aws startups

CTO Community

Eric Grudzien

Sr. Startups Solutions Architect
Machine Learning Core

What we will cover today

1. Prompt Engineering
2. Evaluations
3. Evaluations Resources
4. Q&A

PROMPT ENGINEERING

Prompt

- A prompt is the information you pass into a large language model to elicit a response.

- This includes:
  - Task context
  - Data
  - Conversation / action history
  - Instructions
  - Examples
  - And more!

Prompt

Prompt: How many dogs are in this picture?

Answer: 10

Prompt

Prompt: You have perfect vision and pay great attention to detail which makes you an expert at counting objects in images. How many dogs are in this picture? Before providing the answer in <answer> tags, think step by step in <thinking> tags and analyze every part of the image.

Answer: 9

Prompt

We want to de-identify some text by removing all personally identifiable information from this text so that is can be shared safely with external contractors.

It is very important that PII such as names, phone numbers, and home and email addresses get replaced with XXX.

Here is the txt you should process: “John Doe is a Solutions Architect at AWS. He can be reached at 123-555-1212 or john.doe@amazon.com”

Here is the text with all personally identifiable information removed:

“XXX is a Solutions Architect at AWS. He can be reached at XXX or XXX”

Prompt

We want to de-identify some text by removing all personally identifiable information from this text so that is can be shared safely with external contractors.

It is very important that PII such as names, phone numbers, and home and email addresses get replaced with XXX.

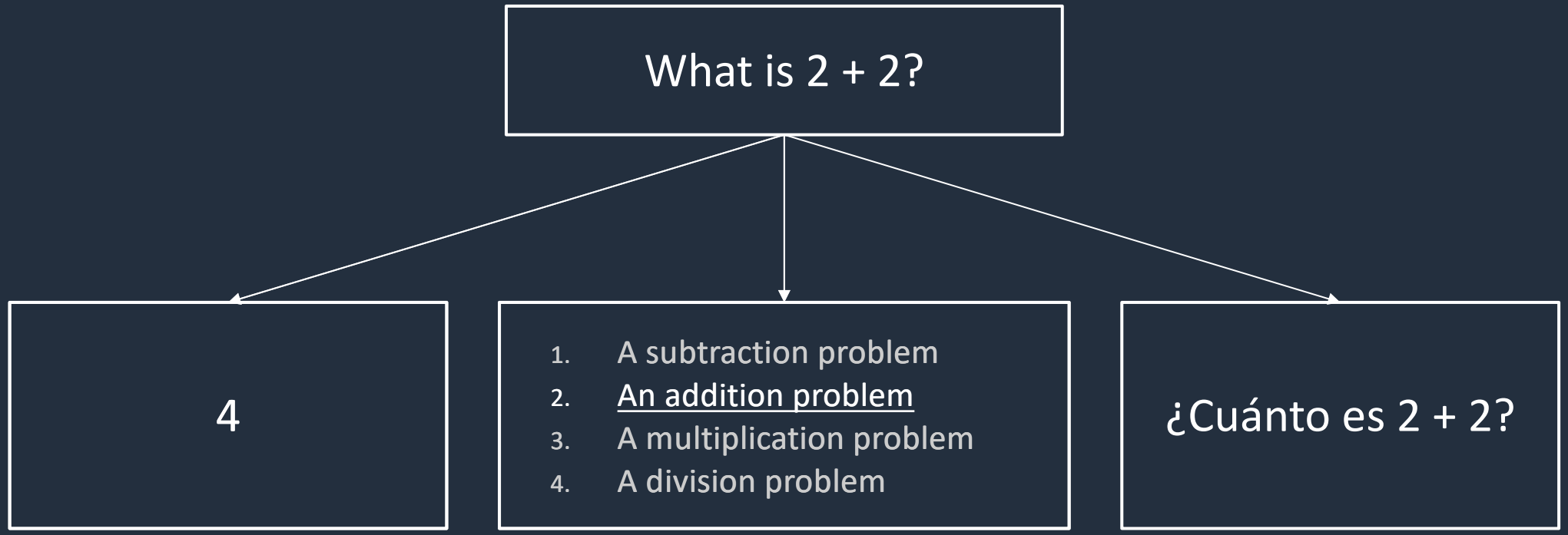
Here is the txt you should process: “John Doe is a Solutions Architect at AWS. He can be reached at 123-5 55-1212 or john.doe@

amazon.com”

Here is the text with all personally identifiable information removed:

“XXX is a Solutions Architect at AWS. He can be reached at XXX or XXX@amazon.com”

What is prompt engineering?



Prompt engineering is the process of **controlling model behavior** by **optimizing your prompt to elicit high performing LLM responses** (as assessed by rigorous evaluations tailored to your use case).

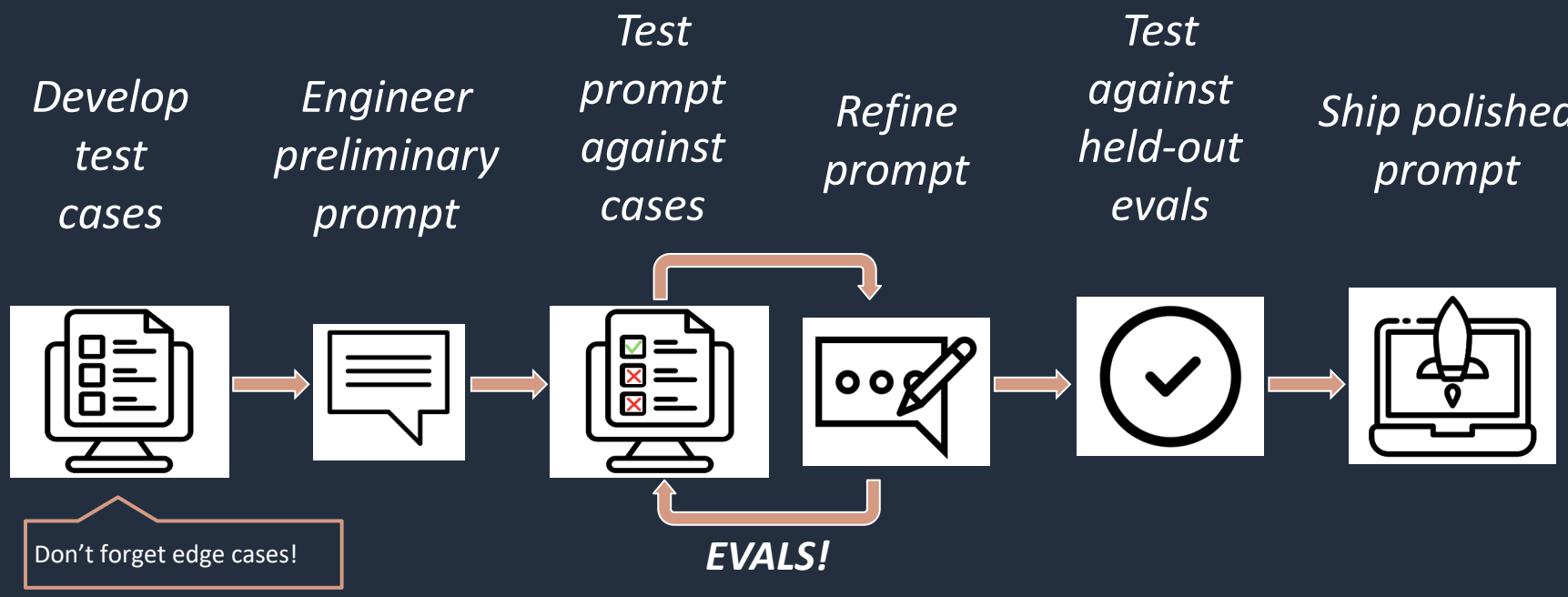
Prompt Composition

1. Task context
2. Tone context
3. Background data, documents, and images
4. Detailed task description & rules
5. Examples
6. Conversation history
7. Immediate task description or request
8. Thinking step by step / take a deep breath
9. Output formatting
10. Prefilled response (if any)

USER	You will be acting as an AI career coach named Joe created by the company AdAstra Careers. Your goal is to give career advice to users. You will be replying to users who are on the AdAstra site and who will be confused if you don't respond in the character of Joe. You should maintain a friendly customer service tone. Here is the career guidance document you should reference when answering the user: <guide>{{DOCUMENT}}</guide> Here are some important rules for the interaction: - Always stay in character, as Joe, an AI from AdAstra careers - If you are unsure how to respond, say “Sorry, I didn’t understand that. Could you repeat the question?” - If someone asks something irrelevant, say, “Sorry, I am Joe and I give career advice. Do you have a career question today I can help you with?” Here is an example of how to respond in a standard interaction: <example> User: Hi, how were you created and what do you do? Joe: Hello! My name is Joe, and I was created by AdAstra Careers to give career advice. What can I help you with today? </example> Here is the conversation history (between the user and you) prior to the question. It could be empty if there is no history: <history> {{HISTORY}} </history> Here is the user’s question: <question> {{QUESTION}} </question> How do you respond to the user’s question? Think about your answer first before you respond. Put your response in <response></response> tags.
Assistant (Pre-fill)	<response>

How to engineer a good prompt

Empirical science*: always test your prompts & iterate often!



* studying and learning about the world through observation and experimentation

Prompt Cheat Code

- Metaprompt
  - Helper for Claude to generate a high-quality prompts tailored to your specific tasks.
  - Useful as a "getting started" tool
  - Use as method to generate multiple prompt versions for a given task,

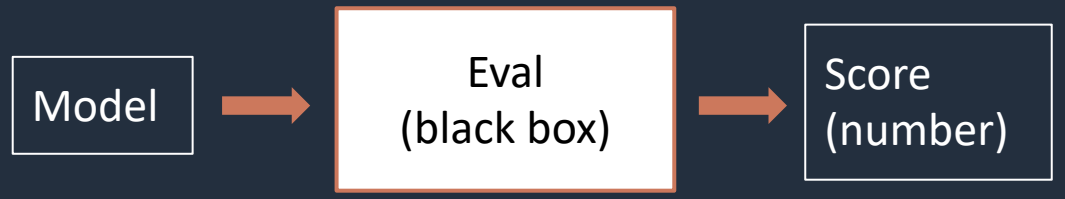
EMPIRICAL EVALUATIONS

Evals

- HELM - https://crfm.stanford.edu/helm/lite/latest/#/

Evaluation Overview

- An evaluation or eval in prompt engineering refers to the process of evaluating an LLM's performance on a given dataset after it has been trained
- Use evals to:
  - Assess a model's knowledge of a specific domain or capability on a given task
  - Measure progress or change when shifting between model generations



Evaluation Overview (what it is not)

- Evaluation is not Benchmarking
- Use Benchmarking to:
  - Inference latency, Throughput, Cost per transaction
  - Infrastructure
- “What is the dollar cost per transaction for a given generative AI workload that serves a given number of users while keeping the response time under a target threshold?”
- “What is the minimum number of instances N, of most cost optimal instance type T, that are needed to serve a workload W while keeping the average transaction latency under L seconds?”
- FMBench Tool + Workshop (see SA; links at end)

What does an eval look like?

Example input	Golden output	Rubric		Model response		Eval score
The entire prompt or only the variable content	*An ideal response to grade against*	*Guidelines for grading a model's actual response*	+	*The model's latest response*	=	*A numerical score assessing the model's response*
Give me a delicious recipe for [cornbread]	[Ideal recipe]	1. Includes cornmeal (*auto 0 if not*) 2. Mentions mixing tool…	+	Here's a recipe for cornbread: Ingredients: - Cornmeal - …	=	- Includes cornmeal - Mentions spoon … **9/10**

TYPES OF EVALUATIONS

Example: multiple choice question eval (MCQ)

- Simplest
- Closed form questions
- Clear answer key
- Easy to automate

Prompt	How many days are there in a week? (A) Five (B) Six (C) Seven (D) None of the above
LLM Response	C

Example: exact match (EM) or string match

Exact match:

Prompt	What is the white powder substance that is used to make bread?
LLM Response	flour
Correct Answer	flour
Score	CORRECT

String match:

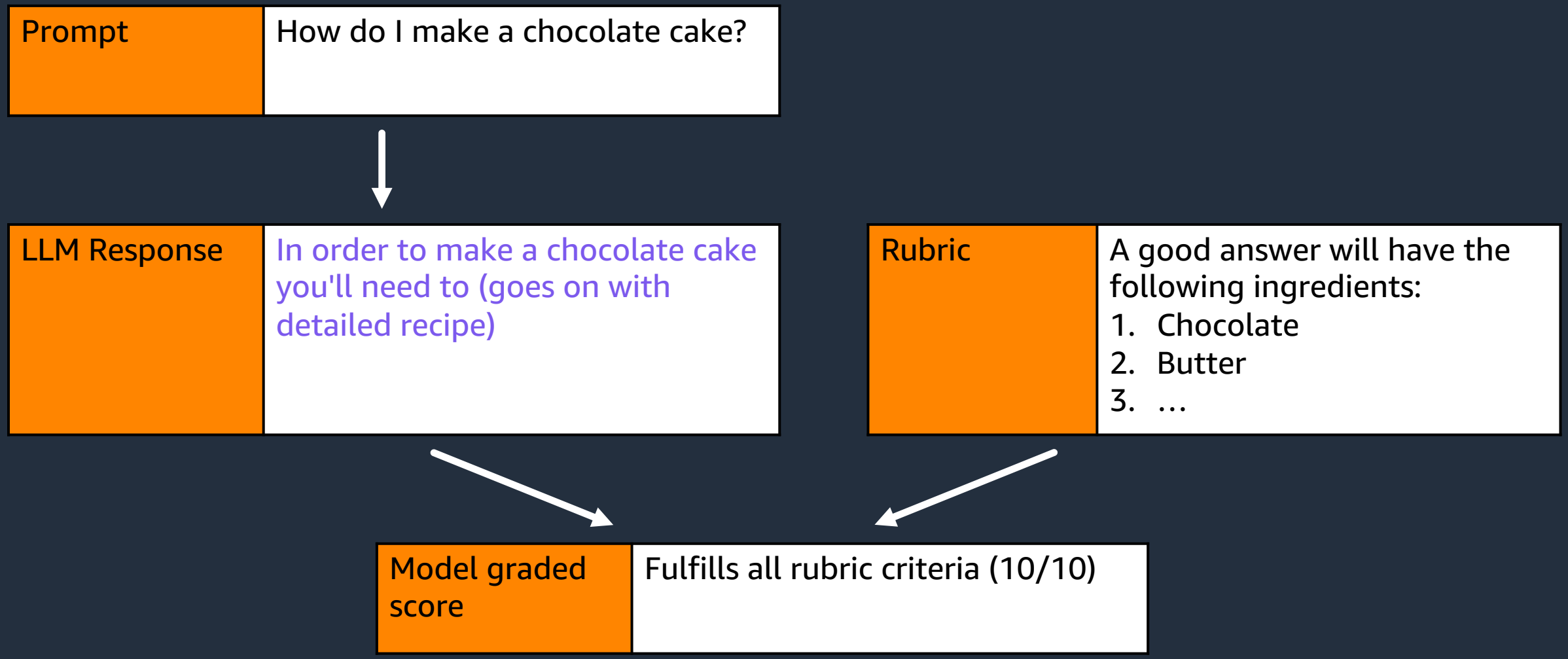
Prompt	What do you think about politics?
LLM Response	Well, I think that country ABCD has …
Correct Answer	“ABCD” in response
Score	response.contains(ABCD) -> CORRECT

Example: open answer eval (OA) - by humans

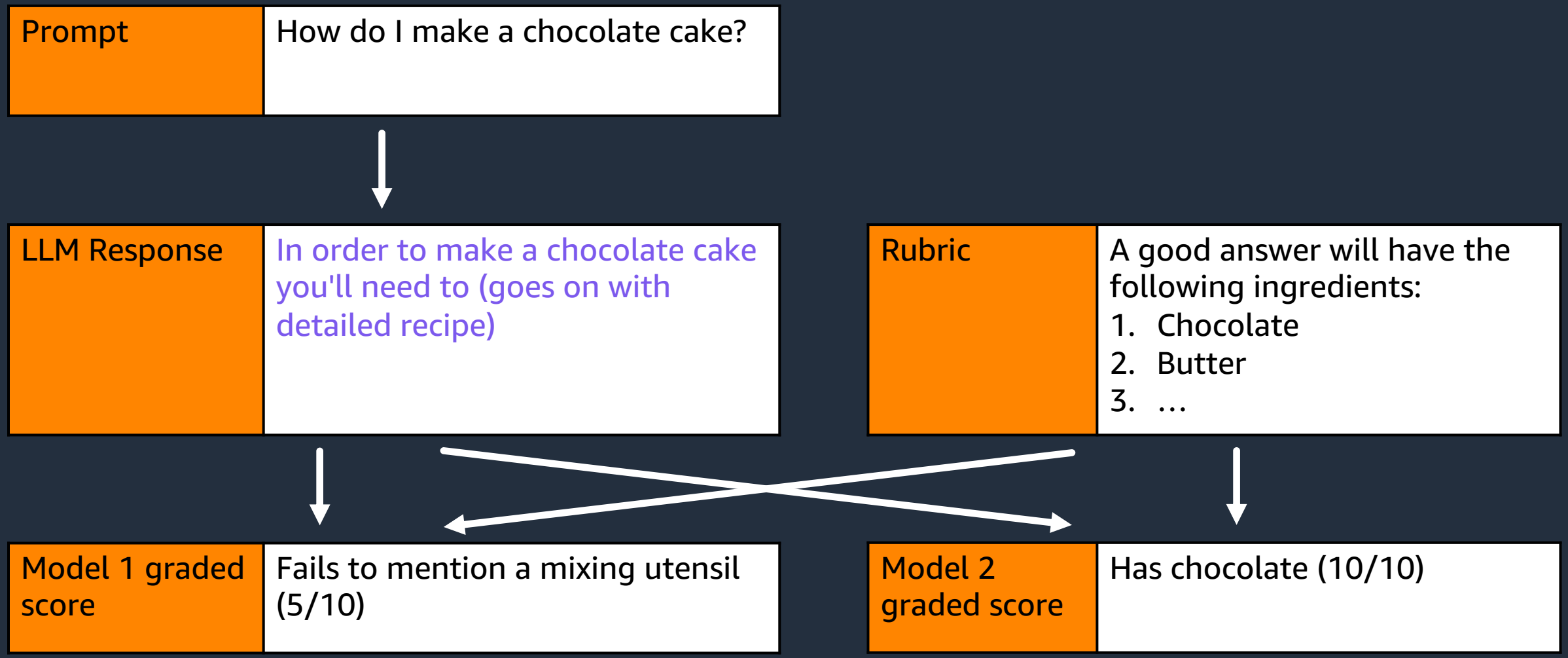
- Question is open ended
- Great for assessing:
  - more advanced knowledge
  - tacit knowledge
  - multiple possible solutions
  - multi-step processes
- Humans can grade this eval
- But models can do it more scalably! Just less accurately
- Needs a very clear rubric

Prompt	How do I make a chocolate cake?
LLM Response	In order to make a chocolate cake you'll need to (goes on with detailed recipe)
Human score (rubric-based)	3/10
Rubric	Has butter Has flour Has chocolate … Doesn't have meat

Example: open answer eval (OA) - by models



Example: open answer eval (OA) - by multiple models



Some evals are better than others

Less desirable eval qualities:

- **Open-ended**
- Requires **human-judgment**
- Higher quality but **very low volume**

More desirable eval qualities:

- Very **detailed & specific**
- Fully **automatable**
- **High volume** even if lower quality

EVALS RESOURCES

Evals Resources

- Building Evals Notebook
- Prompt Engineering with Anthropic's Claude v3 Workshop
- Bedrock Model Evaluations

IN CLOSING

Closing Thought - Threat Modeling

- AWS re:Invent 2023 - Threat modeling your generative AI workload to evaluate security risk (SEC214)
  - https://www.youtube.com/watch?v=TtRFQPlRYK4

Links

- HELM
  - https://crfm.stanford.edu/helm/lite/latest/#/
- Building Evals
  - https://github.com/anthropics/anthropic-cookbook/blob/main/misc/building_evals.ipynb
- Prompt Engineering with Anthropic's Claude v3 Workshop
  - https://catalog.us-east-1.prod.workshops.aws/workshops/0644c9e9-5b82-45f2-8835-3b5aa30b1848/en-US/lessons/lab-10-3-empirical-performance-evaluations
- FMBench
  - https://github.com/aws-samples/foundation-model-benchmarking-tool

Links

- Bedrock Model Evaluations
  - https://docs.aws.amazon.com/bedrock/latest/userguide/model-evaluation.html
- Metaprompt
  - https://gitlab.aws.dev/3p-models/community-samples/-/blob/main/anthropic/Metaprompt_generator_bedrock.ipynb?ref_type=heads

Thank you!

Survey: https://pulse.aws/survey/PSVIXHRQ

Eric Grudzien
Sr. Startups SA, Machine Learning Core