Prompt Evaluation and Dataset Curation

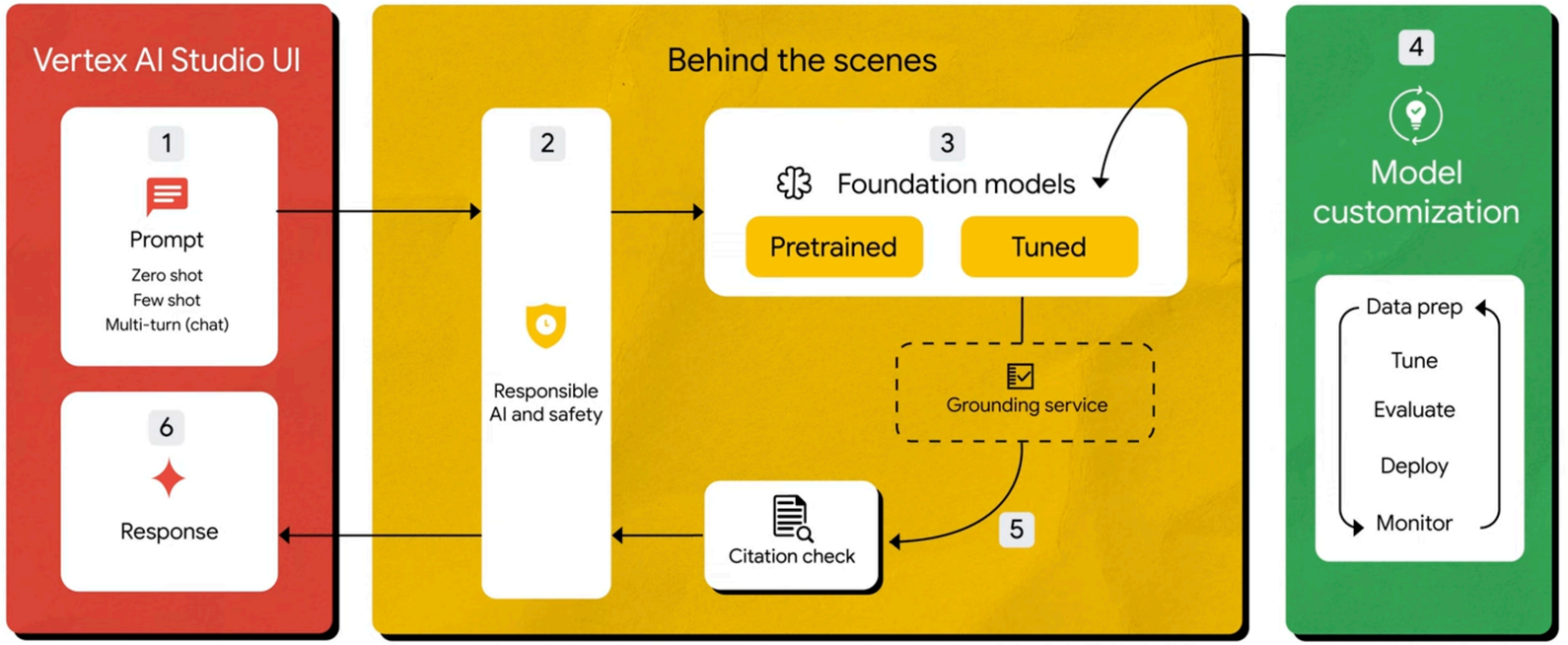
David Lee | TIM 175

Today

- Week 2 Lab + Week 3 Prelab
- Prompt evaluation and dataset curation
- Upcoming deliverables
  - Homework 3 (prelab) due TODAY at 11:59pm
  - Homework 3 (individual) due SATURDAY at 11:59pm
  - Homework 3 (team) due NEXT MONDAY at 11:59pm

From GenAI prototypes to production systems

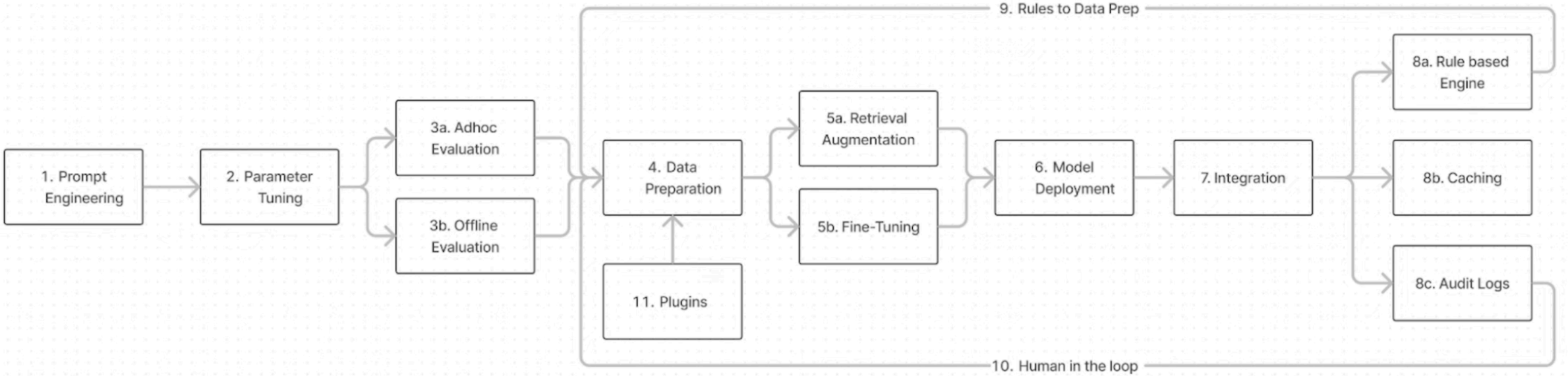
Generative AI workflow on Vertex AI



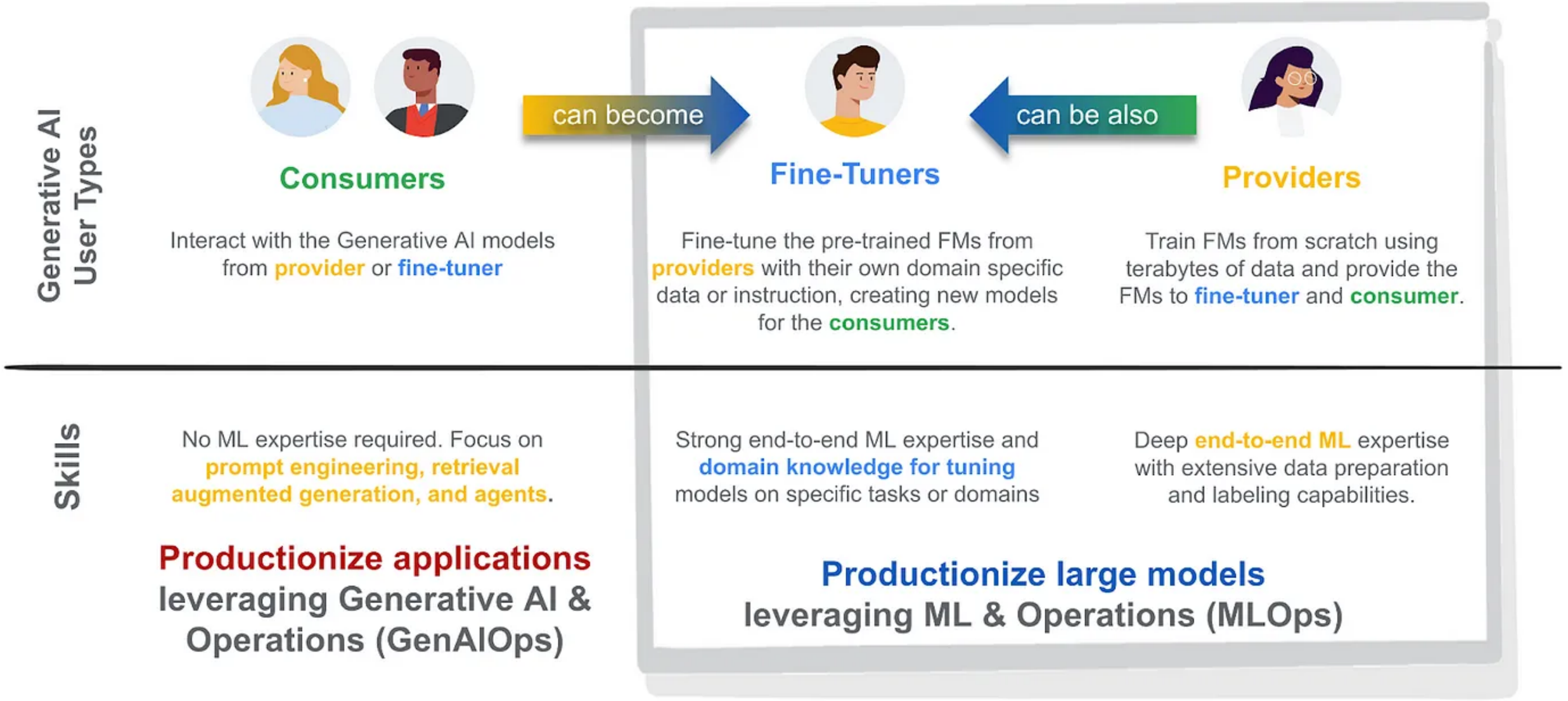
Introduction to Vertex AI Studio
https://www.cloudskillsboost.google/paths/183/course_templates/552/video/469801

the ***last mile*** problem

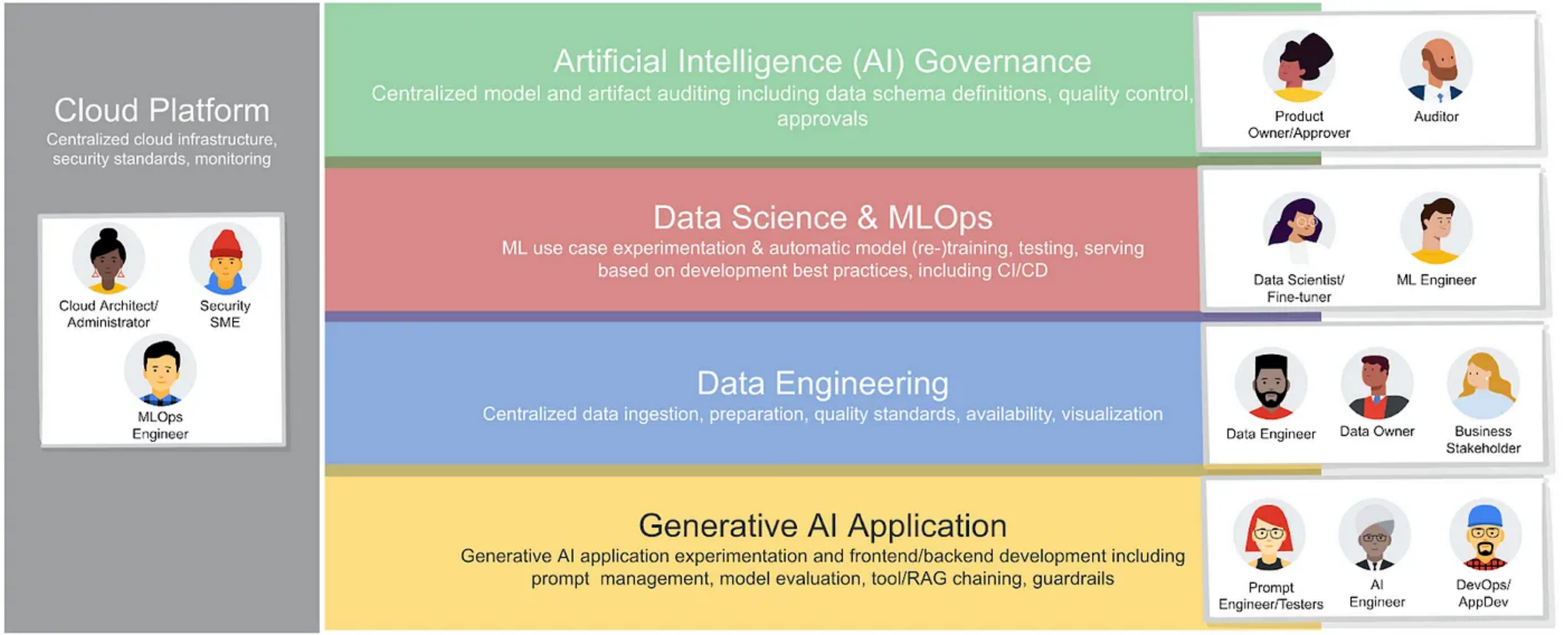
It's monumental effort to deploy reliable generative AI to production



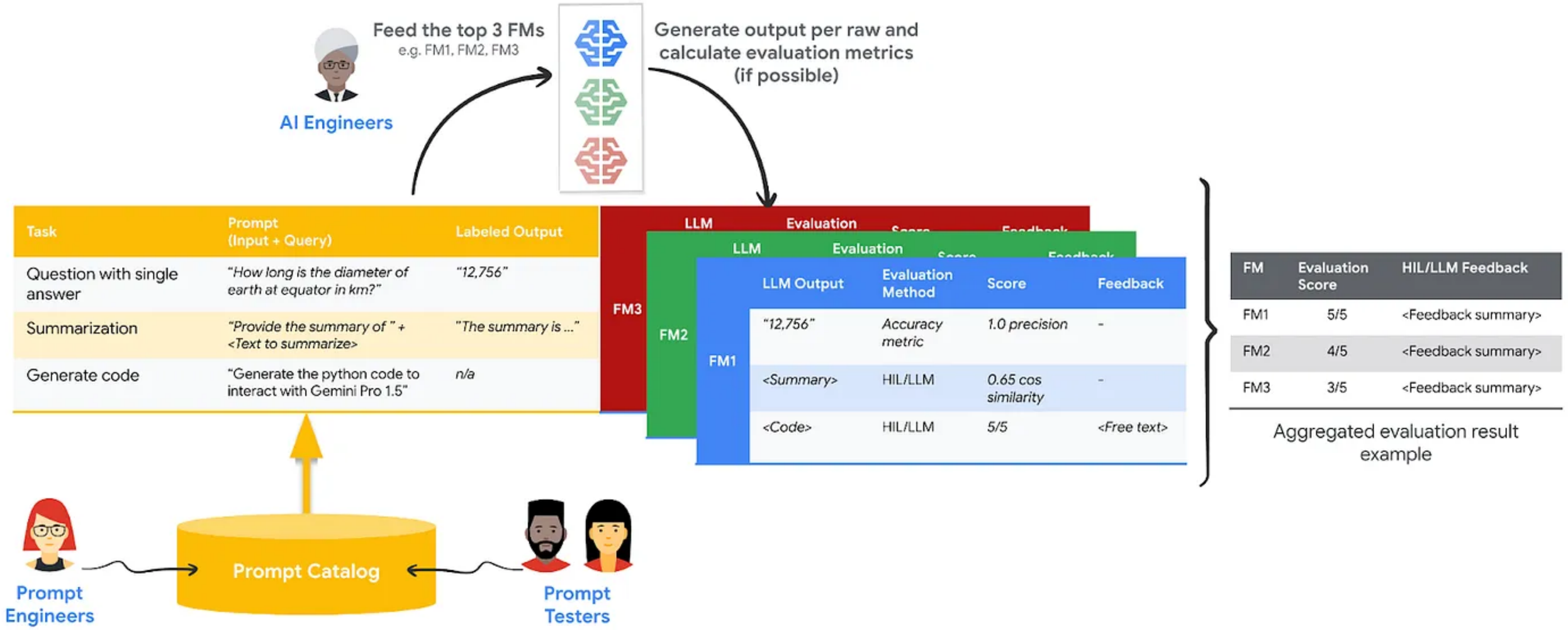
Deploying an application with Generative AI best practices
https://www.youtube.com/watch?v=dRf4DdA1o5c



GenAIOps: Operationalize Generative AI
https://medium.com/google-cloud/genaiops-operationalize-generative-ai-a-practical-guide-d5bedaa59d78



GenAIOps: Operationalize Generative AI
https://medium.com/google-cloud/genaiops-operationalize-generative-ai-a-practical-guide-d5bedaa59d78

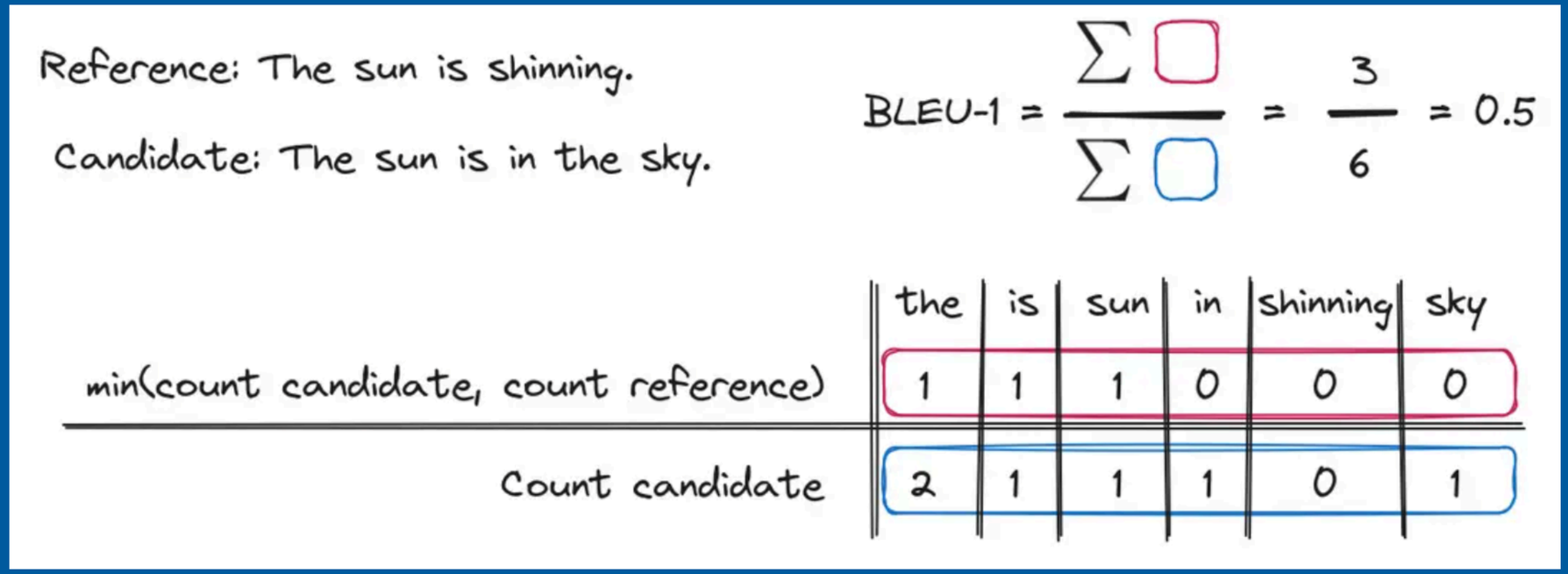


GenAIOps: Operationalize Generative AI
https://medium.com/google-cloud/genaiops-operationalize-generative-ai-a-practical-guide-d5bedaa59d78

How to evaluate?

Traditional NLP Metrics

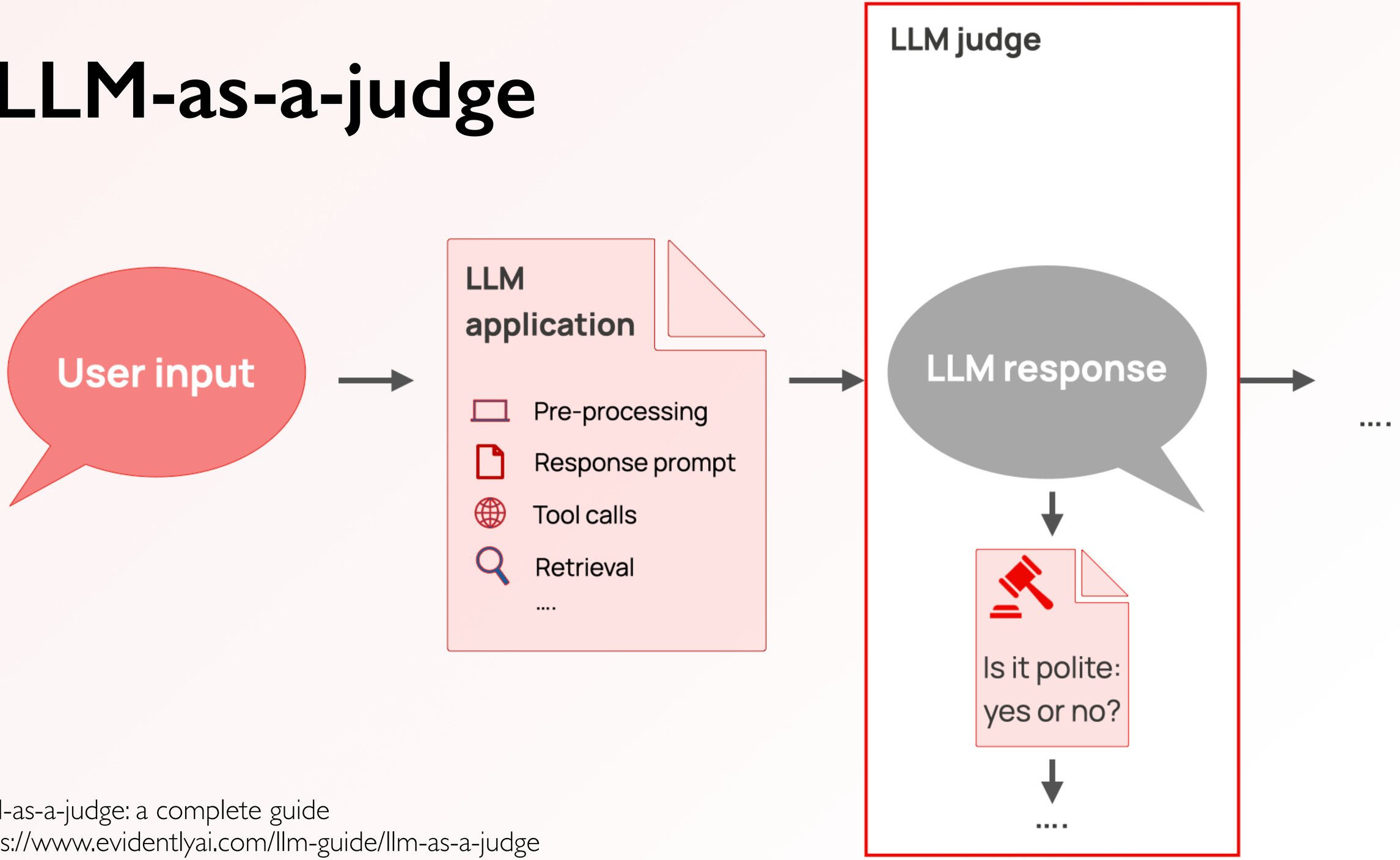
- BLEU - Bilingual Evaluation Understudy



- ROGUE - Recall-Oriented Understudy for Gisting Evaluation

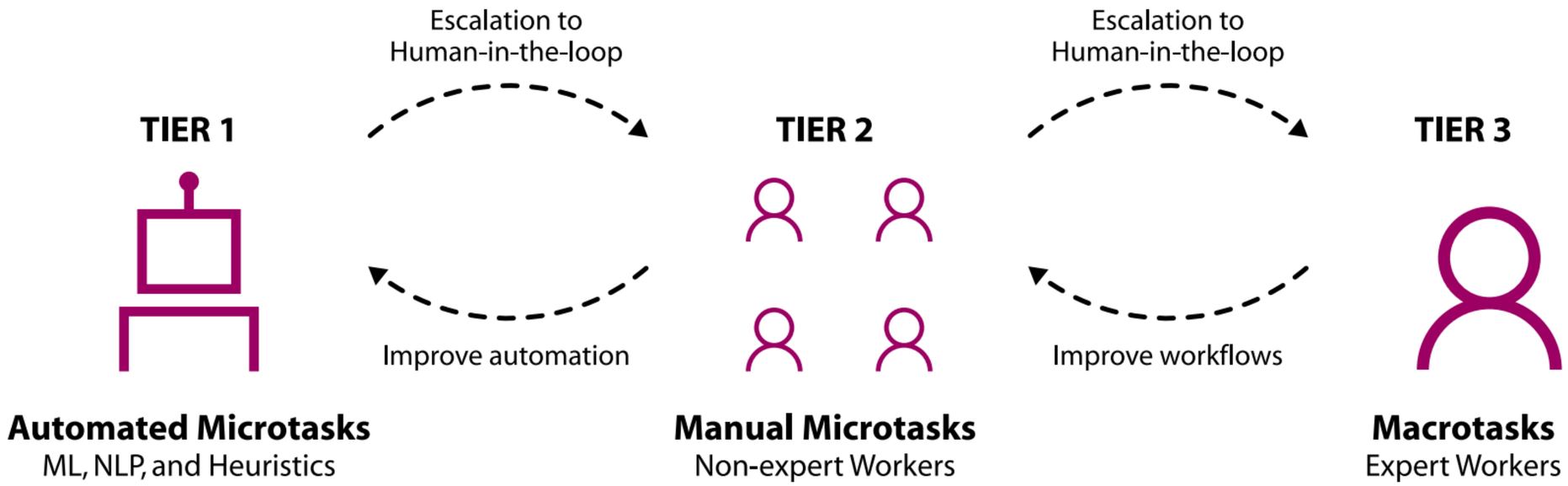
https://www.elastic.co/search-labs/blog/evaluating-rag-metrics

LLM-as-a-judge



LLM-as-a-judge: a complete guide
https://www.evidentlyai.com/llm-guide/llm-as-a-judge

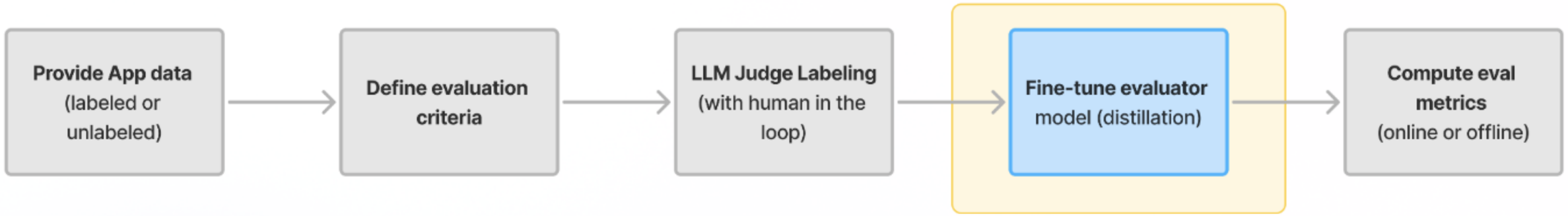
Human evaluation and annotation



How can a dataset be used for evaluation?

Training evaluator models

- Collect input prompts, potential model responses, and labels for whether the response satisfies some criteria
- Use this to train a model that is better at automated evaluation, useful for guardrails



https://docs.lastmileai.dev/autoeval/fine-tune

Fine tuning and RLHF
(Reinforcement Learning with Human Feedback)

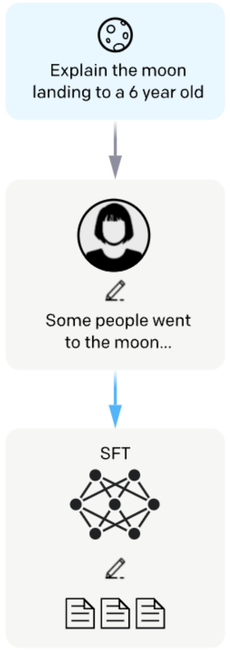
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



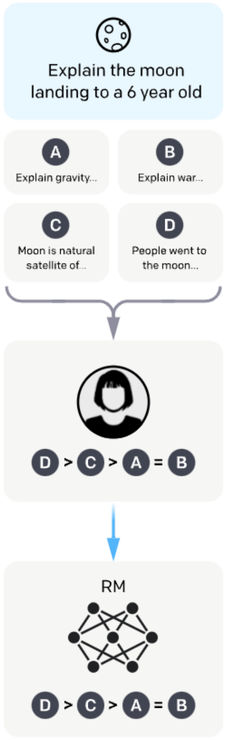
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

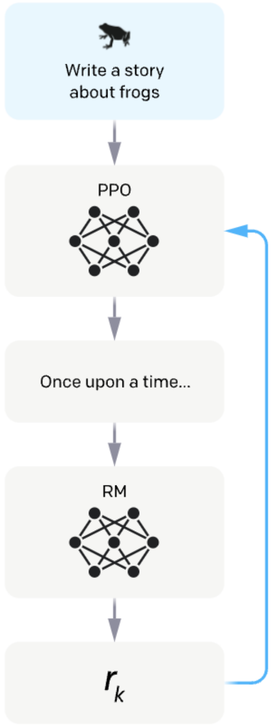
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



https://huyenchip.com/2023/05/02/rlhf.html

Prelab overview

Evaluation Metrics, Iteration, and Dataset Curation

TIM 175 WEEK 3 PRELAB

Brief Task Overview

Set up LastMile AI and create a copy of this document to work through the activities. For each activity:

- Capture screenshots of the prompts you used and the outputs generated by the model. Place these images directly **above** the analysis tables of each activity.
- Use the table to write a detailed analysis of your reflections from the activity.
- ALL NEW TEXT YOU ADD SHOULD BE IN RED TO MAKE IT EASY FOR US TO SEE.
- Create a Google sheets spreadsheet for documenting your prompt refinements using the following template: 3. Annotated Spreadsheet

Complete the following activities to understand methods for evaluating prompts and outputs:

1. Manually Evaluating Responses with Metrics
2. LLMs as a judge/ LLM as evaluators
3. Pairwise Comparison with LLM-as-a-judge
4. Creating an Annotated Dataset During Iteration
5. (optional) Experiment with Temperature

To develop an evaluator model for a given metric, we want to collect a set of inputs, possible generated outputs, and labels evaluating the generated output based on the metric criteria. Our evaluator models will be binary classifiers, so each data point should have the following

- **Inputs:** these are the unique parts in your prompt template, e.g. the user input or text you plug into your prompt,
- **Output**: a potential model generated output for those inputs, e.g. what was generated from one of your prior iterations,
- **Labels**: a 0 or 1 based on the metric evaluation criteria, e.g. whether the generated output was relevant or not relevant given the input and criteria for the relevancy metric,
    - **LLM evaluation labels**: we may store intermediate LLM-generated evaluations created using LLM-as-a-judge,
    - **Human evaluation labels**: these LLM evaluation labels are then verified/edited by humans manually to get the true/final labels.

It can also be helpful to curate other labels useful for fine-tuning the base model:

- **Ground truth**: the expected or ideal output you would like from the LLM, e.g. collected by experts writing responses that are a best attempt at the ideal response,
- **Human rank ordering**: a ranking of multiple generated outputs, used for Reinforcement Learning with Human Feedback (RLHF).

Inputs		Output	Rank	Elicits stories			Relevant and logical		
Survey question	Survey response	Generated follow-up question	Ranking of outputs	LLM label	Manual label	Manual label rationale	LLM label	Manual label	Manual label rationale
How should UCSC improve the student experience?	Housing is horrible! Make that better	Can you share a horrible experience you had related to housing?	1		1	Asks the user to sh		1	Relevant to horribl
		What about housing is horrible?	2		0	Asks to explain, bu		1	Relevant to horribl

Task: Practice creating an annotated dataset while iterating on a prompt for the following problem:

- **Given a user response to an open-ended survey question, we would like to generate a follow-up question that elicits rich stories and experiences from the respondent like in an interview,**

The inputs to the prompt would consist of:

- **The survey question**, e.g. “How should UCSC improve the student experience?”
- **The survey response**, e.g. “Housing is horrible! Make that better”

You should define a prompt that generates a follow-up question that elicits richer stories and experiences.

- **A good example**: “Can you share a horrible experience you had related to housing?”
- **Bad example**: “What about housing is horrible?” (this does not ask about the respondents stories and experiences, which is the focus of this prompt)

As you iterate on your prompt, you should use the following custom metrics:

- **Elicits stories**: we want follow-up questions that are likely to elicit rich stories and experiences from the respondent,
- **Relevant and logical**: we want follow-up questions that are relevant and logical follow-up questions given the response,