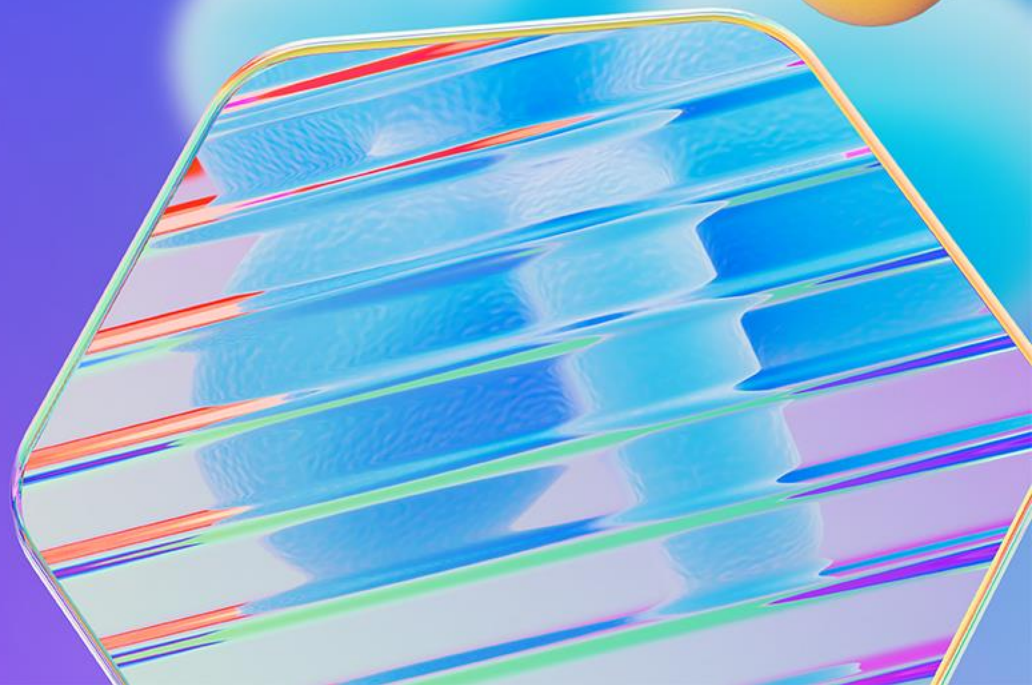
Fine-tuning using Azure AI Studio



Demo - RAFT - 02 - Fine-tuning



Conclusion



Metrics

	Base Llama 2 7B Chat	Fine-tuned Llama 2 7B
Similarity <i>To Llama 3 70B</i>	2.73	4.76 (1.7x)
Groundedness	2.74	3.16 (1.15x)

Scores on a scale of 1-5

Inferencing Costs

Llama 2 7B Fine Tuned

paygo-finetuned-model-inference-hosting:
\$3.09 per hour

paygo-finetuned-model-inference-output-tokens:
\$0.00067 per 1000 tokens

paygo-finetuned-model-inference-input-tokens:
\$0.00052 per 1000 tokens

Llama 3 70B

paygo-inference-output-tokens:
\$0.01134 per 1000 tokens

paygo-inference-input-tokens:
\$0.00378 per 1000 tokens

100x Cheaper for inference costs

Cost

Dataset generation

Model: Llama 3 70B

Data points 18K

Tokens_in 10M

Tokens_out 3.7M

Total ~ \$80

Fine tuning

Duration 1.5 hours

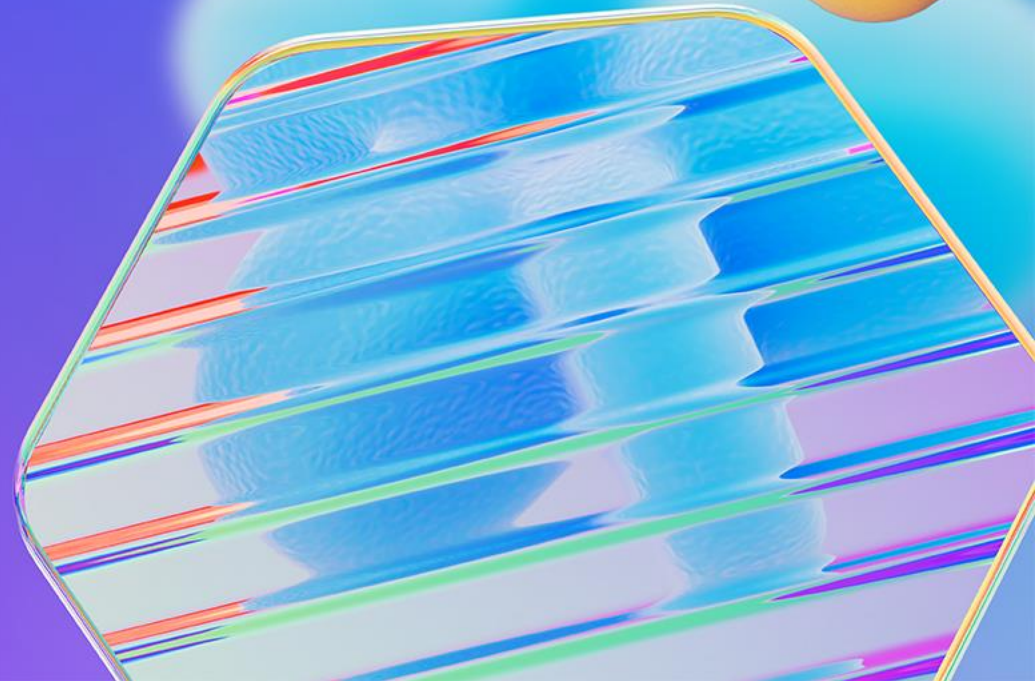
Cost \$35/hour

Total ~ \$50

paygo-inference-output-tokens:
\$0.01134 per 1000 tokens

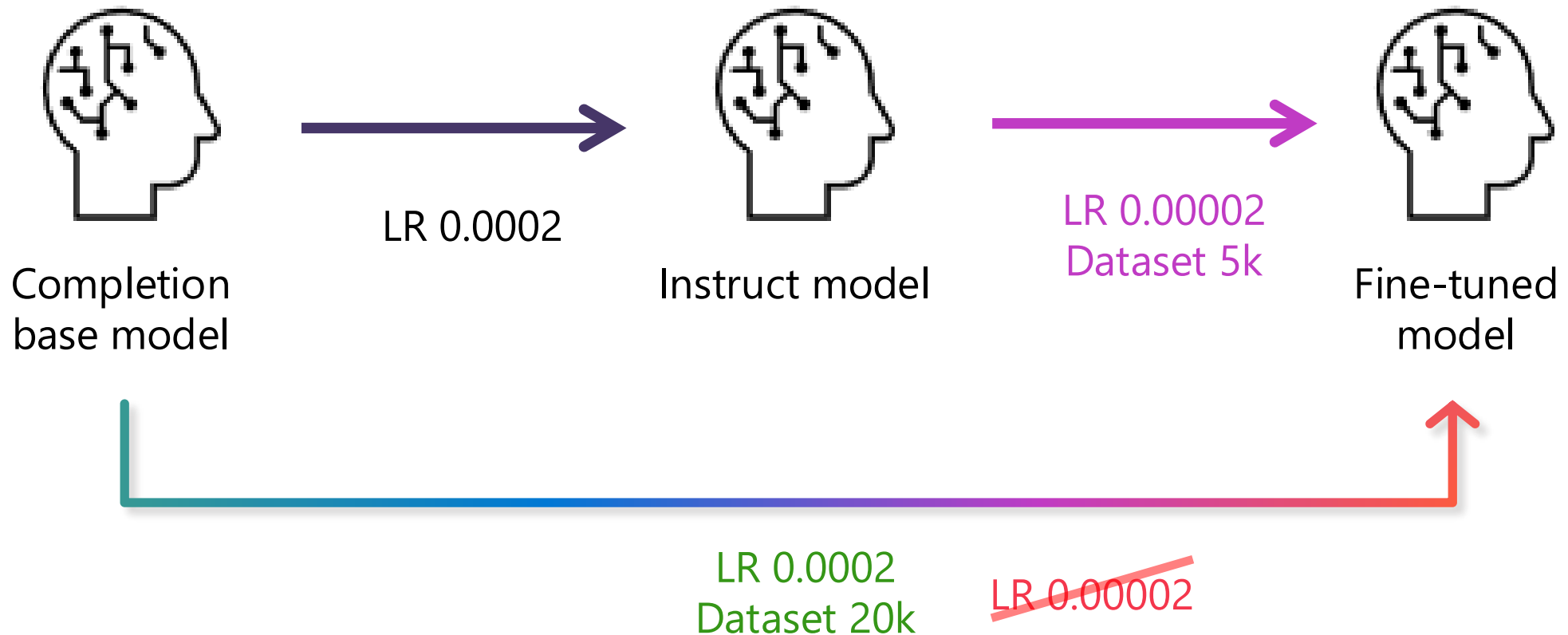
paygo-inference-input-tokens:
\$0.00378 per 1000 tokens

Key Insights



Which fine-tuning learning rate?

- We fine-tuned a Llama 2 7B completion model



Throughput and Load balancing – Data gen

MaaS

Convenient
Pay as You Go
No infra to manage

70 tokens / s

MaaP

Standard_NC96ads_A100_v4
4x Nvidia A100

140 tok/s/VM in average
(35 tok/s/GPU x 4)

Requires managing
provisioning and teardown

Shard

Shard the dataset across
multiple load balanced
endpoints

Increase tok/s

LiteLLM

<https://www.litellm.ai/>

Load balancing on:

2 MaaP NC96ads_A100_v4 +
1 MaaS endpoint

=> 350 tok/s in average

Thank you!



Cedric Vidal

Principal AI Advocate

Microsoft

aka.ms/cvidal



aka.ms/raft-build24

Q&A

