Lessons learnt from MLOps for GenAI at Google Scale

Prakash Chockalingam

BEAM SUMMIT

September 4-5, 2024

Sunnyvale, CA. USA

About me

- Currently Product Lead @ Google
- Building ML Platforms and tools for Googlers to build and deploy models in production.
- Prior to Google:
  - Worked at Meta, Databricks, Netflix, ...
  - Building big data and AI platforms is my area of expertise.

Agenda

01 Gemini & Gemma Models

02 AI @ Google

03 Gemini adaptation workflows

04 Gemini adaptation characteristics & challenges

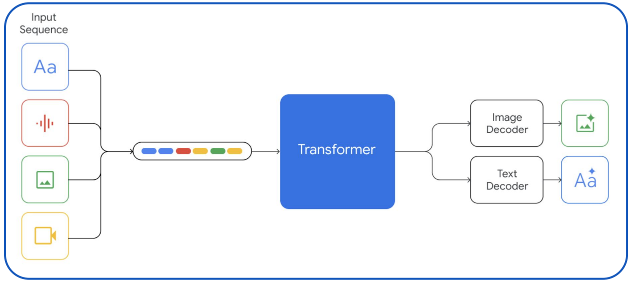
05 Q&A

Gemini 1.0

Family released end 2023, Natively multi-modal

Came in 3 flavors:

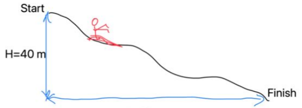
- **01 Ultra** - best in class across modalities
- **02 Pro** - more broadly deployable
- **03 Nano** - for on device



Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?



Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

Start: $E = mgL$

End: $E = \frac{1}{2} mv^2$

$\Rightarrow mgL = \frac{1}{2} mv^2$

$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6\ m.s^{-1}$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.

Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by mgh, where m is the mass of the skier, g is the acceleration due to gravity, and h is the height of the slope. The student used the length of the slope, L, instead of the height, H, in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by mgh.

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2} mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2} mv^2$$

Canceling out the mass m on both sides, we get:

$$gh = \frac{1}{2} v^2$$

Solving for v, we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

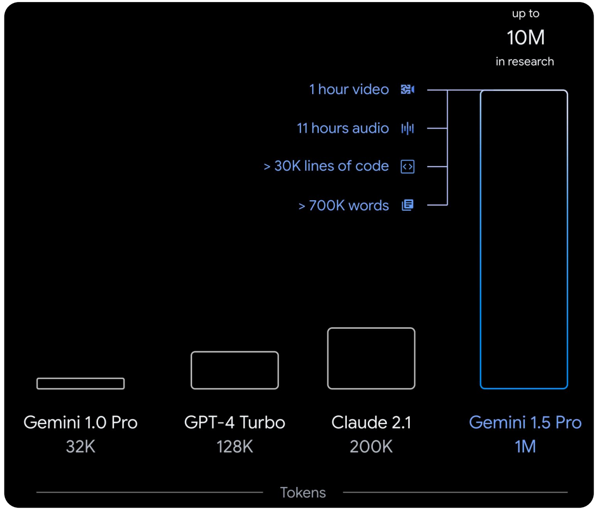
$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

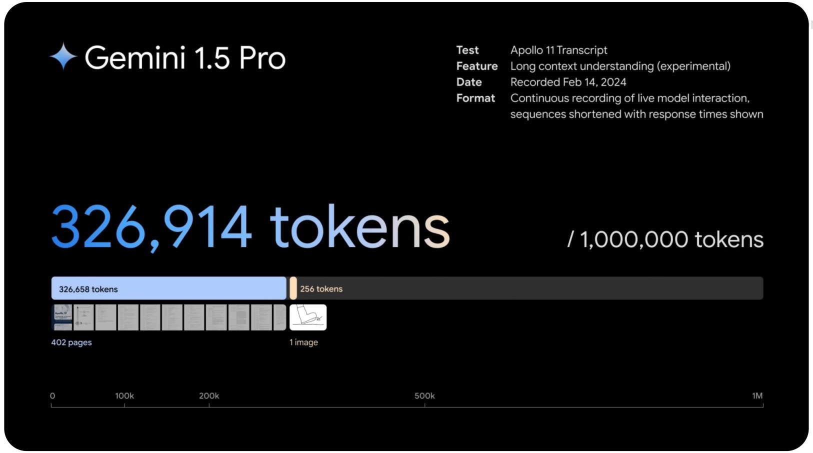
Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.

Gemini 1.5
Released in Feb

World's longest context window

Mind-blowing 10M+ token input possible - launched with 1M cap





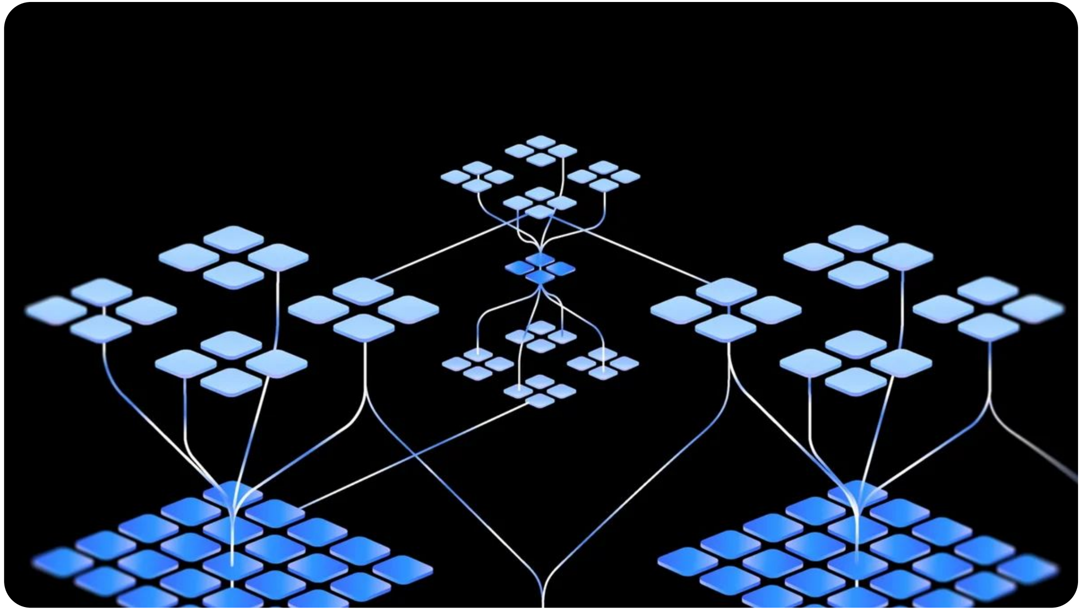
Gemini 1.5 I/O upgrades

Gemini 1.5 Flash

Much faster, smaller model, while preserving great performance on benchmarks: best quality/cost ratio

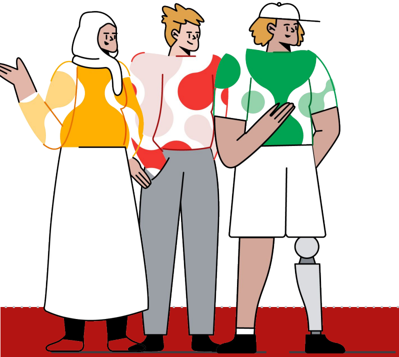
Expanding Gemini 1.5 Pro to 2M context!

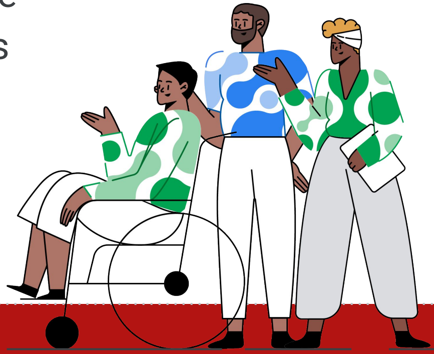
Expanded context window to 2 million tokens



Gemma open models

A family of lightweight, state-of-the art open models built from the same research and technology used to create the Gemini models





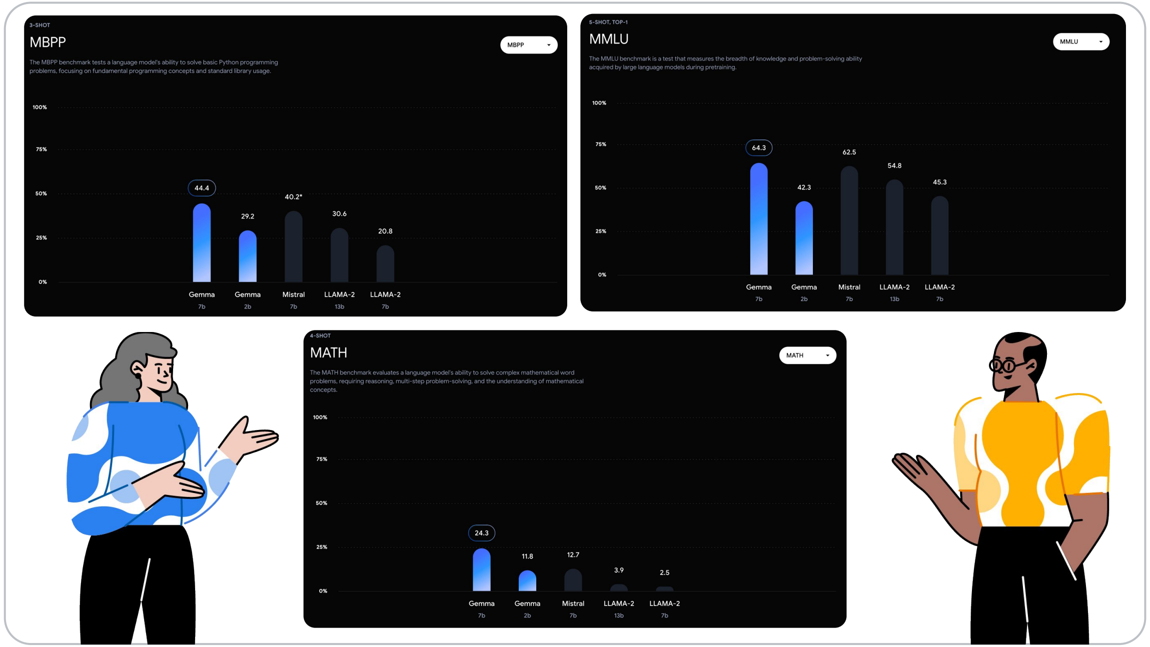
Gemma 1.0

Released in Feb
best open models at size

01 **7B** - most useful, broadly portable/deployable

02 **2B** - better for CPU/on device

Established best performance across several benchmarks, esp impressive on MMLU, MATH and Coding.



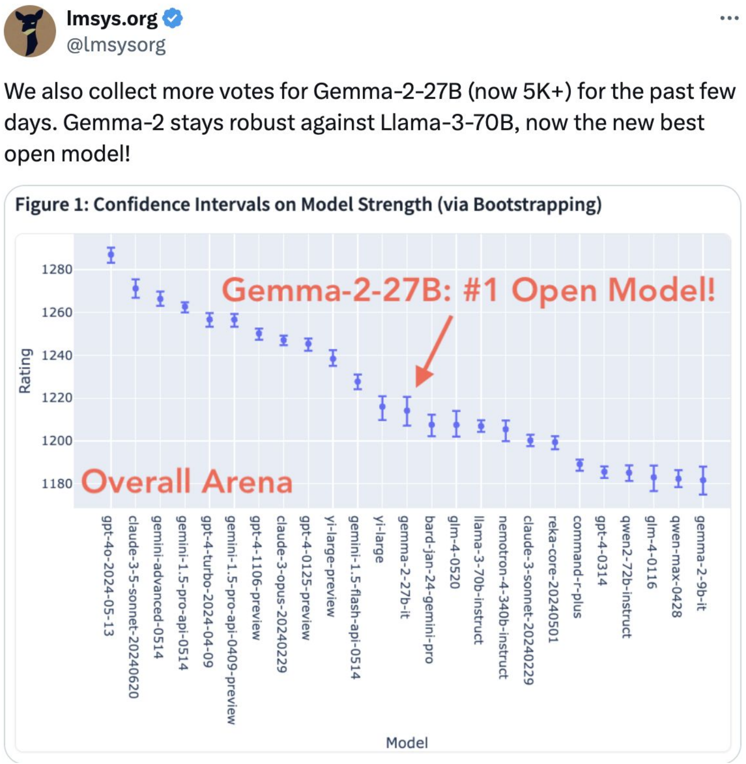
Gemma 2.0

Released in May 2024

01 **27B** - most capable open model from Google

Our goal is to continue delivering small models, that are portable, while being as capable as possible

This reinforces our belief in the power of collaboration and open research to drive innovation

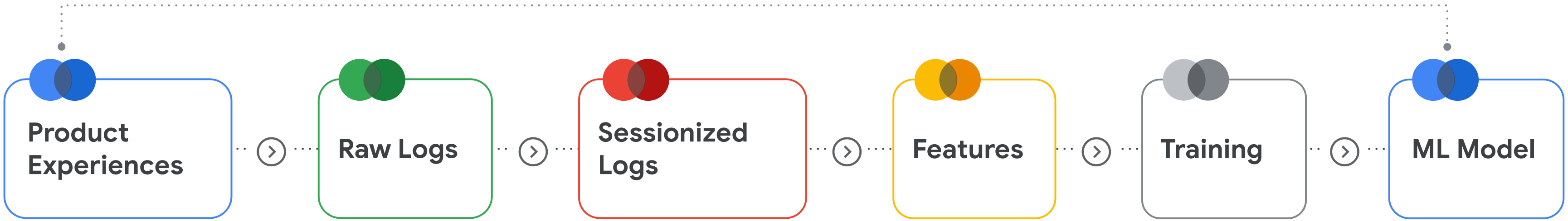


AI @ Google

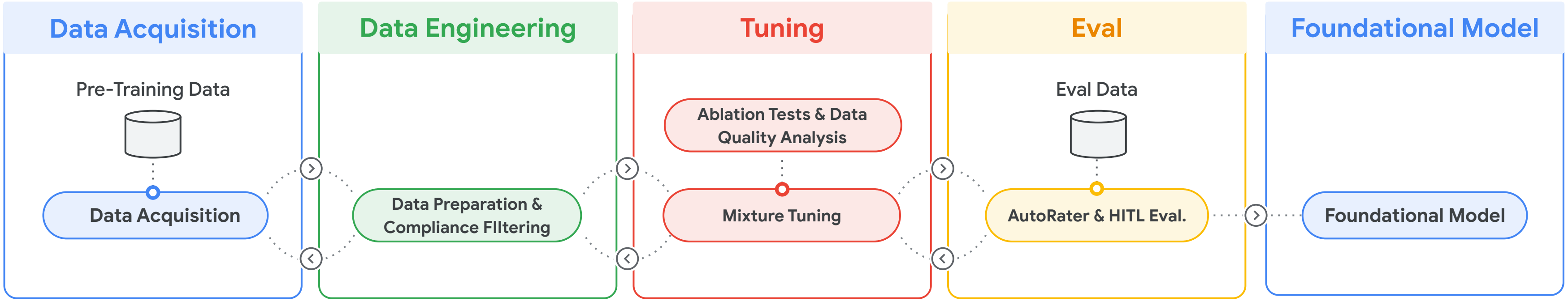
Turn the diversity of AI investments in a multiplier, supercharging progress.

- 8 public products with 1B+ active users
- 100s of teams working on models and applications leveraging generative AI
- Long term view of customer interactions across huge diversity of tasks
- Experience leverage a massive diversity of data sources

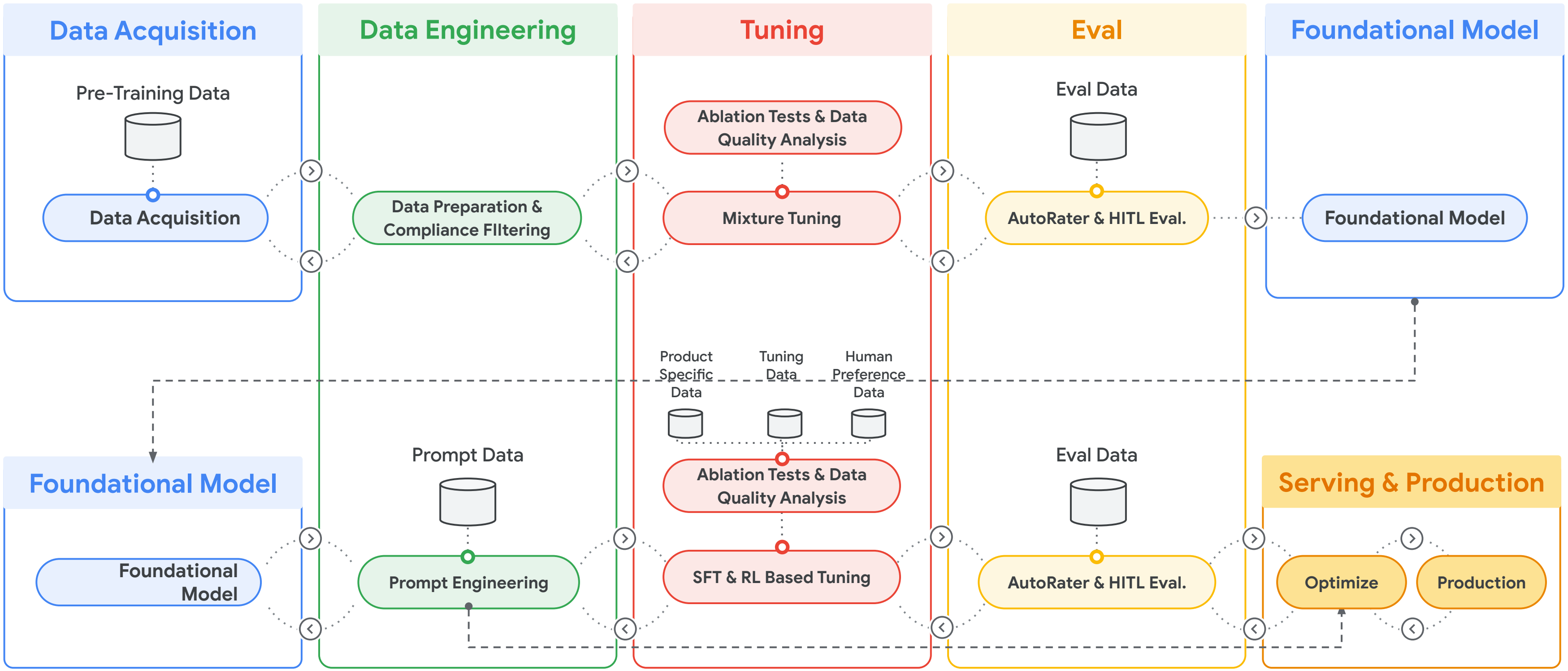
20 years of **ML Workflows** at Google



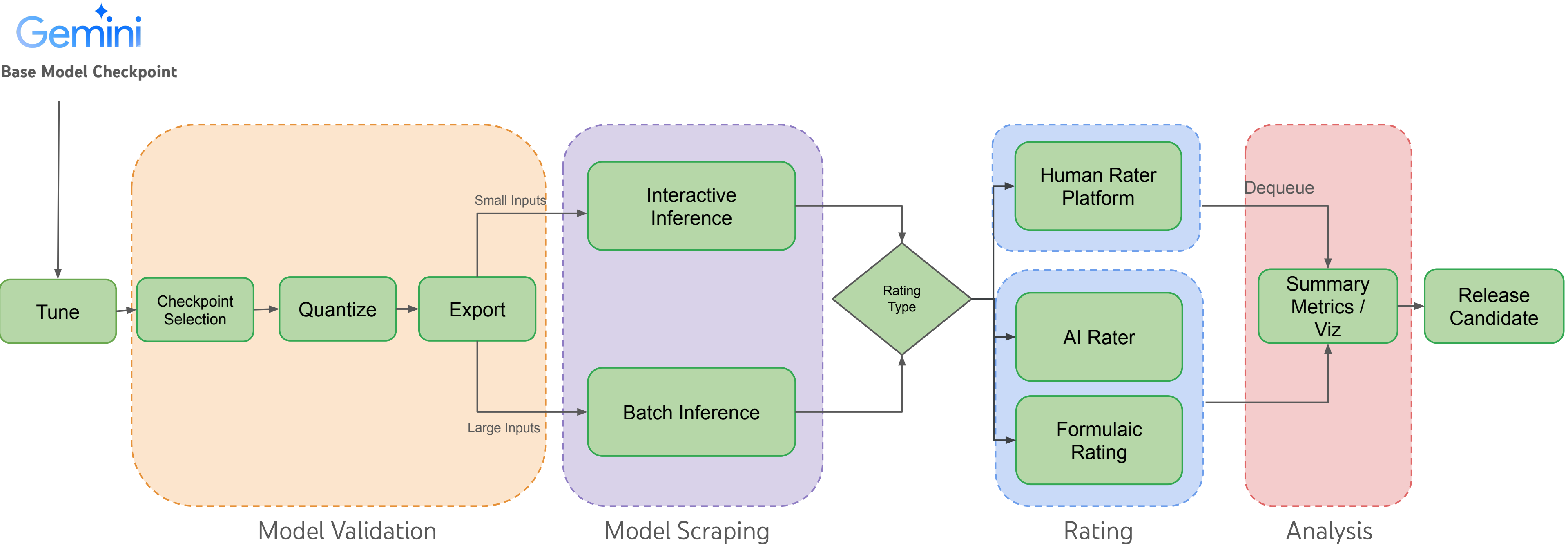
Emerging **AI Workflows** based on Generative AI



Emerging **AI Workflows** based on Generative AI



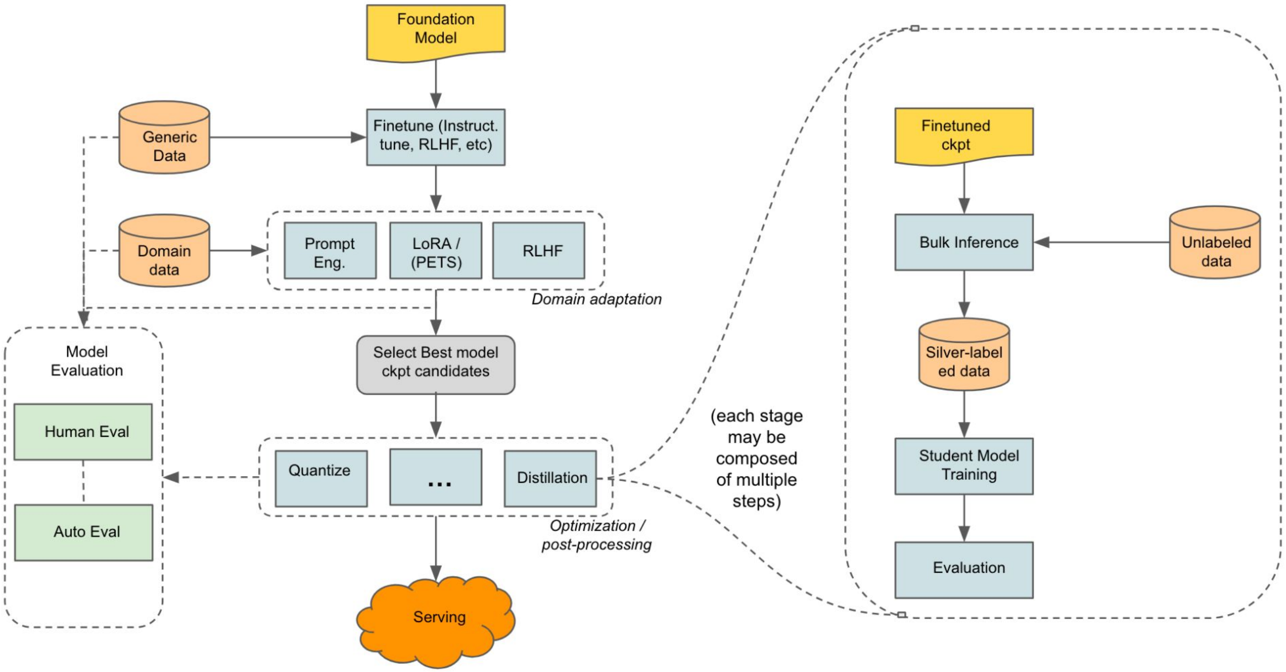
Gemini Adaptation Lifecycle (*A conceptual view*)



Gemini Adaptation Workflows

Types of Workflows:

1. **Training/Tuning workflows** - SFT, RLHF, LoRA, Prompt eng.
2. **Deployment workflows** - Checkpoint Selection, Quantization, Model export Validation
3. **Evaluation workflows** - human eval, auto eval
4. **Distillation workflows** - Student model training



Characteristics & Challenges with GenAI Workflows

1. Heterogenous training phases (SFT, RLHF, LoRA, etc)
   a. Specialized training / tuning frameworks & services to leverage TPU/GPUs efficiently.
2. JAX-based model training, tuning and serving. JAX is efficient.
   a. More JAX-based tooling is needed.
3. Many human touchpoints in human rating workflows
   a. Causing delays in end-to-end latency to have ratings needed.
4. Different types of evals and eval criteria.
   a. Evaluate models for Creativity, Factuality, Safety, Persona, etc. Sophisticated eval tools needed.
5. Different types of model scraping & inference needs - Interactive and Batch
   a. Covering different spectrums in efficiency frontier of cost vs responsiveness.
6. Model deployments and serving are complex
   a. Serving LoRA adapters, 2M context windows, etc.
7. LM application development patterns
   a. RAG, Agentic workflows need very different tooling and have different user journeys.

Thank you!

Questions?

LinkedIn:
https://www.linkedin.com/in/prakashchockalingam/

BEAM SUMMIT