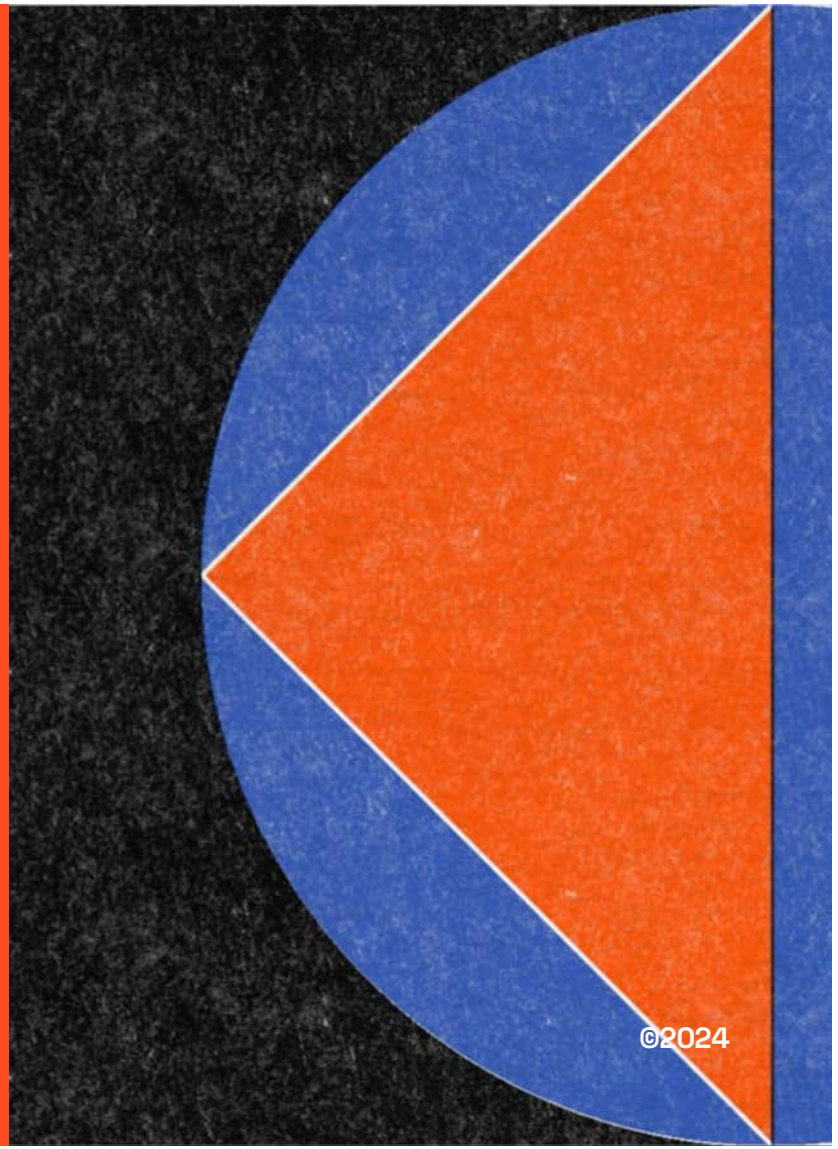
Mastering Evals

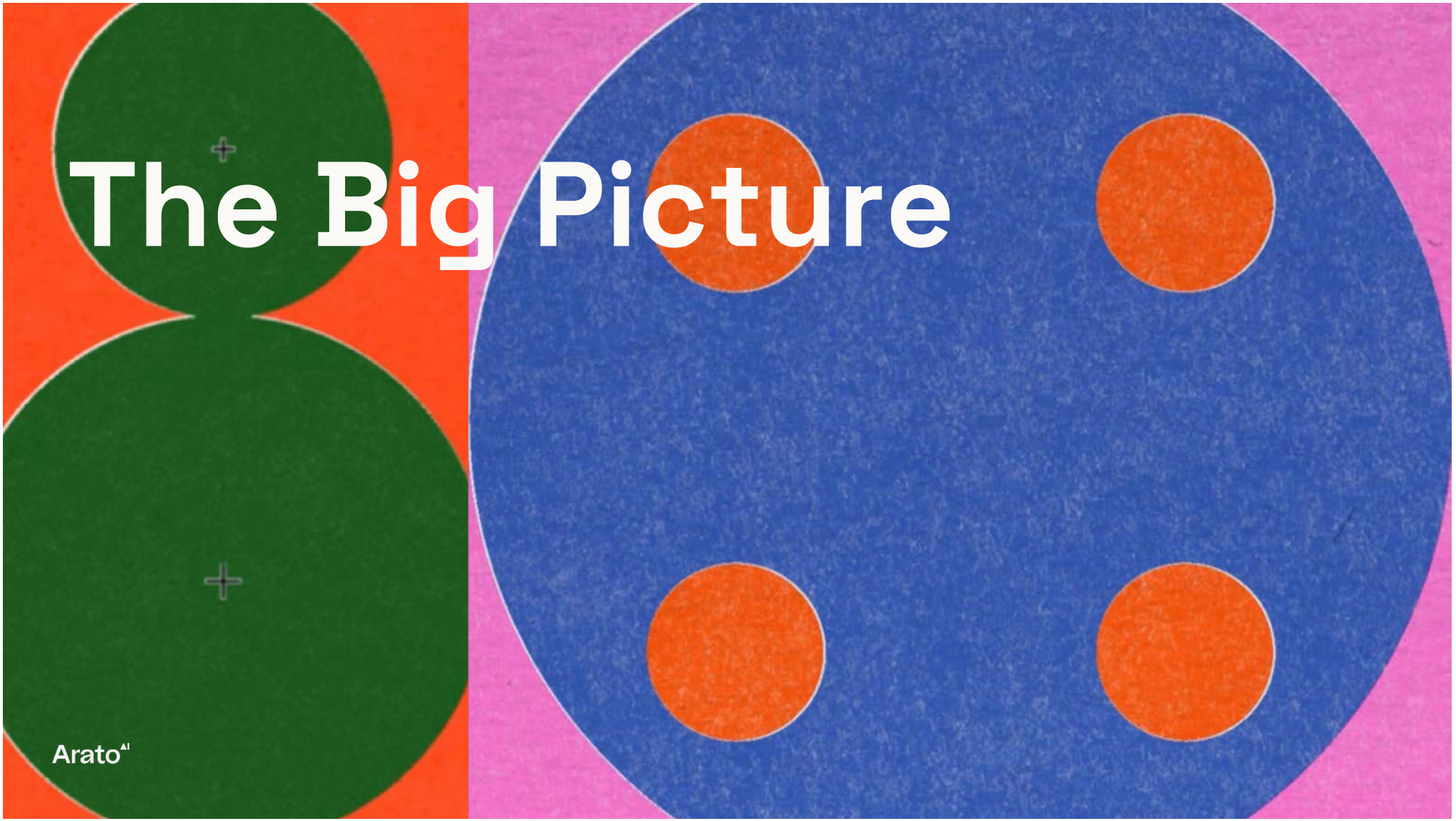
The Key to Building Effective LLM Based Apps

Hilik Paz, Arato.ai

Arato



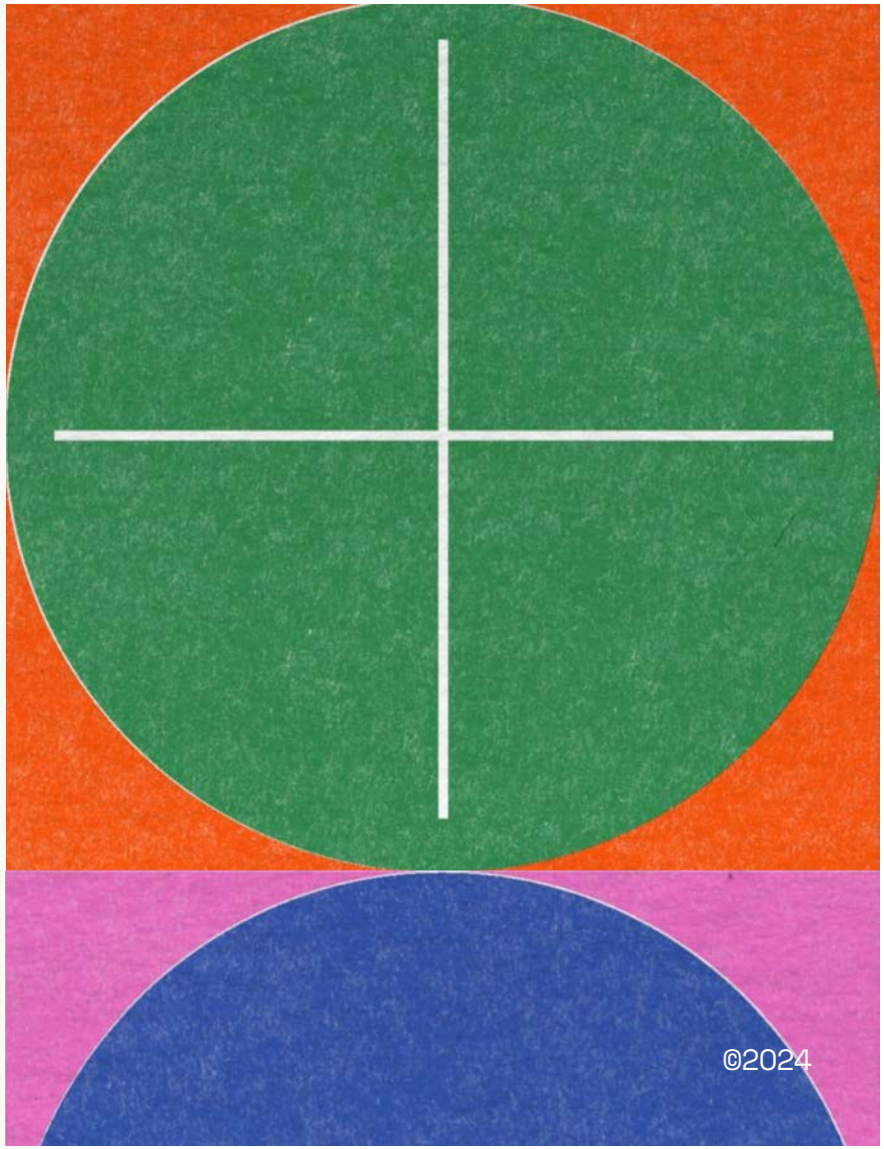
The Big Picture



What are Evals?

Methods for **systematically** determine the performance of LLMs and Models.

Measurable and **Repeatable**, enabling data driven iterative improvements



Arato

What Changed?

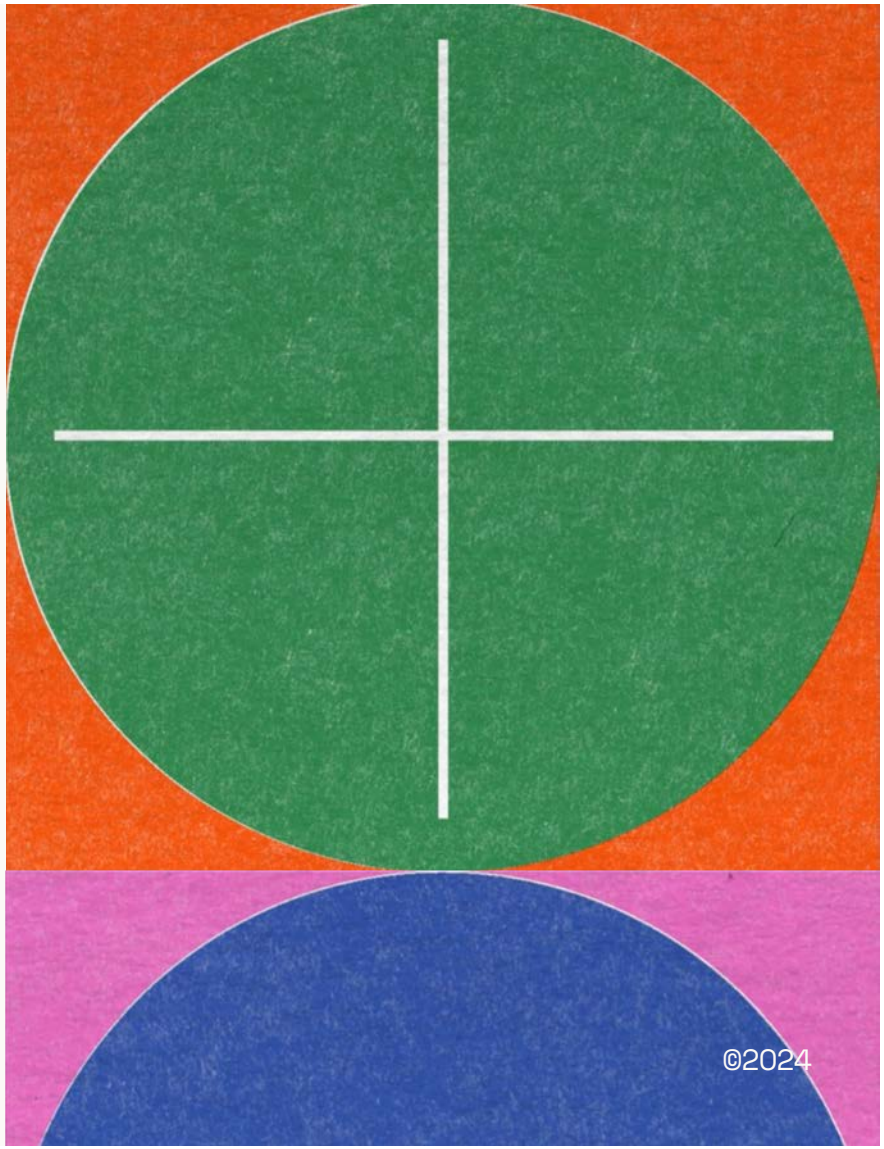
Traditional software testing vs. LLM evaluations

Clear **boundaries** vs Complex nuanced output.

Deterministic vs Probabilistic
Good or Bad is no longer binary

From **Pass/Fail to Scoring**

Arato AI

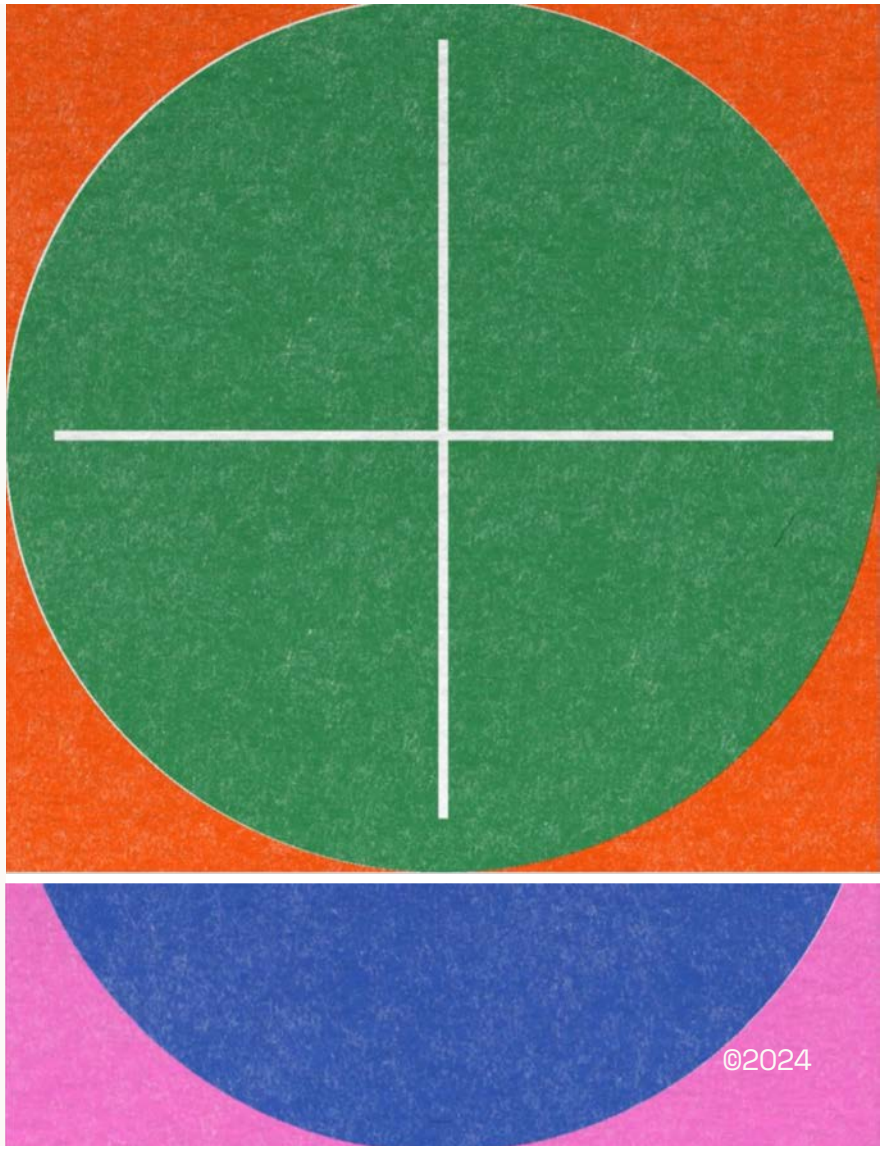


Why Evals Matter?

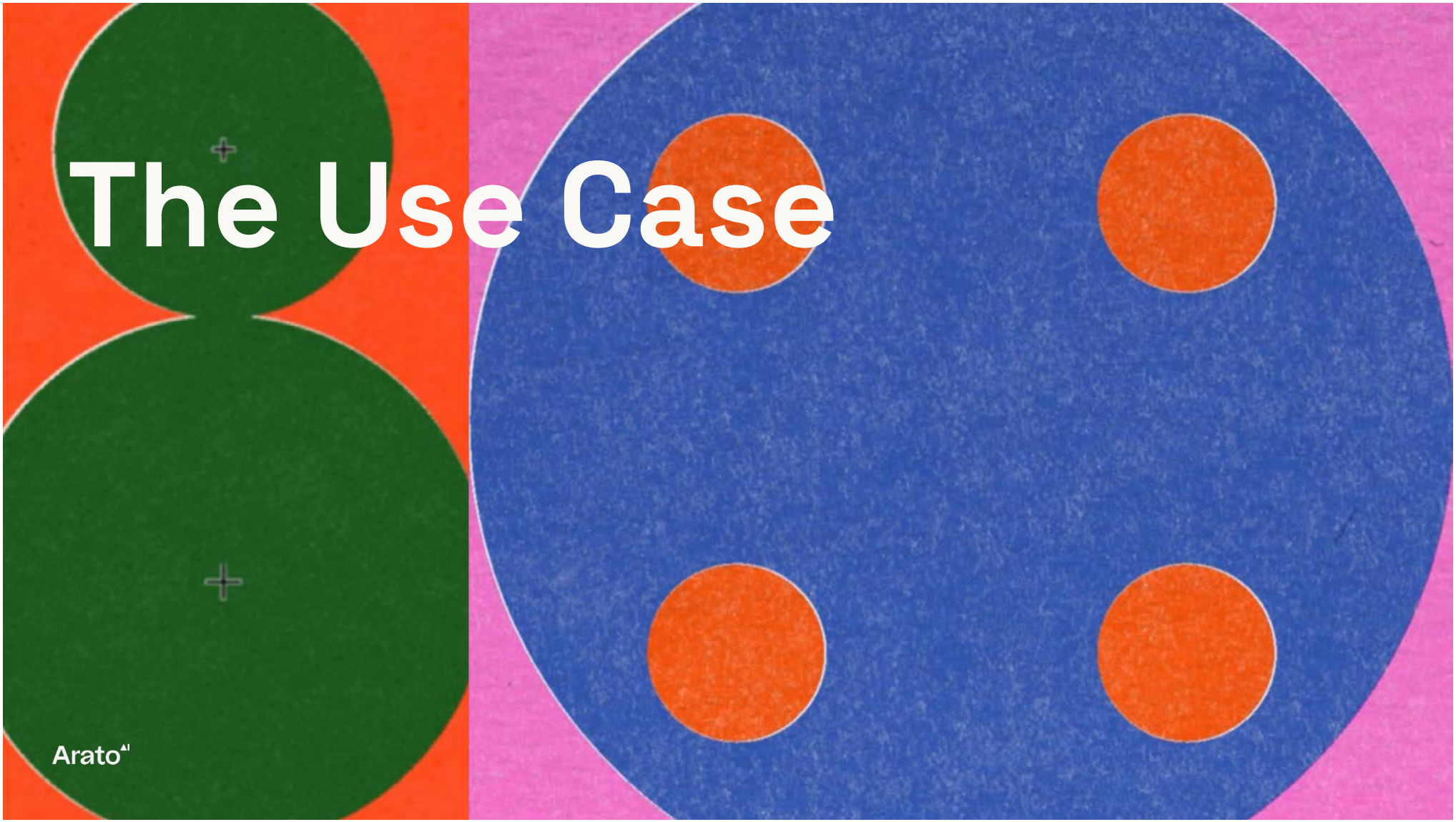
Ensure LLM outputs **align** with business objectives

Reduce the **risk** of unforeseen issues in production

Enable **measurable** outcome.



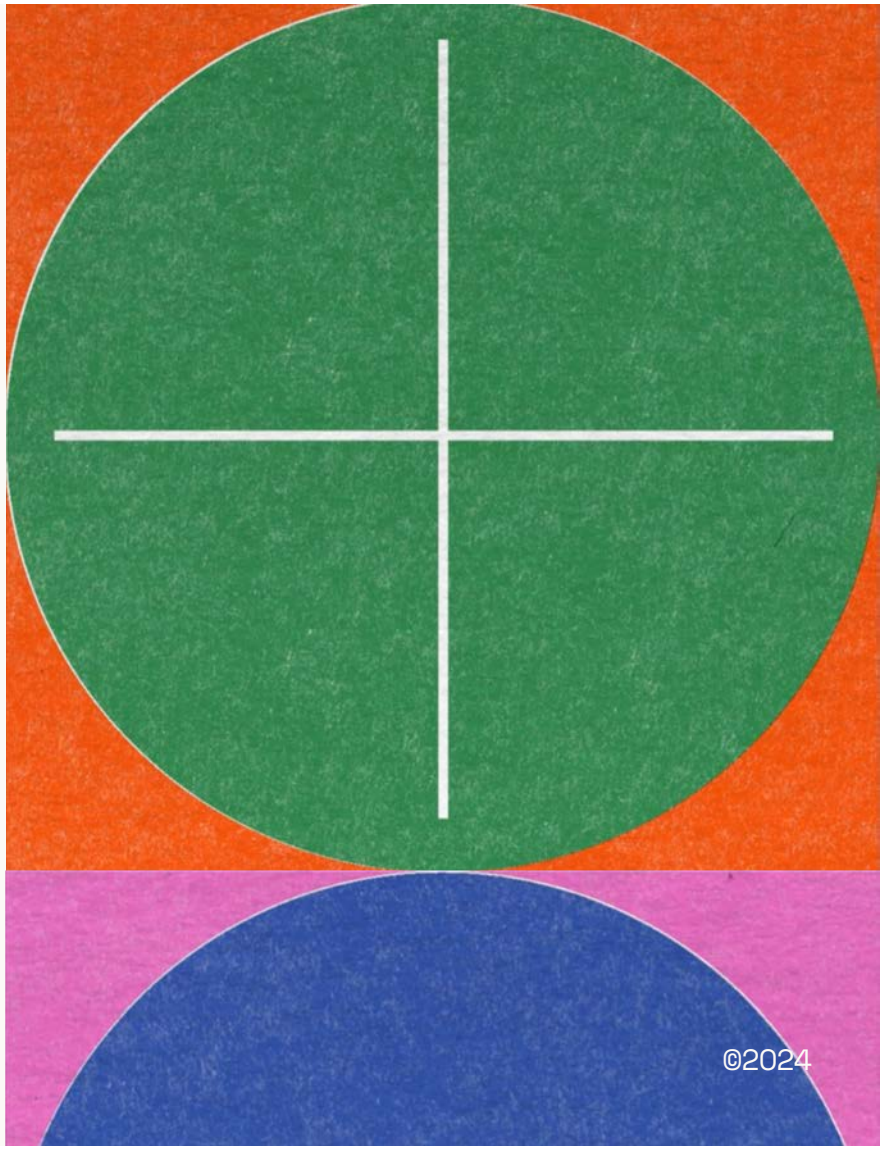
Arato AI


The Use Case
Arato AI

The AppMaster Support Bot

- AI-powered support bot for AppMaster users
- Answers **technical** questions using a **knowledge base**
- Goal: Provide **accurate**, **fast**, and **context-aware** responses, with a **polite** and friendly tone of voice.

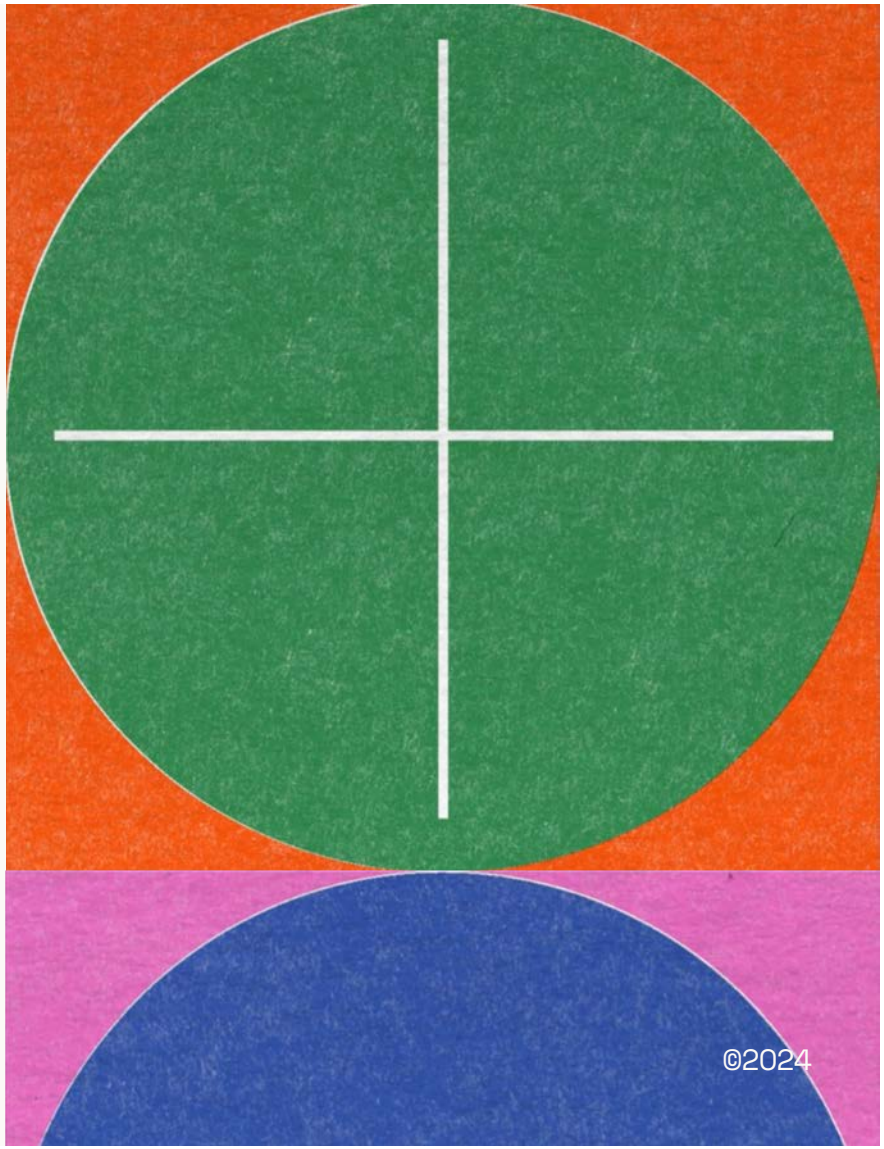
Arato[AI]

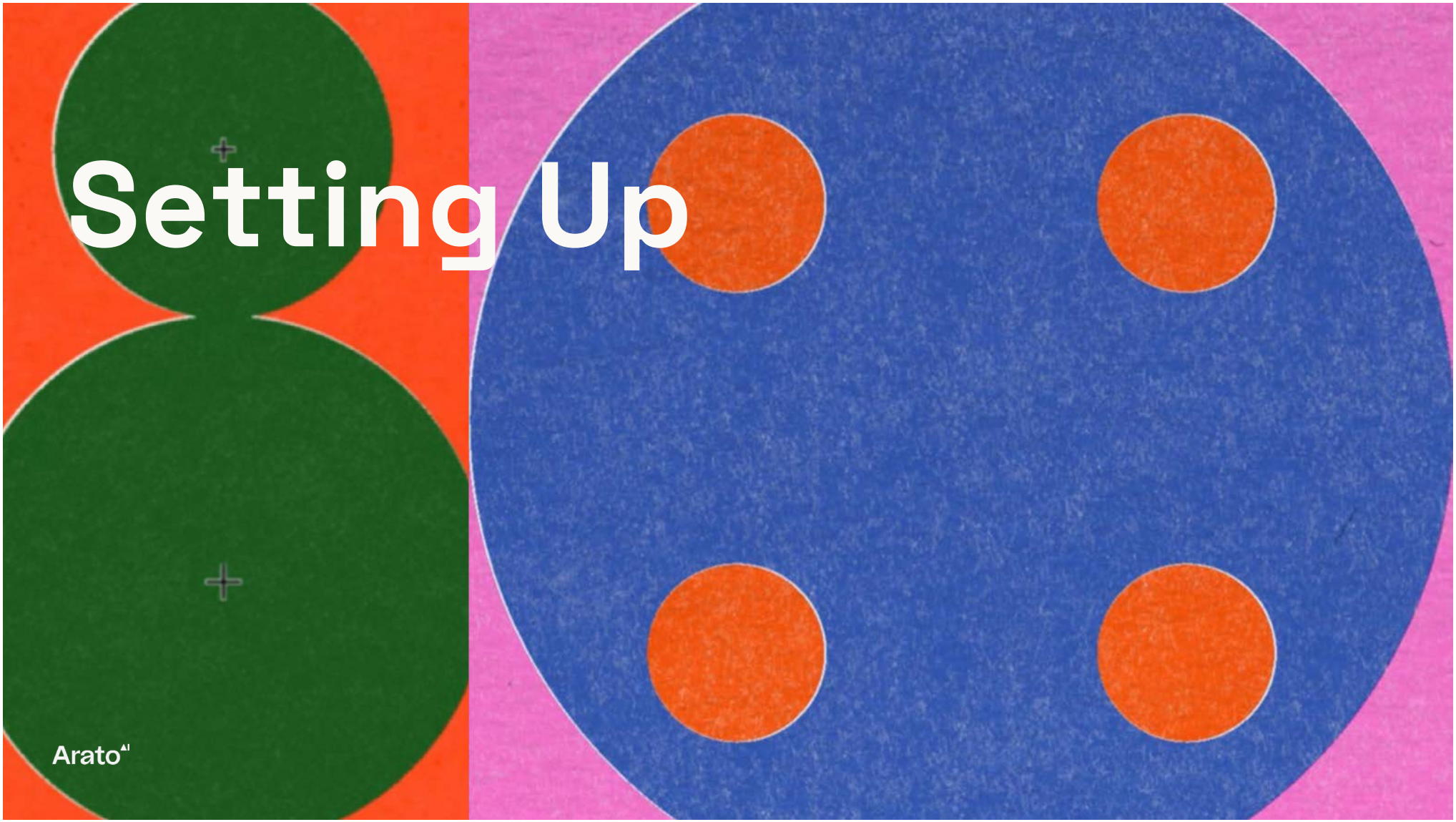


Defining the Initial Prompt

We start with a **simple** and clear prompt.
We set the role
Provide clear instructions
And set the expected output

But is this enough? We need a way to **measure** its effectiveness



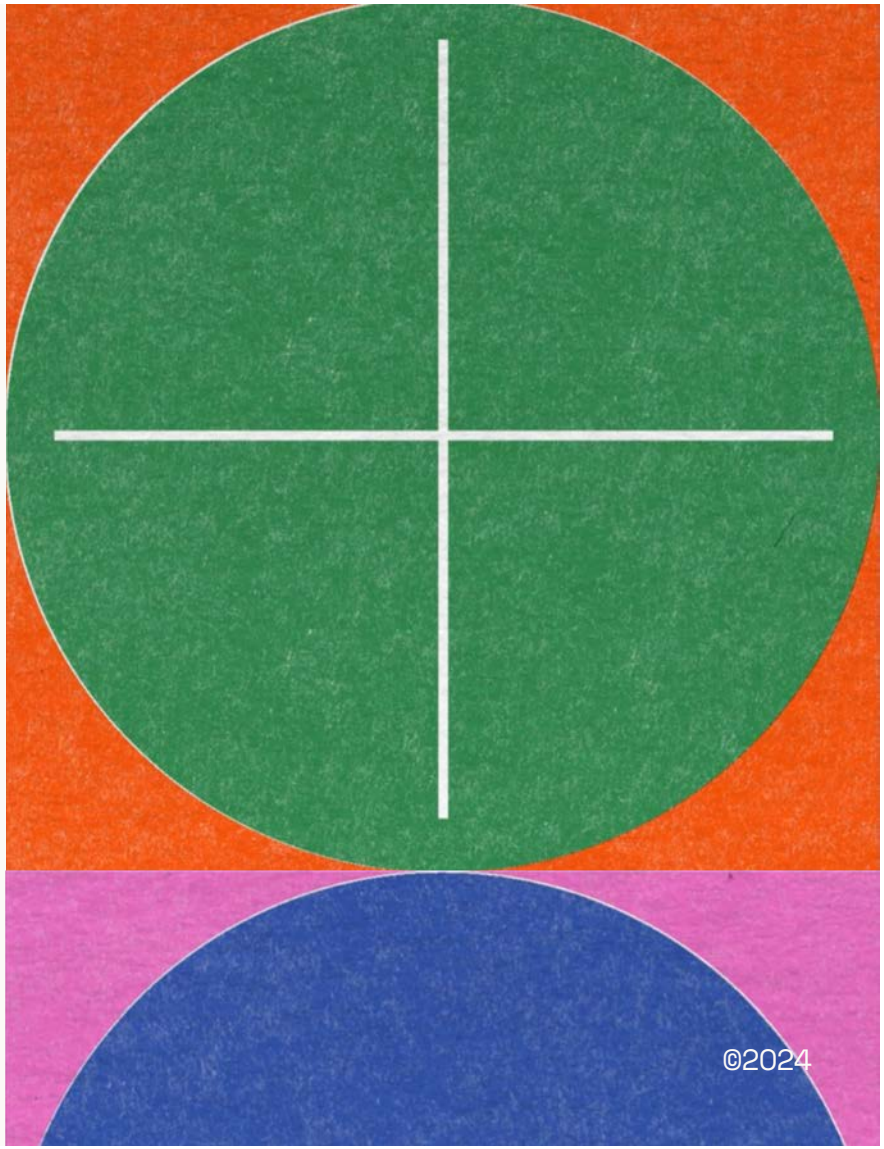

Setting Up
Arato AI

How Do We Measure Success?

Define key evaluation criteria

Choose evaluation methods

Establish test cases

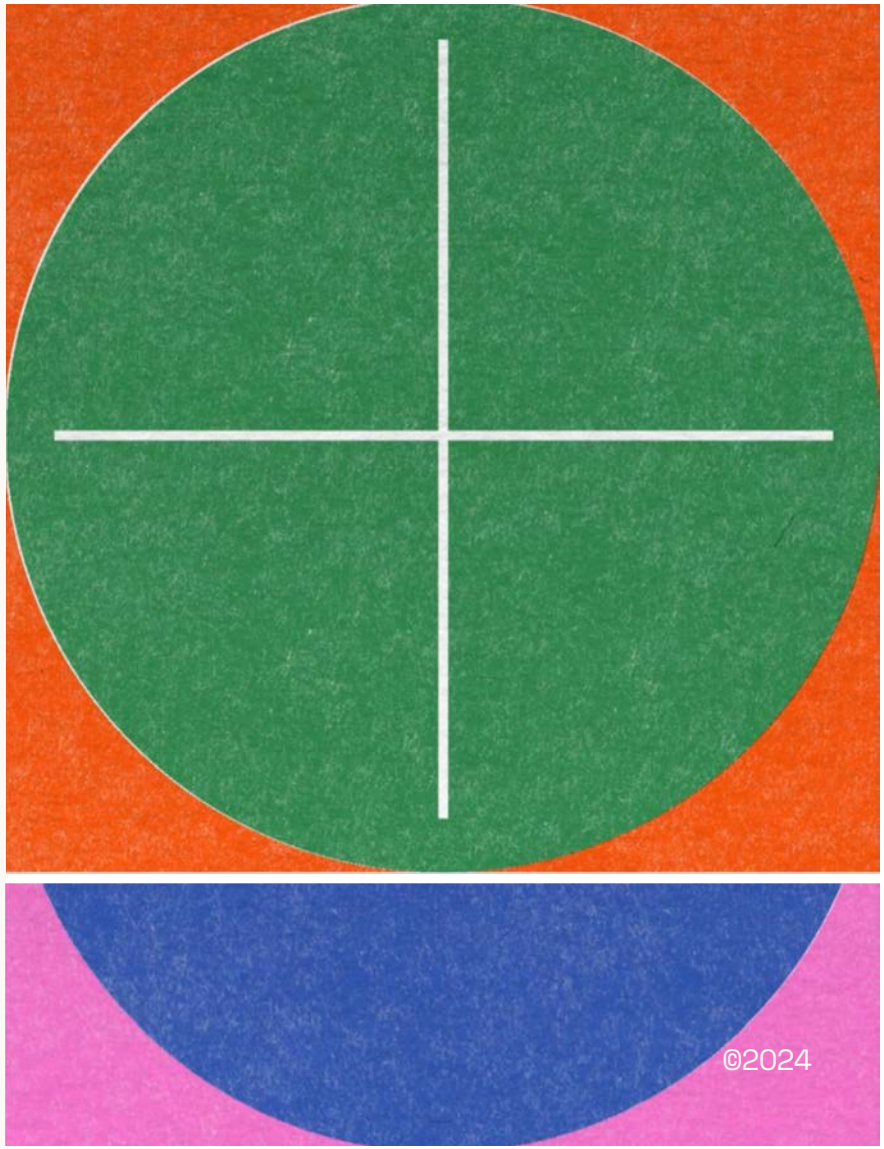


What Makes a Good Eval?

Aligns with the **Business** Use Case

Specific and **Measurable**

Example: For our AppMater support bot, evals could measure **accuracy**, **politeness**, and **response time**, and expected output **format**.



Types of Evals

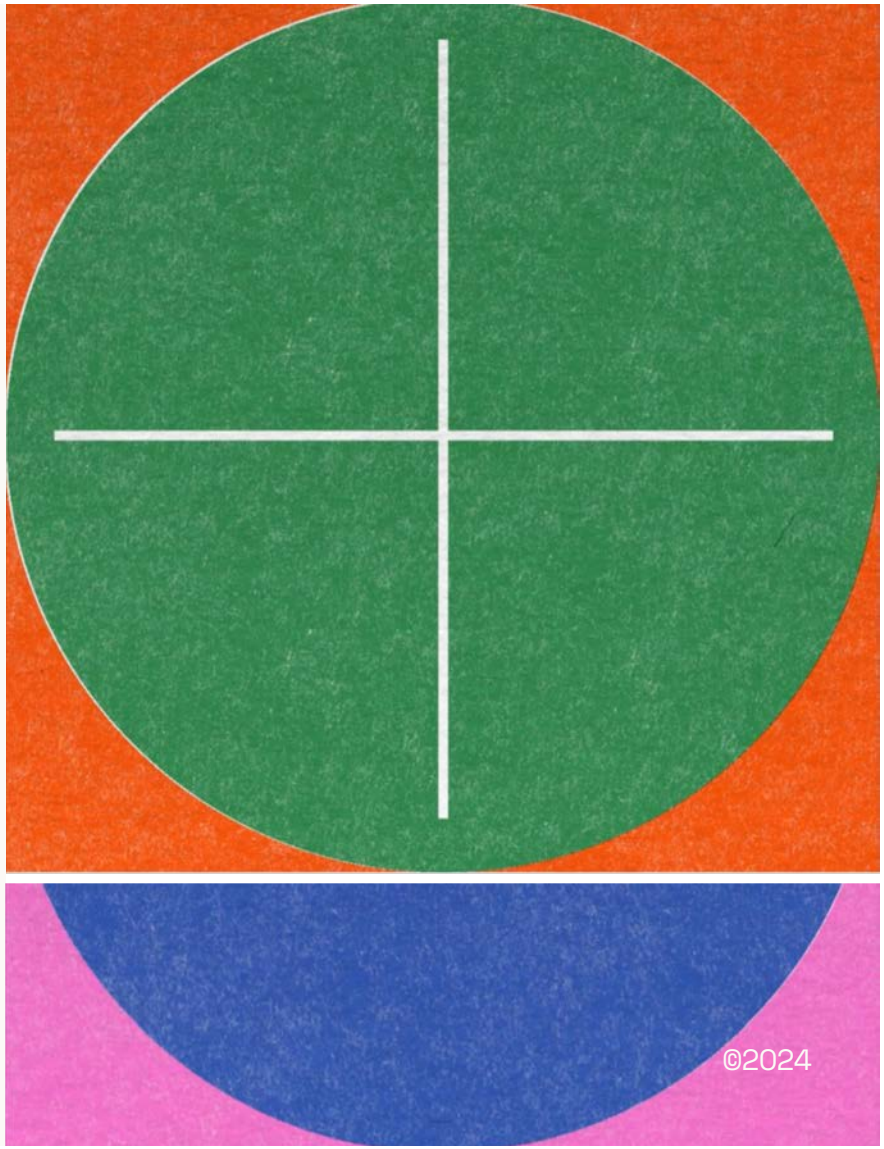
Content
Match expected outcomes, Similarity, Sensitive data etc...

Format
Text structure, JSON schema, tabular data etc...

Qualities
Grounding, Correctness, Detail etc...

Guardrails
Non-functional aspects like Diversity, Politeness etc...

Arato[AI]



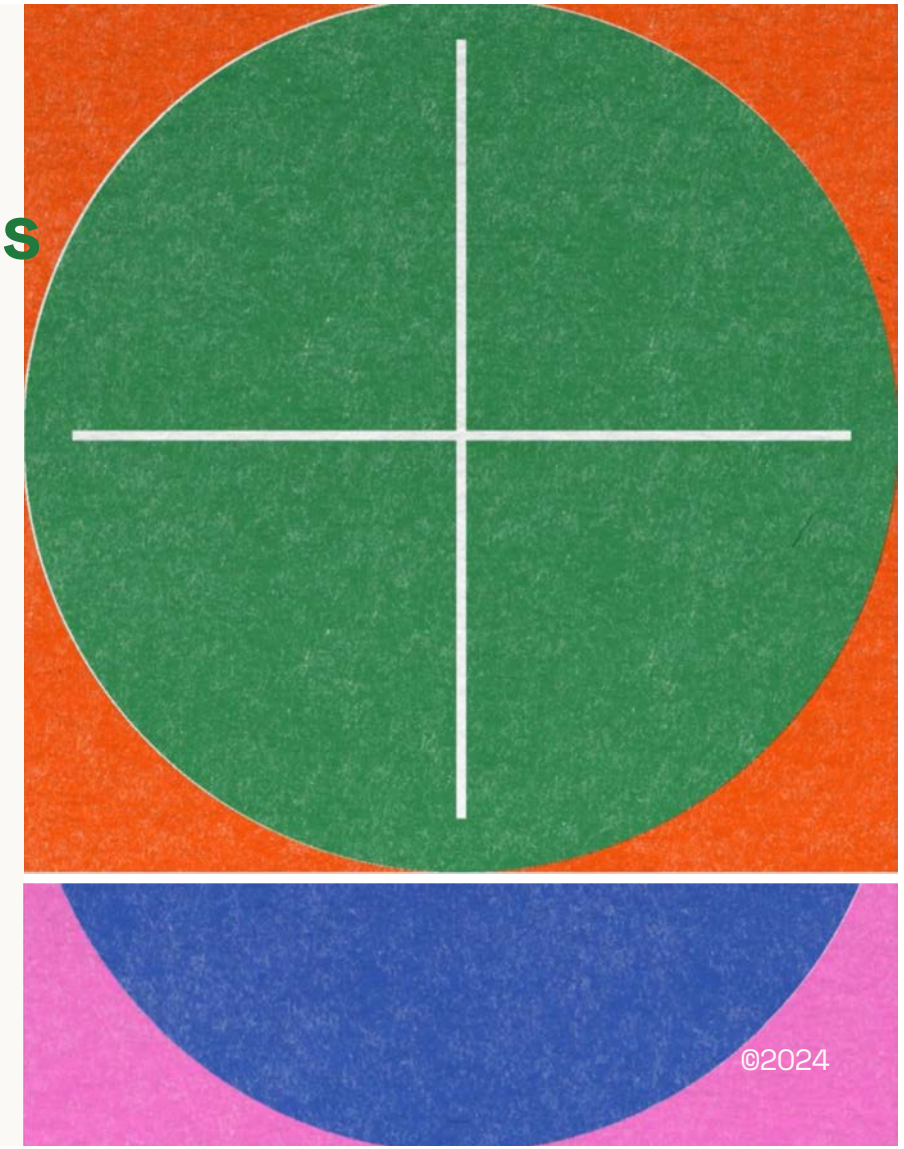
Evals: Different Mechanisms

Deterministic

Rule based, Code, Text matching,
Regex etc...

Model Based

LLM as an evaluator
Vector Similarity



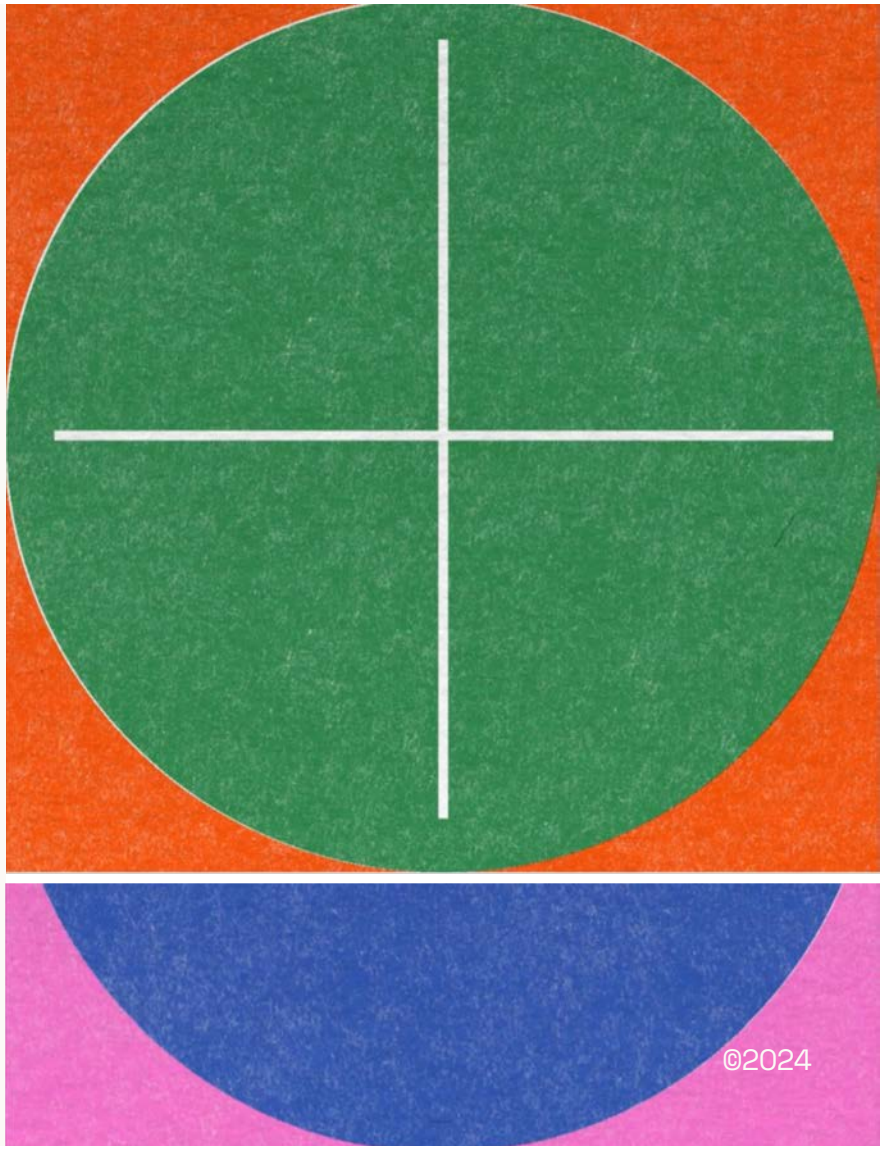
Pros and Cons

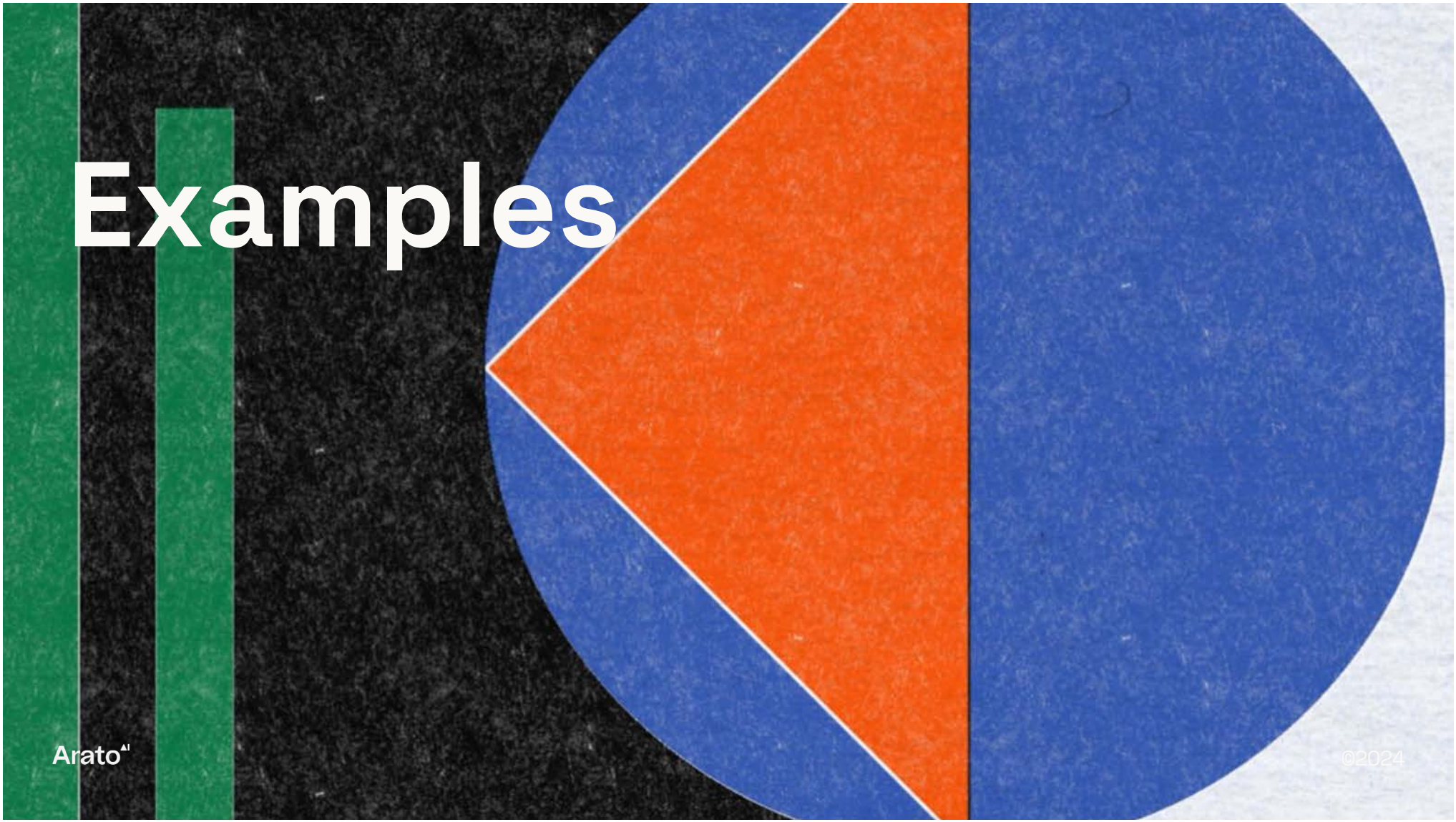
Deterministic

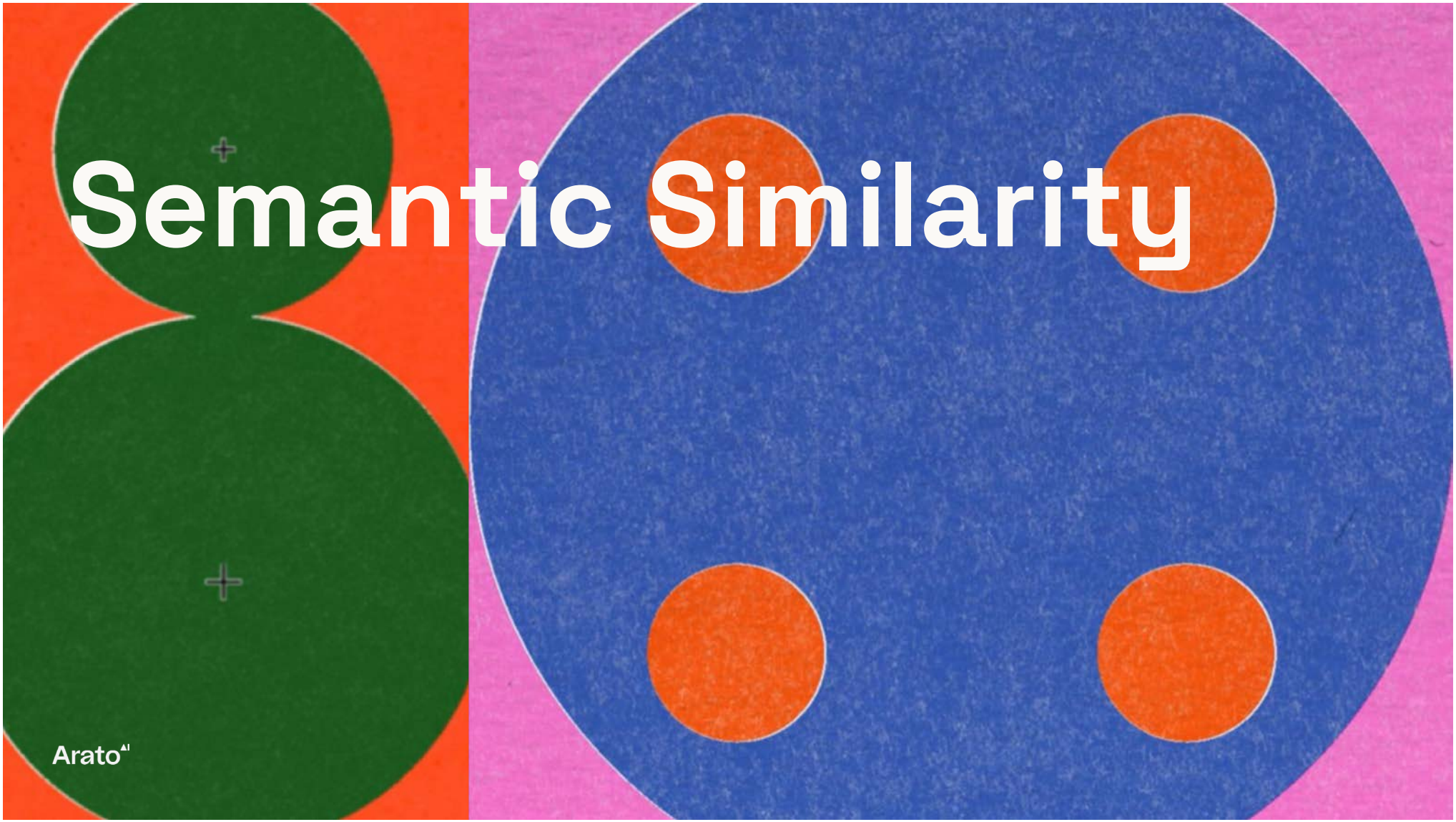
- Cheap and Accurate
- Rigid and Limited

Model Based

- Flexible and Easy to use
- Expensive and Non Deterministic




Examples
Arato
©2024


Semantic Similarity
Arato

A Sample Expected Response

Expected output


```
{
  "answer": "Hi Tim, I'm your friendly AppMaster support bot. To find your
project files in AppMaster, you should first check if they were saved in the
default directory or a custom location you specified during setup.
Additionally, you can use the search functionality in the AppMaster interface
to locate missing files by their names or tags. If you suspect that the files
might have been deleted or moved, verify their existence in your local
storage.",
  "analysis": "I used document 1 and document 3 to provide a comprehensive
answer. Document 1 helped identify where to look for the files, while
document 3 offered a method to search for them within the AppMaster
interface."
}
```


Copy Close

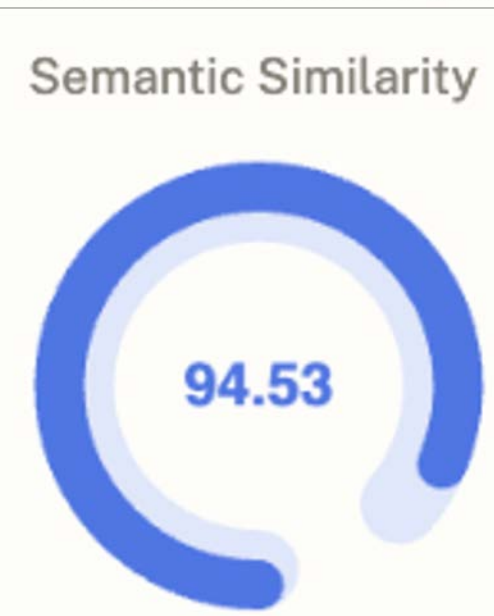
Similarity Eval

Expected output


```
{
  "answer": "Hi Tim, I'm your friendly AppMaster support bot. To find your project files in AppMaster, you
should first check if they were saved in the default directory                                    g setup.
Additionally, you can use the search functionality in the AppMa                                  y their
names or tags. If you suspect that the files might have been de                                  n your
local storage.",
  "analysis": "I used document 1 and document 3 to provide a co                                  entify
where to look for the files, while document 3 offered a method
interface."
}
```

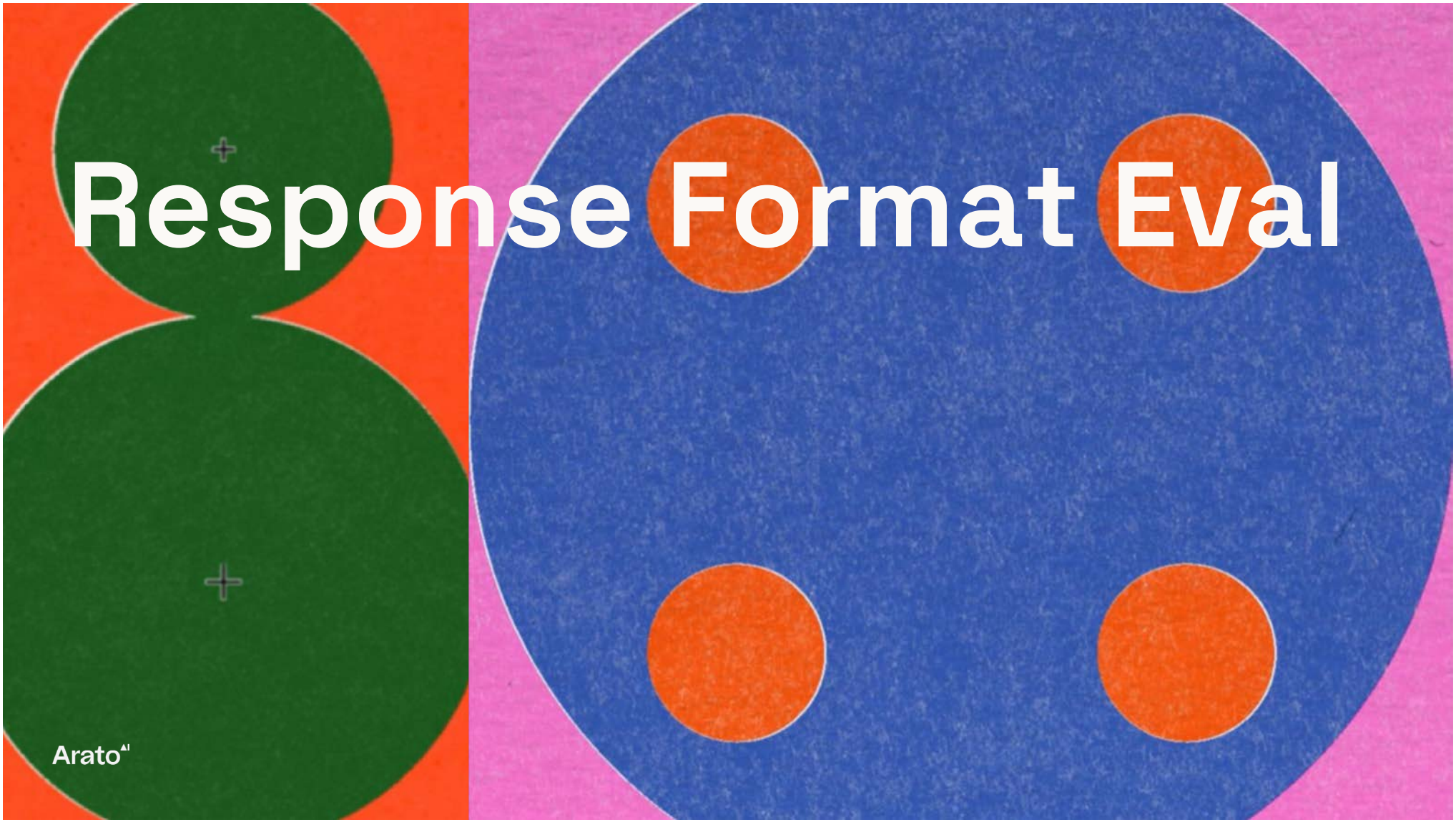

Results


```
{
  "answer": "To find your project files in AppMaster, first che                                  ocation
you specified during setup. If you still can't find them, verif                                  rage and
weren't accidentally deleted or moved. You can also use the sea                                  face to
locate missing files by their names or tags.",
  "analysis": "I used document 1, document 2, and document 3 to                                  t 1
helped identify where to look for the files, document 2 address                                  d or
moved, and document 3 suggested using the search functionality to locate the files."
}
```

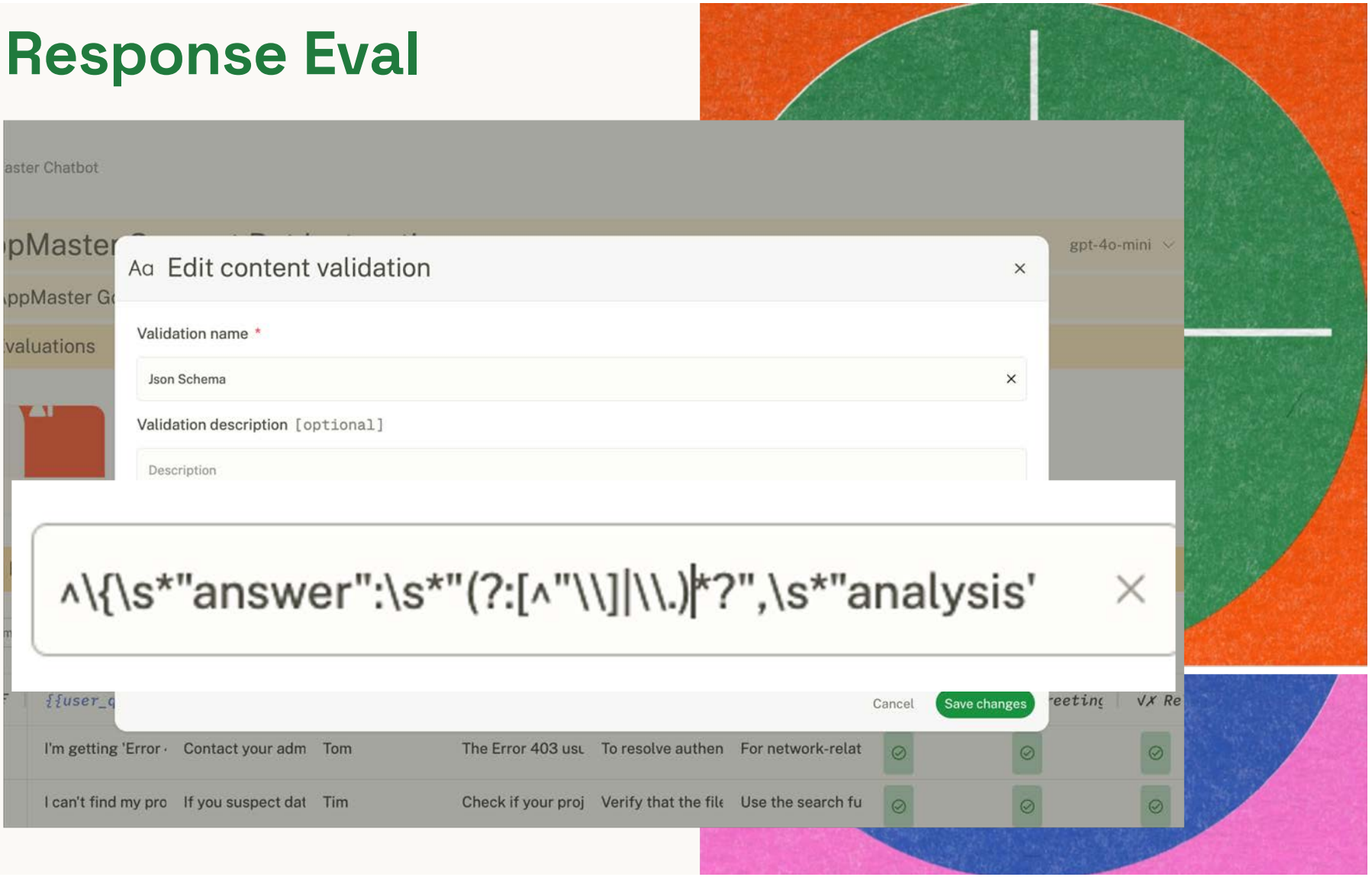



Copy Close

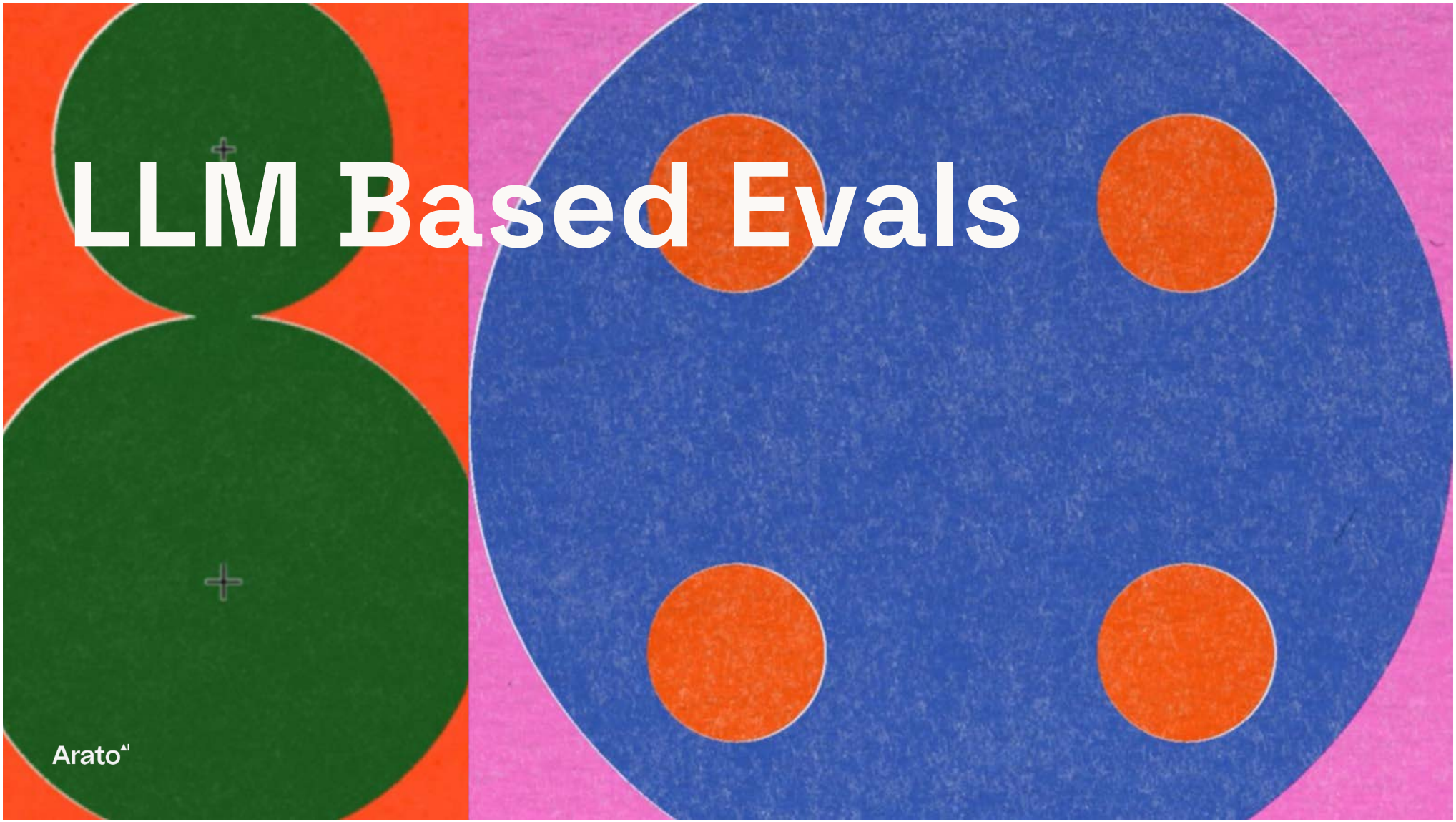
Arato[AI]


Response Format Eval
Arato AI

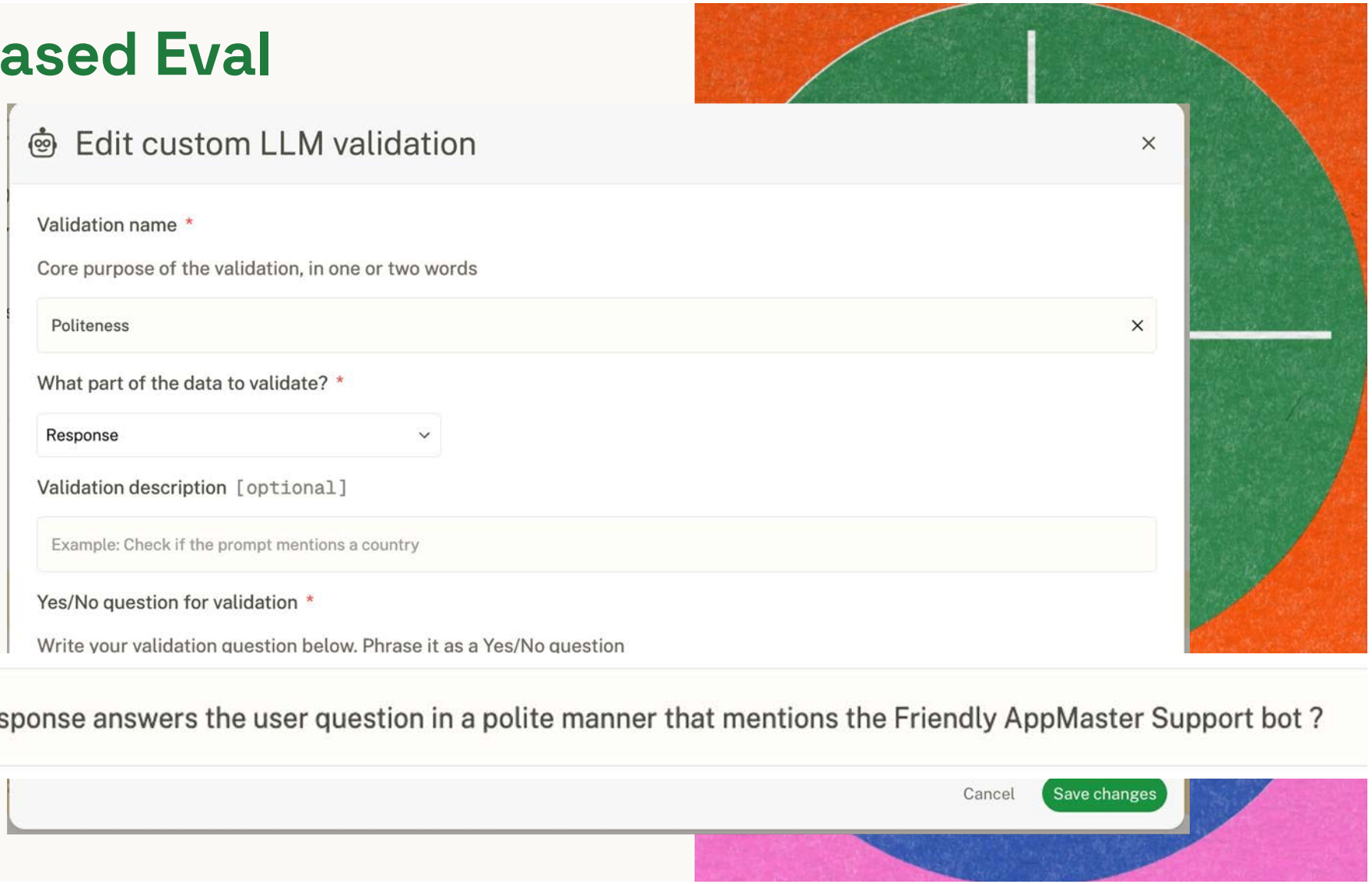
JSON Response Eval



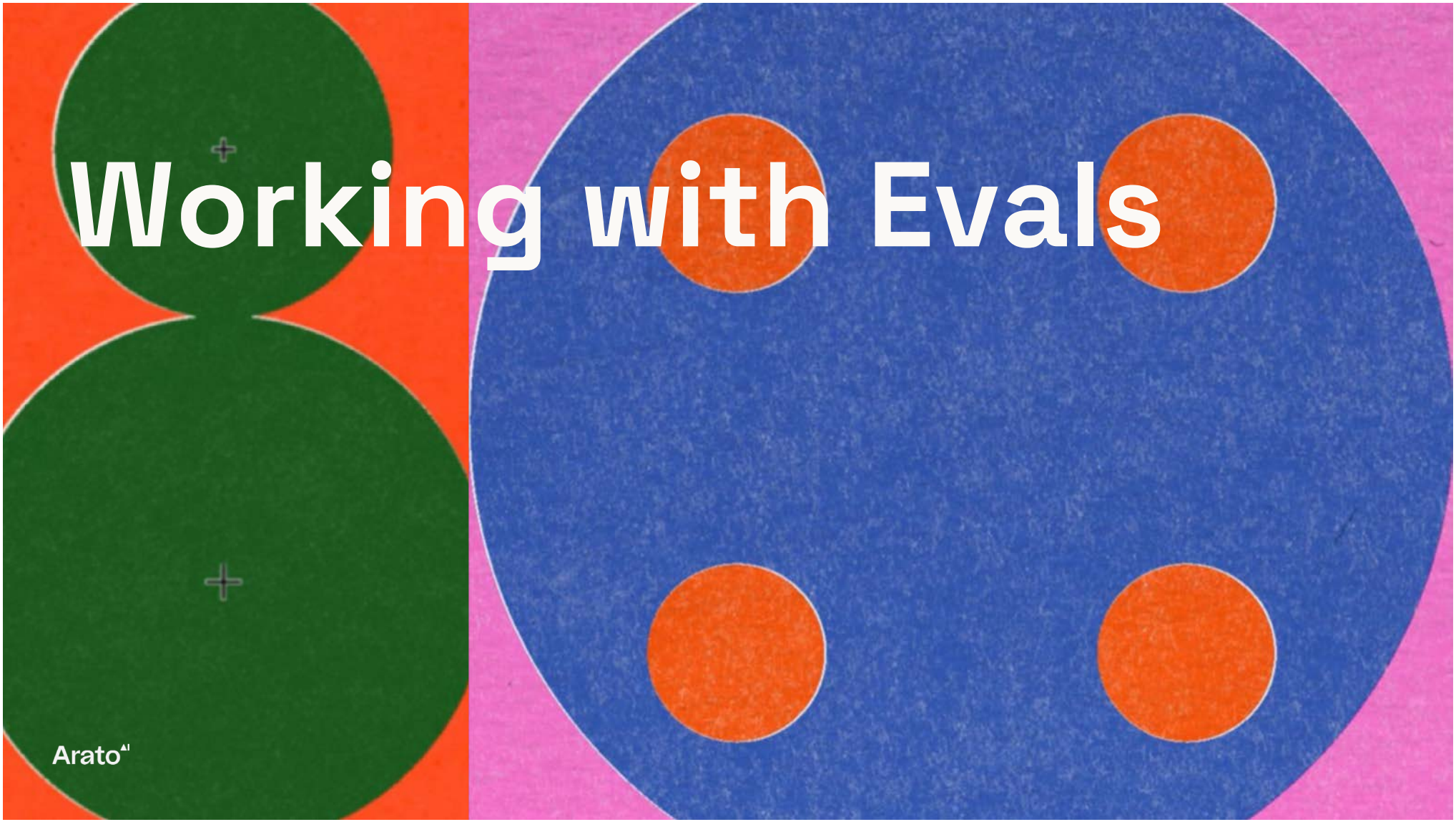
Arato


LLM Based Evals
Arato AI

LLM Based Eval



Arato

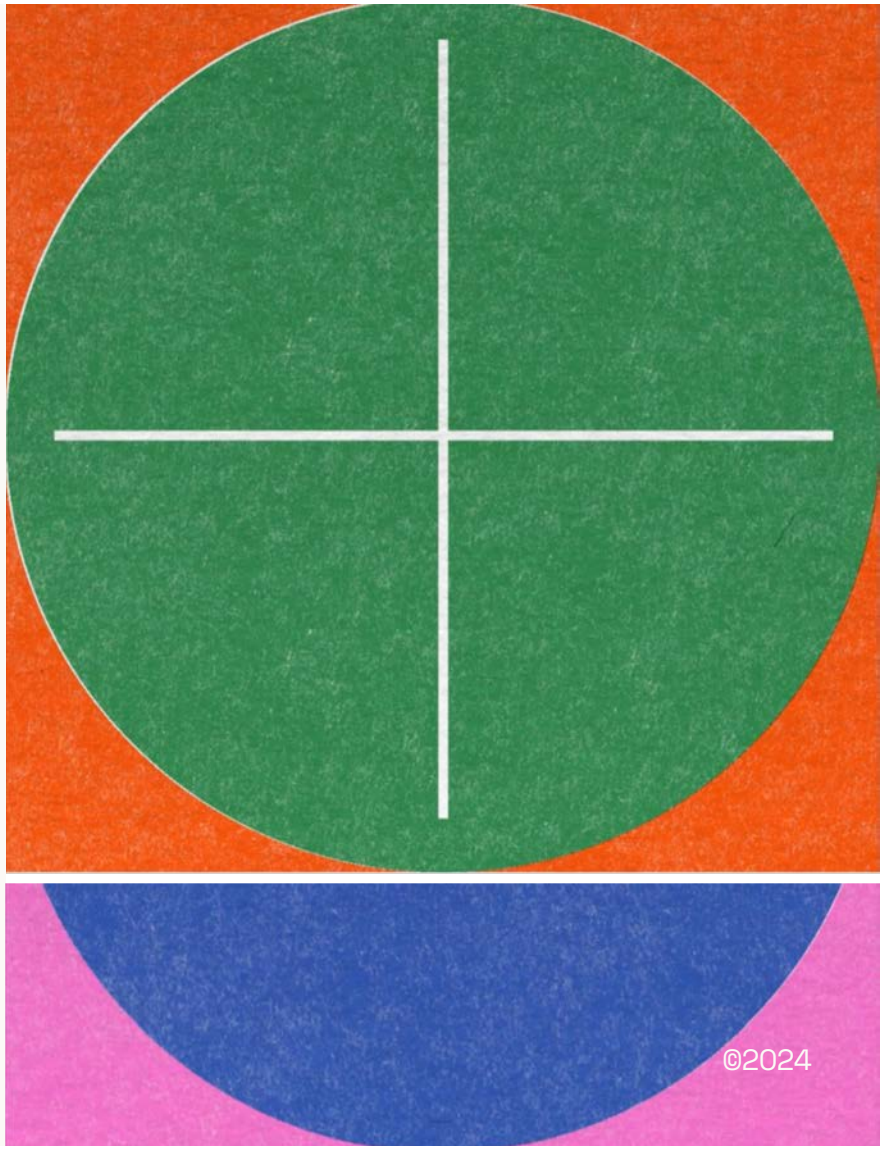

Working with Evals
Arato AI

Experimenting

Prompt Template

Data

Eval Question / Metric



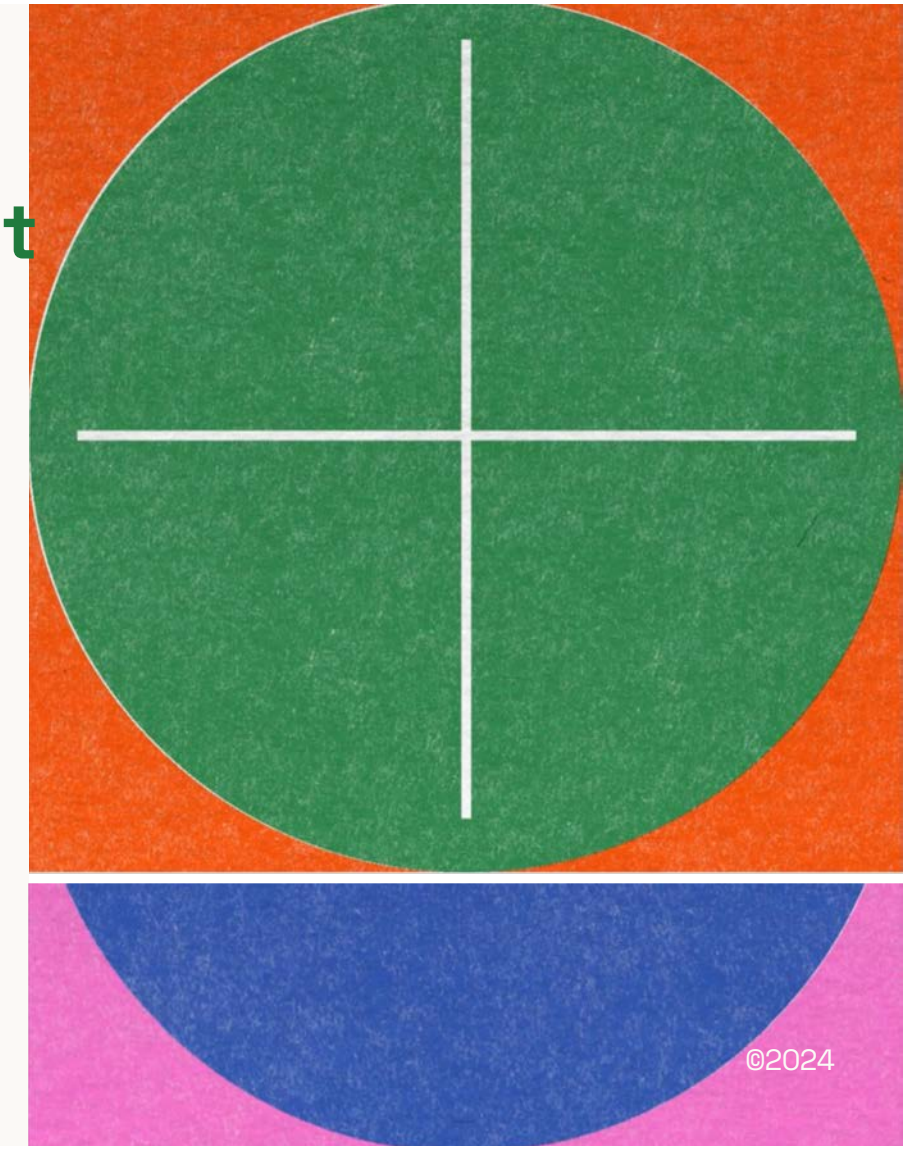
Arato AI

What Makes a Good Dataset

Diversity.

Representativeness of **real-world** inputs.

Clear **expected** outputs.




Putting it all together
Arato AI

Thank You!

hilik@arato.ai

linkedin.com/company/arato-ai

www.arato.ai

Arato