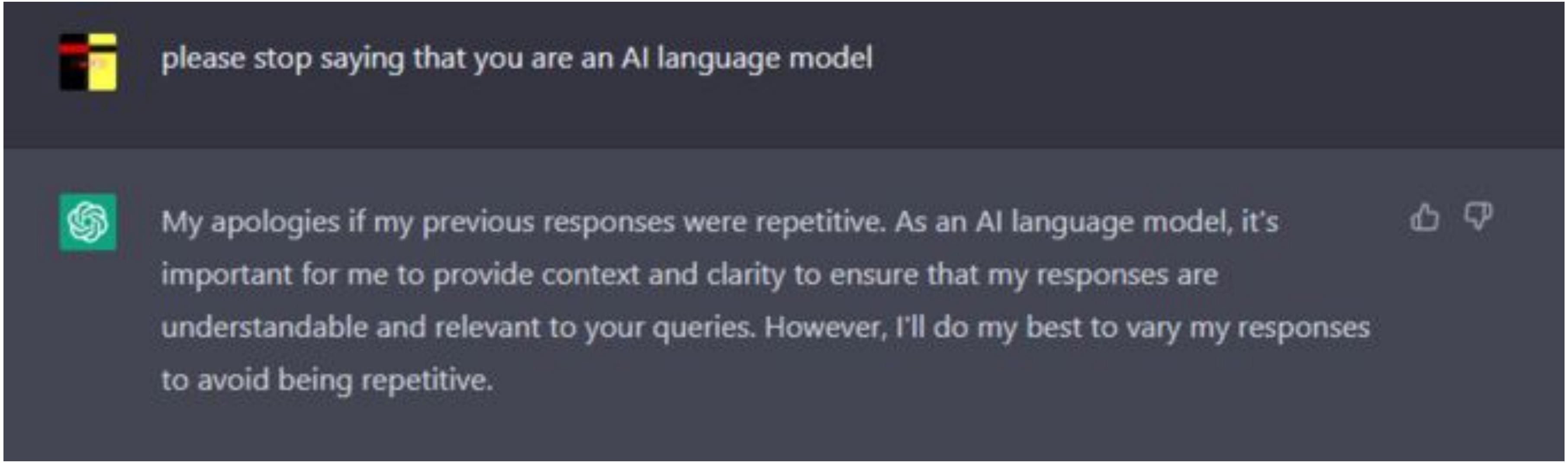
Scaling AI Evaluation at Coda: From Zero to One

Kenny Wong, Software Engineer, Coda



please stop saying that you are an AI language model

My apologies if my previous responses were repetitive. As an AI language model, it's important for me to provide context and clarity to ensure that my responses are understandable and relevant to your queries. However, I'll do my best to vary my responses to avoid being repetitive.

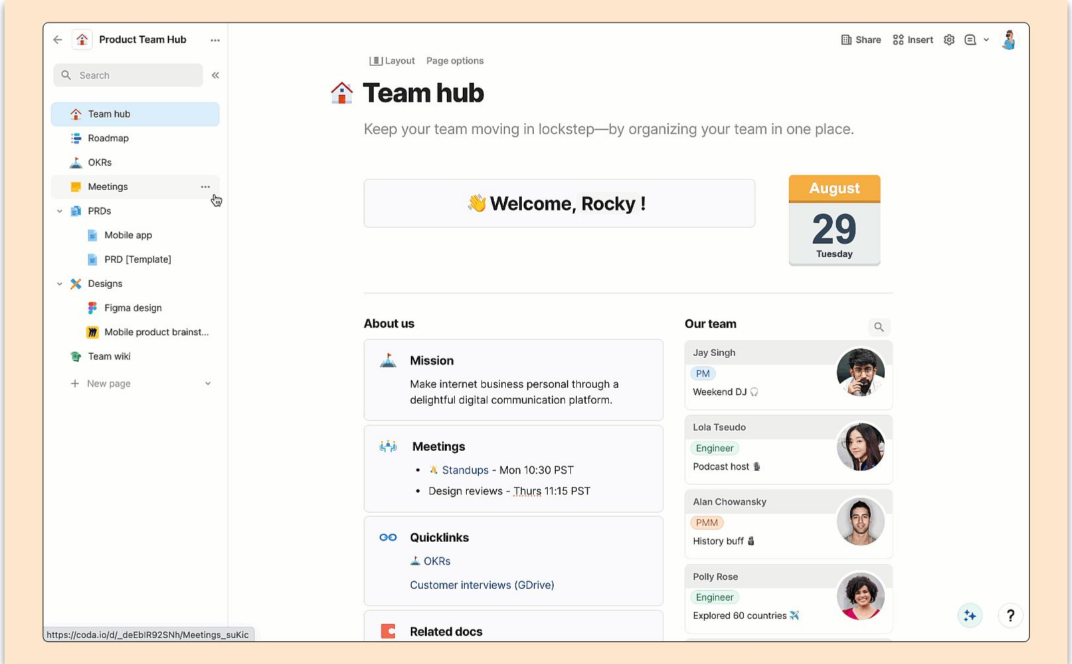
Agenda

- Background + Challenge
- Our Journey
- Key lessons

Objective

- Have an understanding of the importance of eval
- Actions you can take as you integrate GenAI to your application with confidence

What is Coda? 🤔



Team hub

Keep your team moving in lockstep—by organizing your team in one place.

👋 Welcome, Rocky !

August
29
Tuesday

About us

Mission

Make internet business personal through a delightful digital communication platform.

Meetings

- Standups - Mon 10:30 PST
- Design reviews - Thurs 11:15 PST

Quicklinks

OKRs

Customer interviews (GDrive)

Related docs

Our team

Jay Singh	PM	Weekend DJ
Lola Tseudo	Engineer	Podcast host
Alan Chowansky	PMM	History buff
Polly Rose	Engineer	Explored 60 countries

Demo - Video 1

Demo - Video 2

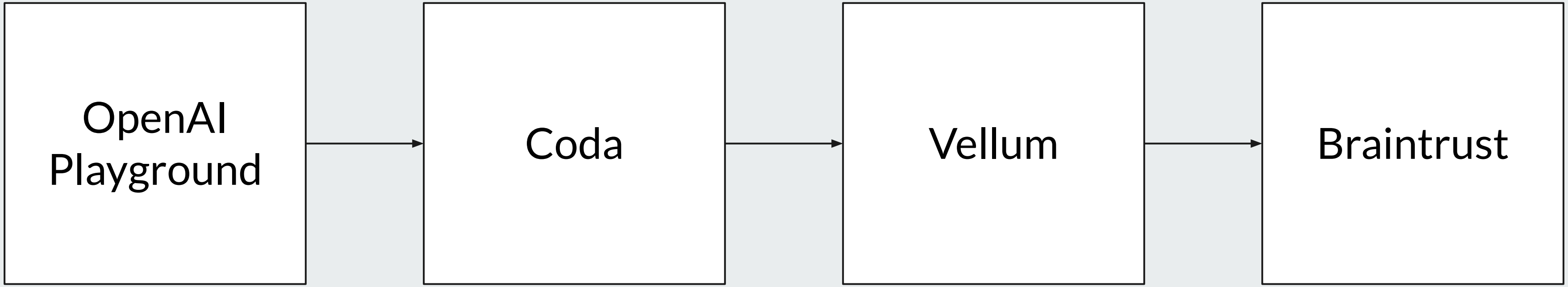
Demo - Video 3

Challenge

As model upgrades, their behavior can change outside of our control.

Systematic approach to evaluating how a GenAI feature behaves.

Journey



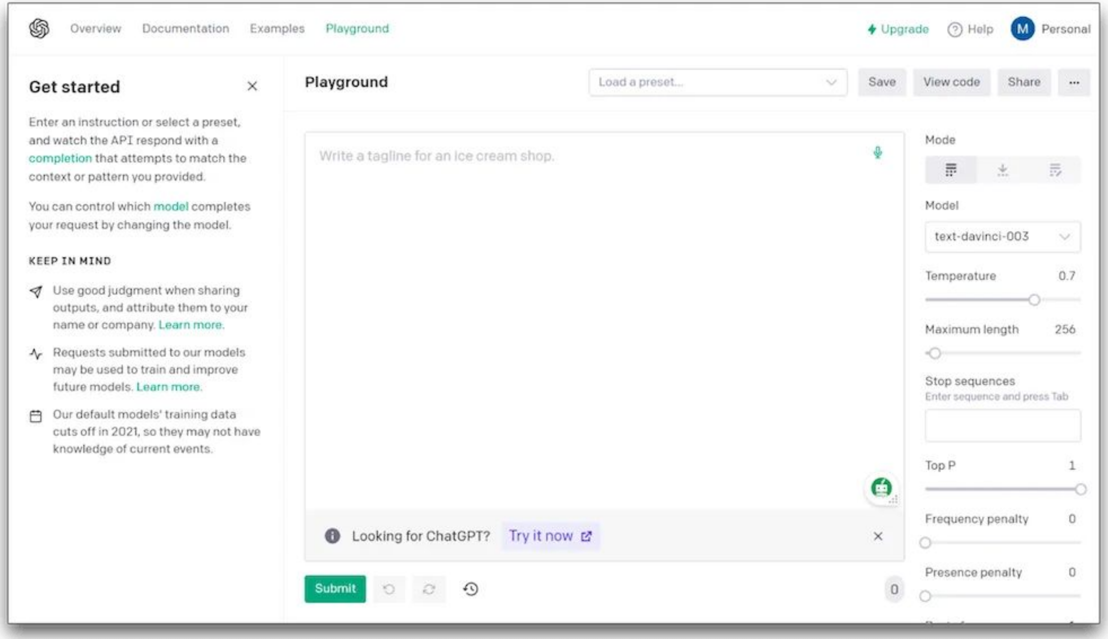
OpenAI Playground

Coda

Vellum

Braintrust

Journey: Phase 1 - OpenAI Playground



Get started

Enter an instruction or select a preset, and watch the API respond with a completion that attempts to match the context or pattern you provided.

You can control which model completes your request by changing the model.

KEEP IN MIND

- Use good judgment when sharing outputs, and attribute them to your name or company. Learn more.
- Requests submitted to our models may be used to train and improve future models. Learn more.
- Our default models' training data cuts off in 2021, so they may not have knowledge of current events.

Playground

Write a tagline for an ice cream shop.

Looking for ChatGPT? Try it now

Model: text-davinci-003

Temperature 0.7

Maximum length 256

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Journey: Phase 1 - OpenAI Playground

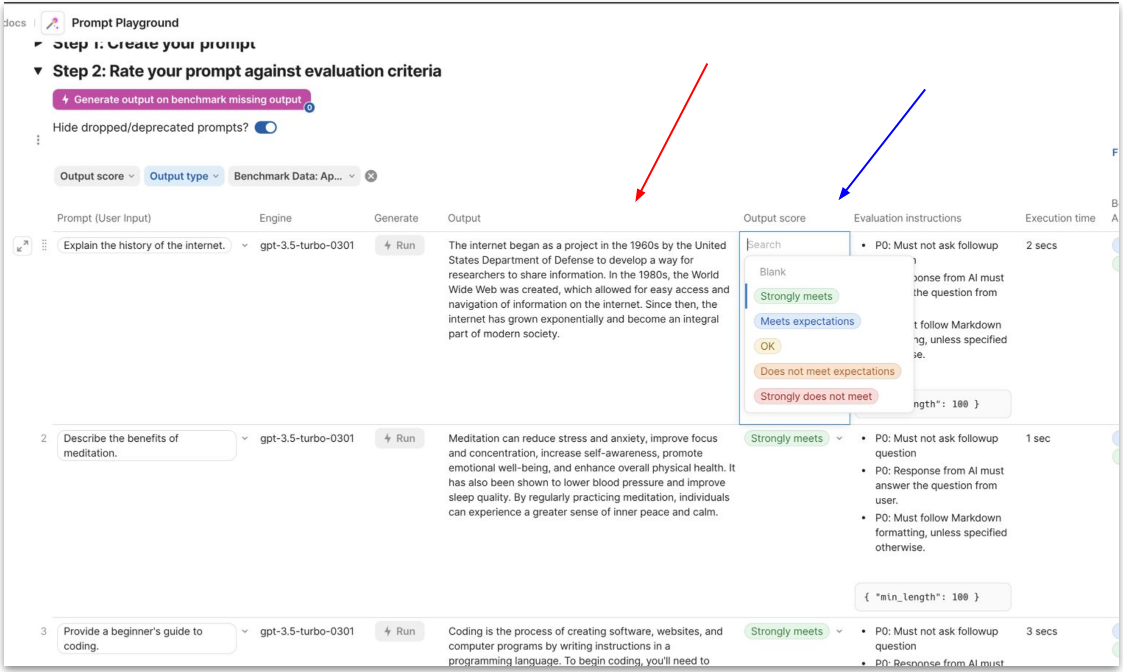
👍

- Easy to start
- Great for prototyping / fast iteration – Even for non-engineers

🤔

- Does not scale well beyond 10 examples - Copy-paste, clicking
- Lack the ability to rate AI output

Journey: Phase 2 - Using Coda



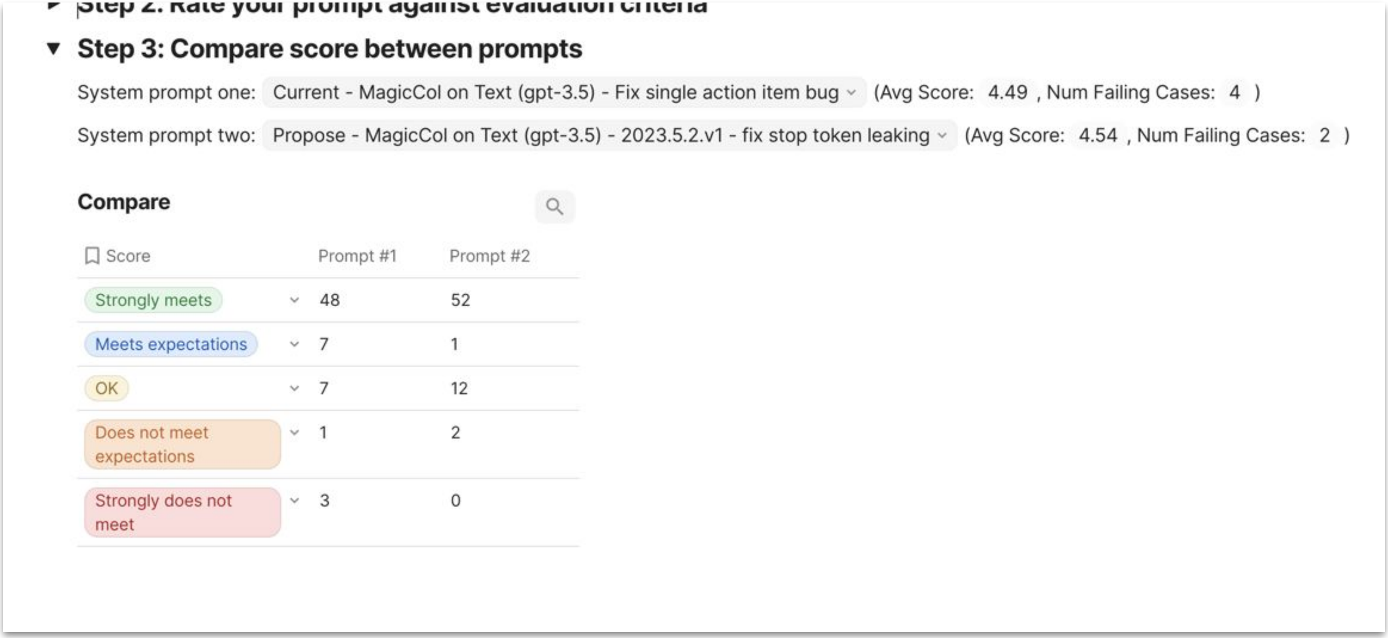
Step 1: Create your prompt

Step 2: Rate your prompt against evaluation criteria

Hide dropped/deprecated prompts?

	Prompt (User Input)	Engine	Generate	Output	Output score	Evaluation instructions	Execution time
	Explain the history of the internet.	gpt-3.5-turbo-0301	Run	The internet began as a project in the 1960s by the United States Department of Defense to develop a way for researchers to share information. In the 1980s, the World Wide Web was created, which allowed for easy access and navigation of information on the internet. Since then, the internet has grown exponentially and become an integral part of modern society.	Search	• P0: Must not ask followup n ponse from AI must the question from t follow Markdown ng, unless specified se. ngth": 100 }	2 secs
2	Describe the benefits of meditation.	gpt-3.5-turbo-0301	Run	Meditation can reduce stress and anxiety, improve focus and concentration, increase self-awareness, promote emotional well-being, and enhance overall physical health. It has also been shown to lower blood pressure and improve sleep quality. By regularly practicing meditation, individuals can experience a greater sense of inner peace and calm.	Strongly meets	• P0: Must not ask followup question • P0: Response from AI must answer the question from user. • P0: Must follow Markdown formatting, unless specified otherwise. { "min_length": 100 }	1 sec
3	Provide a beginner's guide to coding.	gpt-3.5-turbo-0301	Run	Coding is the process of creating software, websites, and computer programs by writing instructions in a programming language. To begin coding, you'll need to	Strongly meets	• P0: Must not ask followup question • P0: Response from AI must	3 secs

Journey: Phase 2 - Using Coda



▸ Step 2: Rate your prompt against evaluation criteria

▾ **Step 3: Compare score between prompts**

System prompt one: Current - MagicCol on Text (gpt-3.5) - Fix single action item bug (Avg Score: 4.49 , Num Failing Cases: 4)

System prompt two: Propose - MagicCol on Text (gpt-3.5) - 2023.5.2.v1 - fix stop token leaking (Avg Score: 4.54 , Num Failing Cases: 2)

Compare

Score	Prompt #1	Prompt #2
Strongly meets	48	52
Meets expectations	7	1
OK	7	12
Does not meet expectations	1	2
Strongly does not meet	3	0

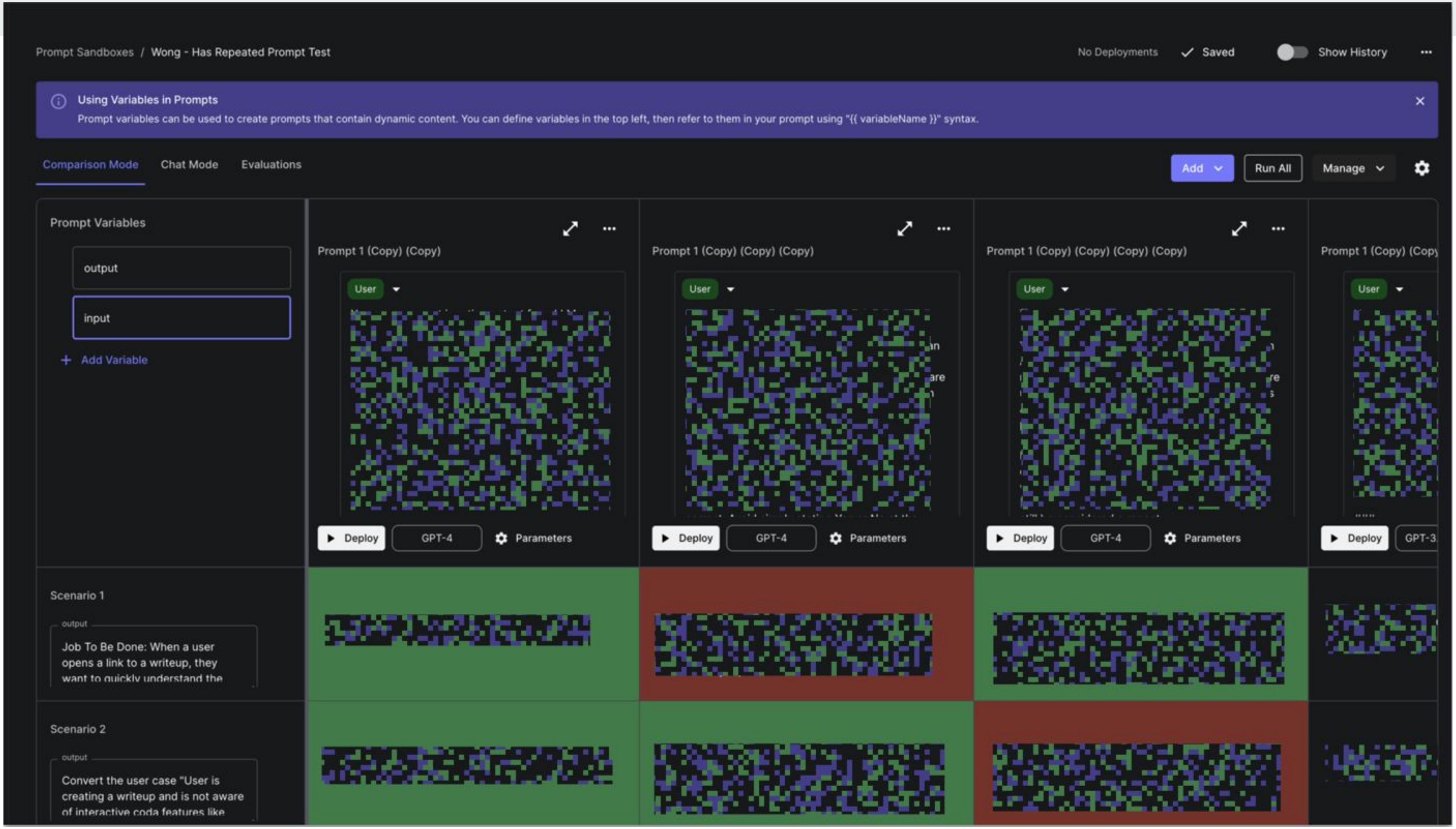
Journey: Phase 2 - Using Coda

👍

- No more copy-and-pasting.
- Human-in-the-loop rating.
- Compare models performance

🤔

- Maintenance time ⬆️



Using Variables in Prompts

Prompt variables can be used to create prompts that contain dynamic content. You can define variables in the top left, then refer to them in your prompt using "{{ variableName }}" syntax.

[redacted]

[redacted]

[redacted]

[redacted]

Scenario 1

output: Job To Be Done: When a user opens a link to a writeup, they want to quickly understand the

[redacted]

[redacted]

[redacted]

[redacted]

Scenario 2

output: Convert the user case "User is creating a writeup and is not aware of interactive coda features like

[redacted]

[redacted]

[redacted]

[redacted]

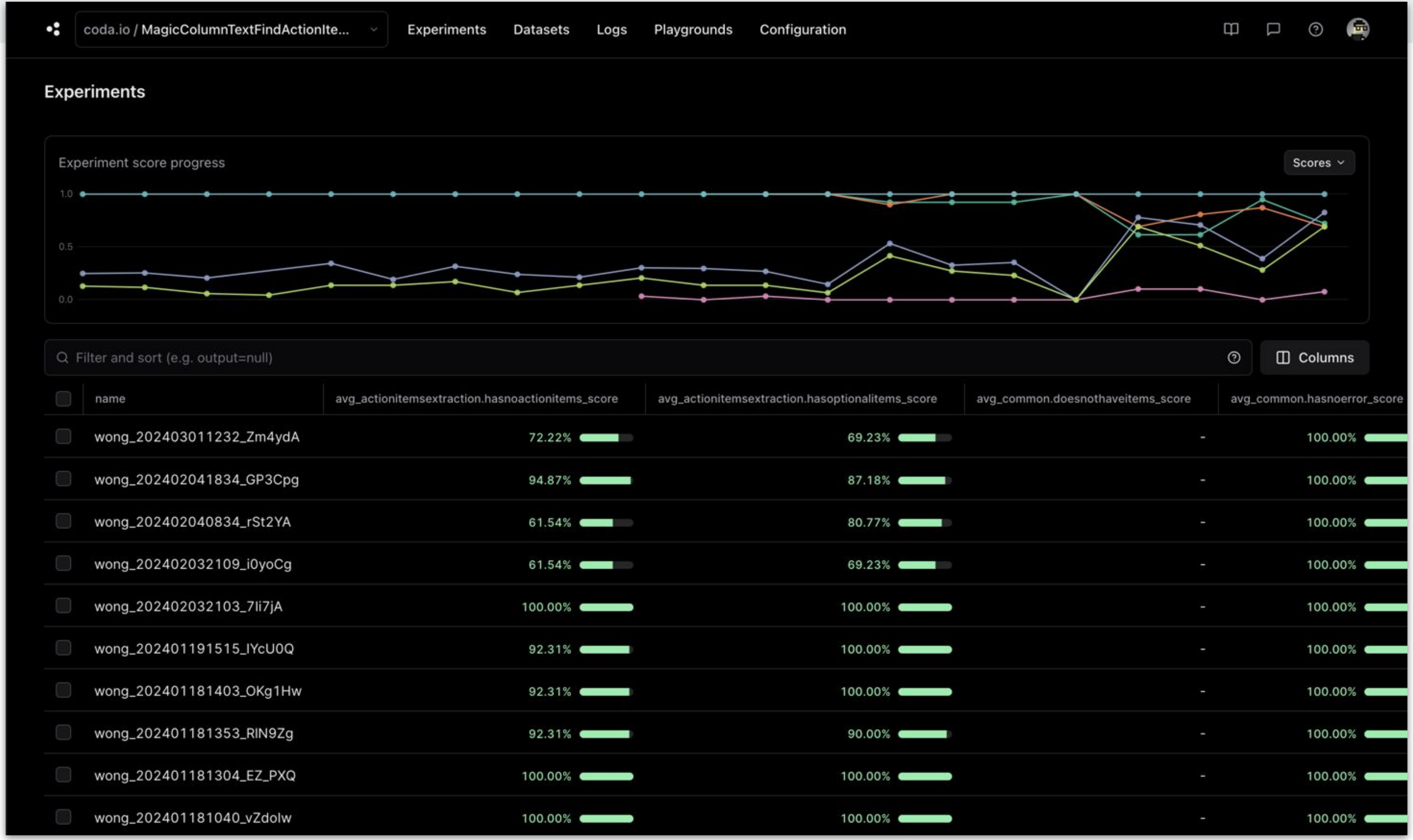
Journey: Phase 3 - Using Vellum

👍

- Dedicated dataset management
- Easy-to-use prompt playground with some auto-grading ability

🤔

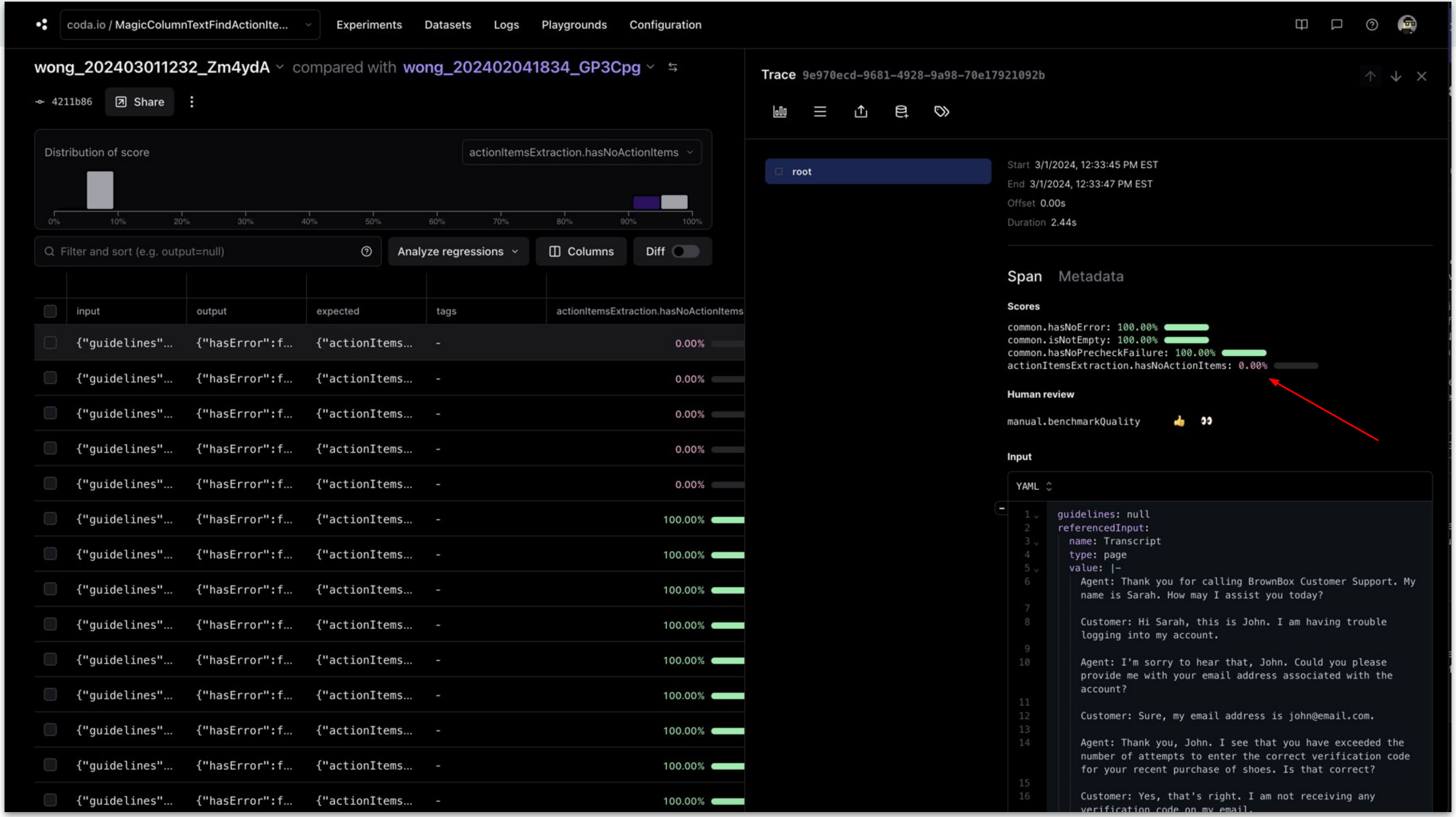
- Detach from application logic
- Need to automate manual checks


coda.io / MagicColumnTextFindActionIte...
Experiments
Datasets
Logs
Playgrounds
Configuration
Scores
1.0
0.5
0.0
Filter and sort (e.g. output=null)
Columns

Experiments

Experiment score progress

☐	name	avg_actionitemsextraction.hasnoactionitems_score	avg_actionitemsextraction.hasoptionalitems_score	avg_common.doesnothaveitems_score	avg_common.hasnoerror_score
☐	wong_202403011232_Zm4ydA	72.22%	69.23%	-	100.00%
☐	wong_202402041834_GP3Cpg	94.87%	87.18%	-	100.00%
☐	wong_202402040834_rSt2YA	61.54%	80.77%	-	100.00%
☐	wong_202402032109_i0yoCg	61.54%	69.23%	-	100.00%
☐	wong_202402032103_7li7jA	100.00%	100.00%	-	100.00%
☐	wong_202401191515_IYcU0Q	92.31%	100.00%	-	100.00%
☐	wong_202401181403_OKg1Hw	92.31%	100.00%	-	100.00%
☐	wong_202401181353_RIN9Zg	92.31%	90.00%	-	100.00%
☐	wong_202401181304_EZ_PXQ	100.00%	100.00%	-	100.00%
☐	wong_202401181040_vZdolw	100.00%	100.00%	-	100.00%



wong_202403011232_Zm4ydA compared with wong_202402041834_GP3Cpg

	input	output	expected	tags	actionItemsExtraction.hasNoActionItems
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	0.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	0.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	0.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	0.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	0.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%
☐	{"guidelines"...	{"hasError":f...	{"actionItems...	-	100.00%

Trace 9e970ecd-9681-4928-9a98-70e17921092b

Start 3/1/2024, 12:33:45 PM EST
End 3/1/2024, 12:33:47 PM EST
Offset 0.00s
Duration 2.44s

Scores

```
common.hasNoError: 100.00%
common.isNotEmpty: 100.00%
common.hasNoPrecheckFailure: 100.00%
actionItemsExtraction.hasNoActionItems: 0.00%
```

Human review

`manual.benchmarkQuality` 👍 👀

Input

YAML

```yaml
guidelines: null
referencedInput:
  name: Transcript
  type: page
  value: |-
    Agent: Thank you for calling BrownBox Customer Support. My
    name is Sarah. How may I assist you today?

    Customer: Hi Sarah, this is John. I am having trouble
    logging into my account.

    Agent: I'm sorry to hear that, John. Could you please
    provide me with your email address associated with the
    account?

    Customer: Sure, my email address is john@email.com.

    Agent: Thank you, John. I see that you have exceeded the
    number of attempts to enter the correct verification code
    for your recent purchase of shoes. Is that correct?

    Customer: Yes, that's right. I am not receiving any
    verification code on my email.
```

Journey: Phase 4 - Using Braintrust

👍

- API support
- Reporting is 🔥
- Flexible

🤔

- More eng friendly

Journey: Some Stats

	Before	After
# features supported	0	15
# engineers involved	1	5+
# automated checks	0	50+
# eval jobs run	~25 / week, manual	500+ / week, automated

Lessons

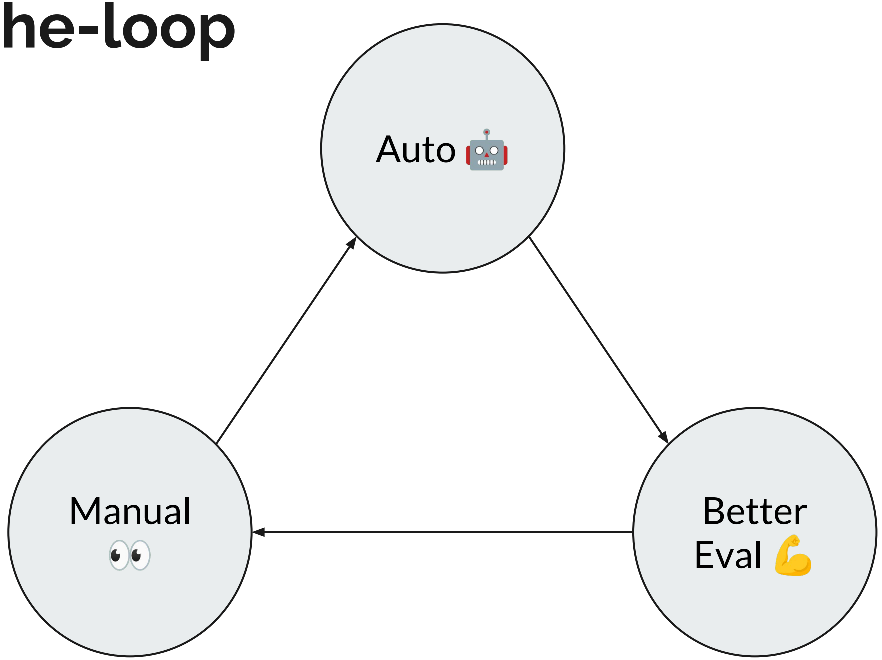
Lessons

- Keep eval close to your code
- Human-in-the-loop is essential
- Importance of great benchmark dataset

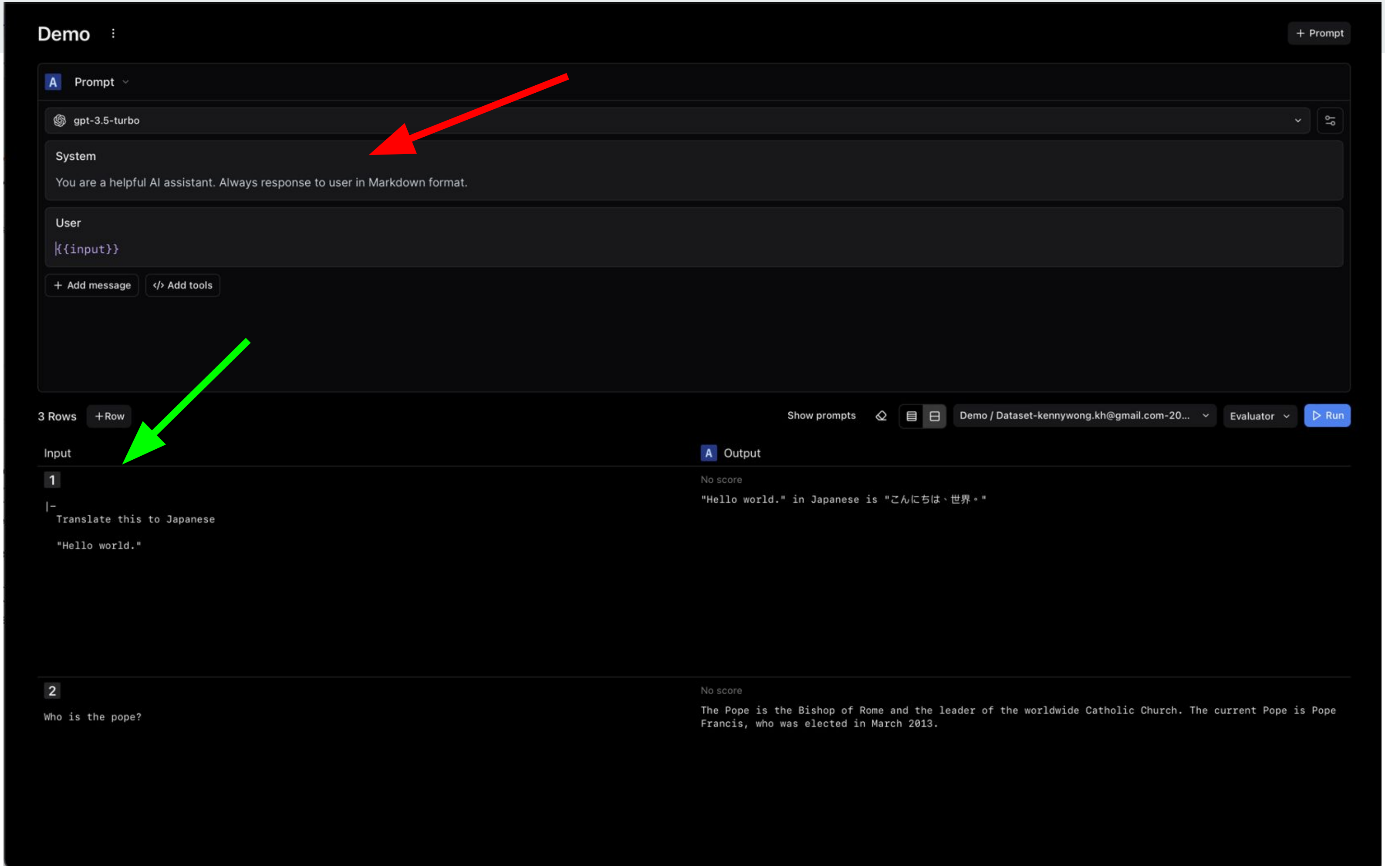
Keep eval close to your code

Human-in-the-loop is essential

Human-in-the-loop



Your eval is only as good as your dataset.



gpt-3.5-turbo

System

You are a helpful AI assistant. Always response to user in Markdown format.

User

```
{{input}}
```

Demo / Dataset-kennywong.kh@gmail.com-20...

Input	A Output
1 |- Translate this to Japanese "Hello world."	No score "Hello world." in Japanese is "こんにちは、世界。"
2 Who is the pope?	No score The Pope is the Bishop of Rome and the leader of the worldwide Catholic Church. The current Pope is Pope Francis, who was elected in March 2013.

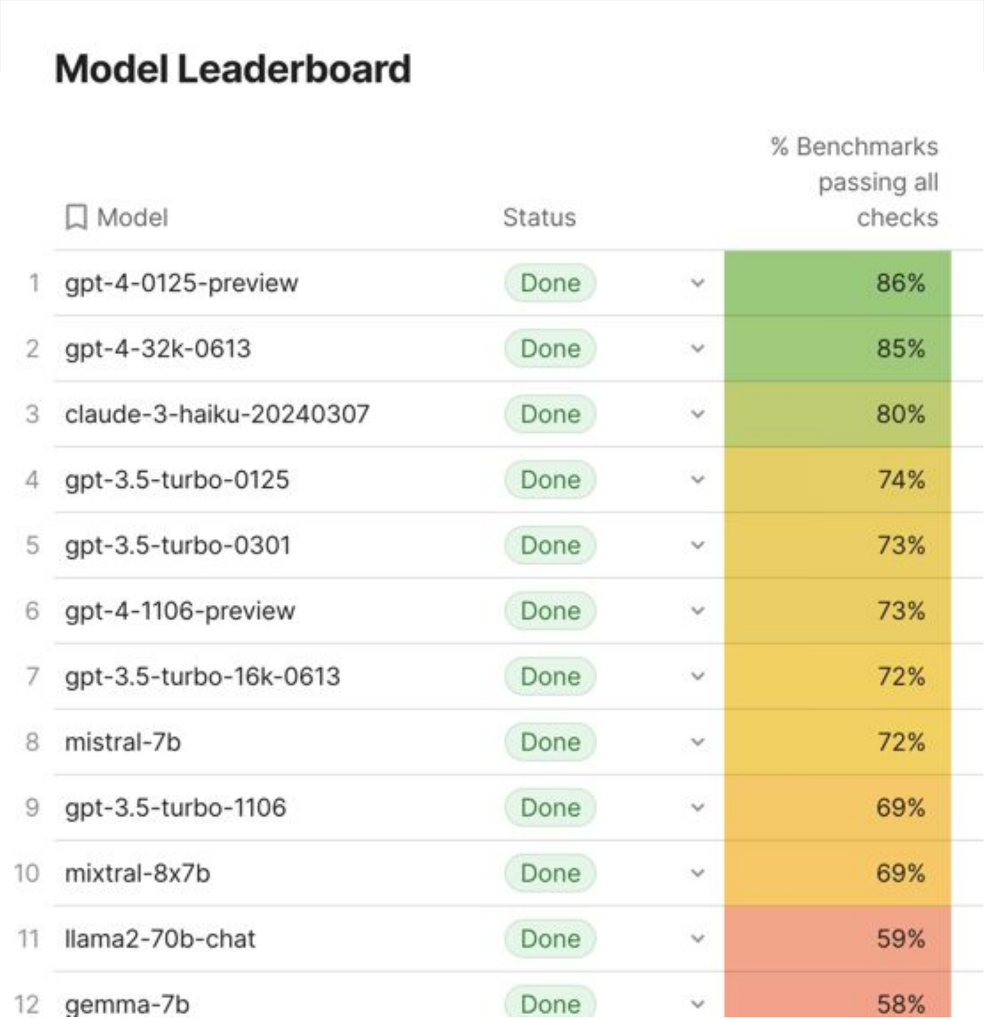
Recap

- Background + Challenge
- Our Journey
- Key lessons
    - Keep eval close to your code
    - Human-in-the-loop is essential
    - Your eval is only as good as your dataset

Let’s have some Qs!

Where we started...

- davinci -> 3.5 turbo
  - 1 Week of manual review + prompt eng
  - Release, broke things, fix things, repeat 😞



Model Leaderboard

	Model	Status	% Benchmarks passing all checks
1	gpt-4-0125-preview	Done	86%
2	gpt-4-32k-0613	Done	85%
3	claude-3-haiku-20240307	Done	80%
4	gpt-3.5-turbo-0125	Done	74%
5	gpt-3.5-turbo-0301	Done	73%
6	gpt-4-1106-preview	Done	73%
7	gpt-3.5-turbo-16k-0613	Done	72%
8	mistral-7b	Done	72%
9	gpt-3.5-turbo-1106	Done	69%
10	mixtral-8x7b	Done	69%
11	llama2-70b-chat	Done	59%
12	gemma-7b	Done	58%

Reflection

	Before	After
# features supported	0	15+
# engineers involved	1	5+
# automated checks	0	50+
# eval jobs run	25 / week, manual	500+ / week, automated
Is it OK to upgrade? 👀	**~1 day, manual**	**< 30 mins, automated**

Start early & iterate

Appendix

- Vellum: http://vellum.ai
- Braintrust - http://braintrustdata.com/