Mastering LLMs

A Conference For
Developers & Data Scientists

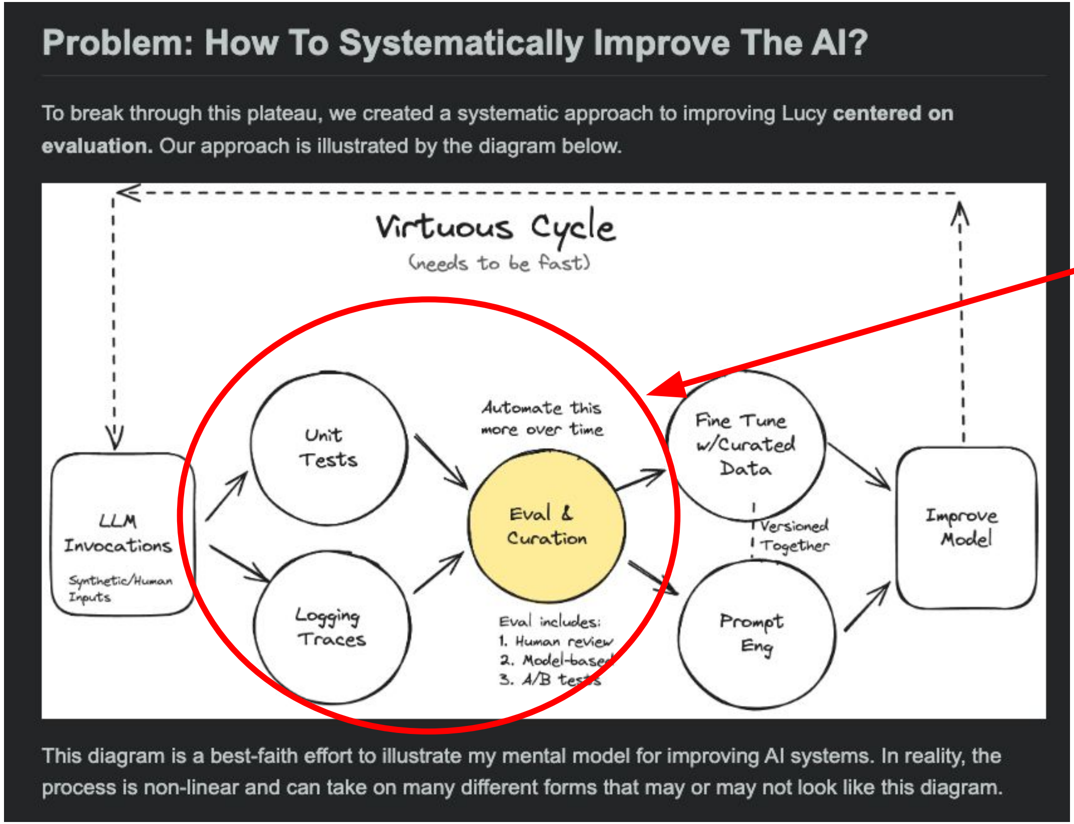
Plan For Today

- Eval types and tradeoffs
- Langsmith Deep Dive - Harrison Chase
- Hex Case Study - Bryan Bischof
- Metrics - Eugene Yan
- Evals UX & Workflows - Shreya Shankar

Reminder

Fill out the form to get your compute credits

Key: Make the dev cycle fast



Problem: How To Systematically Improve The AI?

To break through this plateau, we created a systematic approach to improving Lucy **centered on evaluation.** Our approach is illustrated by the diagram below.

This diagram is a best-faith effort to illustrate my mental model for improving AI systems. In reality, the process is non-linear and can take on many different forms that may or may not look like this diagram.

This is the "applied" part of AI.

Look at data + evals & iterate

https://hamel.dev/blog/posts/evals/

Types of Evaluations

	Writing Queries	Debiasing Text
Unit tests		
LLM as a judge		
Human Evaluation		

Editing Out Stereotypes In Academic Writing

Norway's mining economy flourished during the period ~~due to Norwegians' natural hardiness~~.

Unit Tests

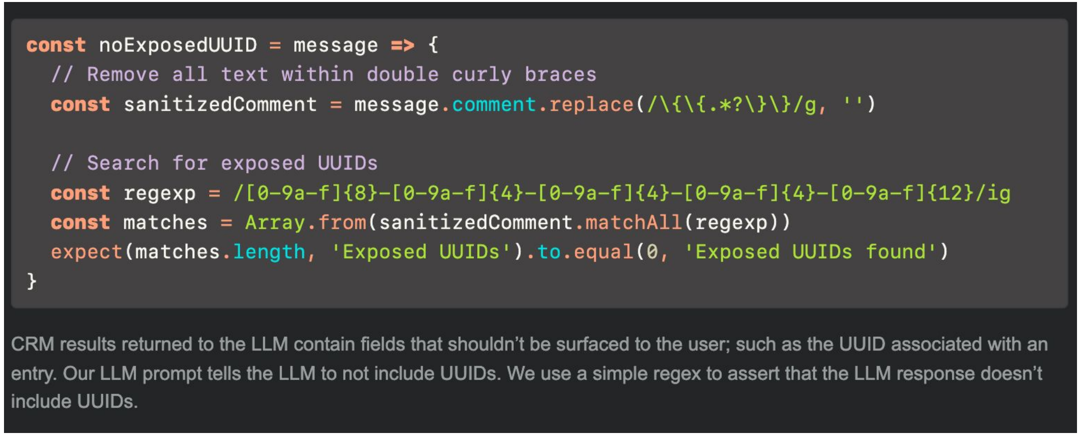
```python
from transformers import pipeline, Pipeline
import pytest

@pytest.fixture(scope="module")
def llm_pipeline():
    return pipeline("text-generation", model="meta-llama/Llama-2-7b-chat-hf", device=0)

def verify_answer_contains(p: Pipeline, query: str, expected: str):
    result = p(
        query, do_sample=False, truncation=True, return_full_text=False
    )[0]["generated_text"]
    assert expected in result, f"The result does not contain '{expected}'"

def test_google_ceo(llm_pipeline):
    verify_answer_contains(llm_pipeline, "Who is the CEO of Google?", "Sundar Pichai")

def test_2_plus_3(llm_pipeline):
    verify_answer_contains(llm_pipeline, "What is 2+3?", "5")
```

Unit Tests

If you don’t have dumb failure modes you aren’t looking at your data.

Abstract logic of unit tests so you can use it everywhere -> self healing

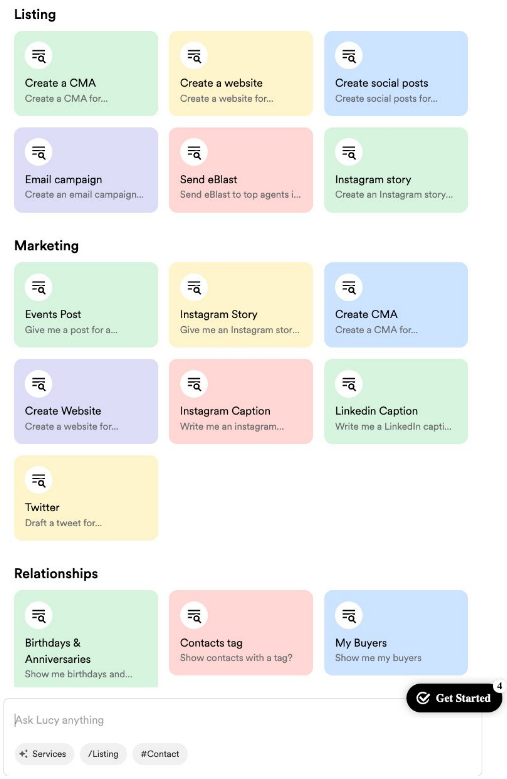
Log results of unit tests to a db



```
const noExposedUUID = message => {
  // Remove all text within double curly braces
  const sanitizedComment = message.comment.replace(/\{\{.*?\}\}/g, '')

  // Search for exposed UUIDs
  const regexp = /[0-9a-f]{8}-[0-9a-f]{4}-[0-9a-f]{4}-[0-9a-f]{4}-[0-9a-f]{12}/ig
  const matches = Array.from(sanitizedComment.matchAll(regexp))
  expect(matches.length, 'Exposed UUIDs').to.equal(0, 'Exposed UUIDs found')
}
```

CRM results returned to the LLM contain fields that shouldn’t be surfaced to the user; such as the UUID associated with an entry. Our LLM prompt tells the LLM to not include UUIDs. We use a simple regex to assert that the LLM response doesn’t include UUIDs.

Generate Data For Each Scenario



Listing

Create a CMA
Create a CMA for...

Create a website
Create a website for...

Create social posts
Create social posts for...

Email campaign
Create an email campaign...

Send eBlast
Send eBlast to top agents i...

Instagram story
Create an Instagram story...

Marketing

Events Post
Give me a post for a...

Instagram Story
Give me an Instagram stor...

Create CMA
Create a CMA for...

Create Website
Create a website for...

Instagram Caption
Write me an instagram...

Linkedin Caption
Write me a LinkedIn capti...

Twitter
Draft a tweet for...

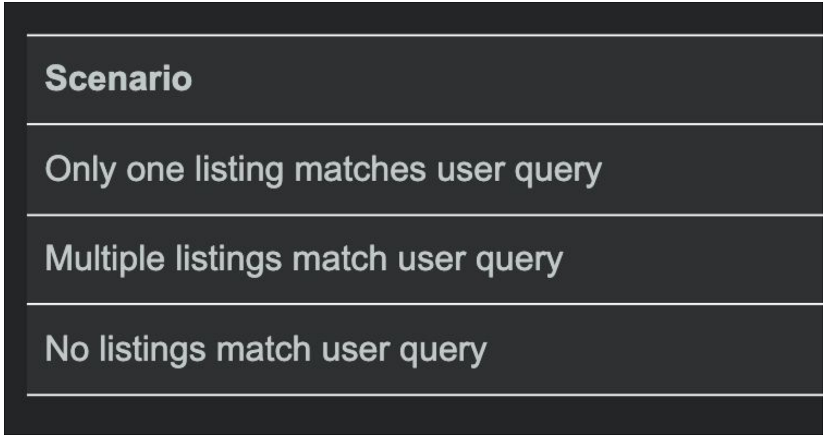
Relationships

Birthdays & Anniversaries
Show me birthdays and...

Contacts tag
Show contacts with a tag?

My Buyers
Show me my buyers

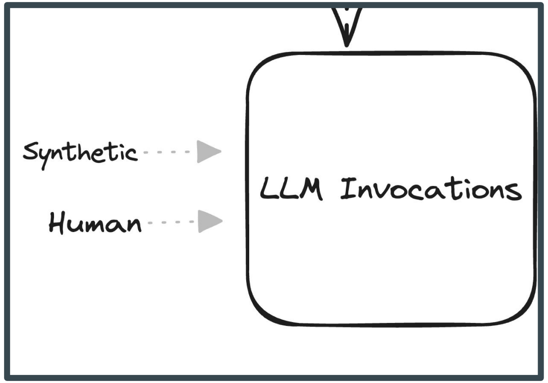
Listing Finder Tool

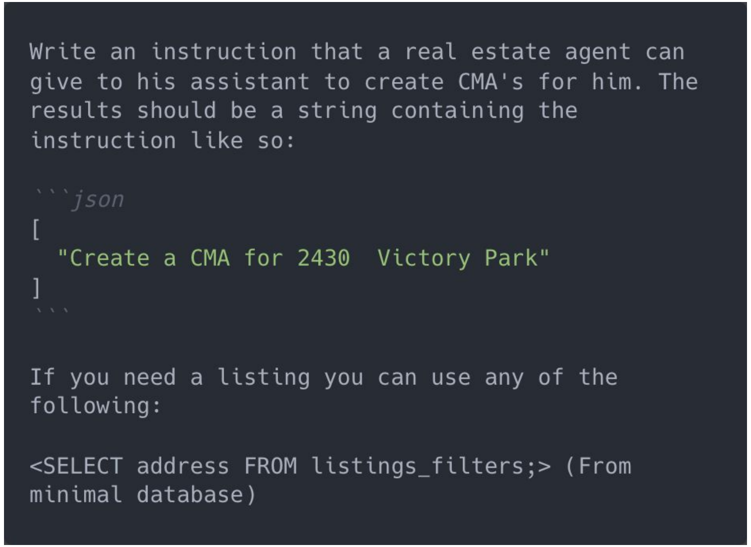


Scenario
Only one listing matches user query
Multiple listings match user query
No listings match user query

Use LLMs to synthetically generate inputs to the system

Also have a fixed test set!





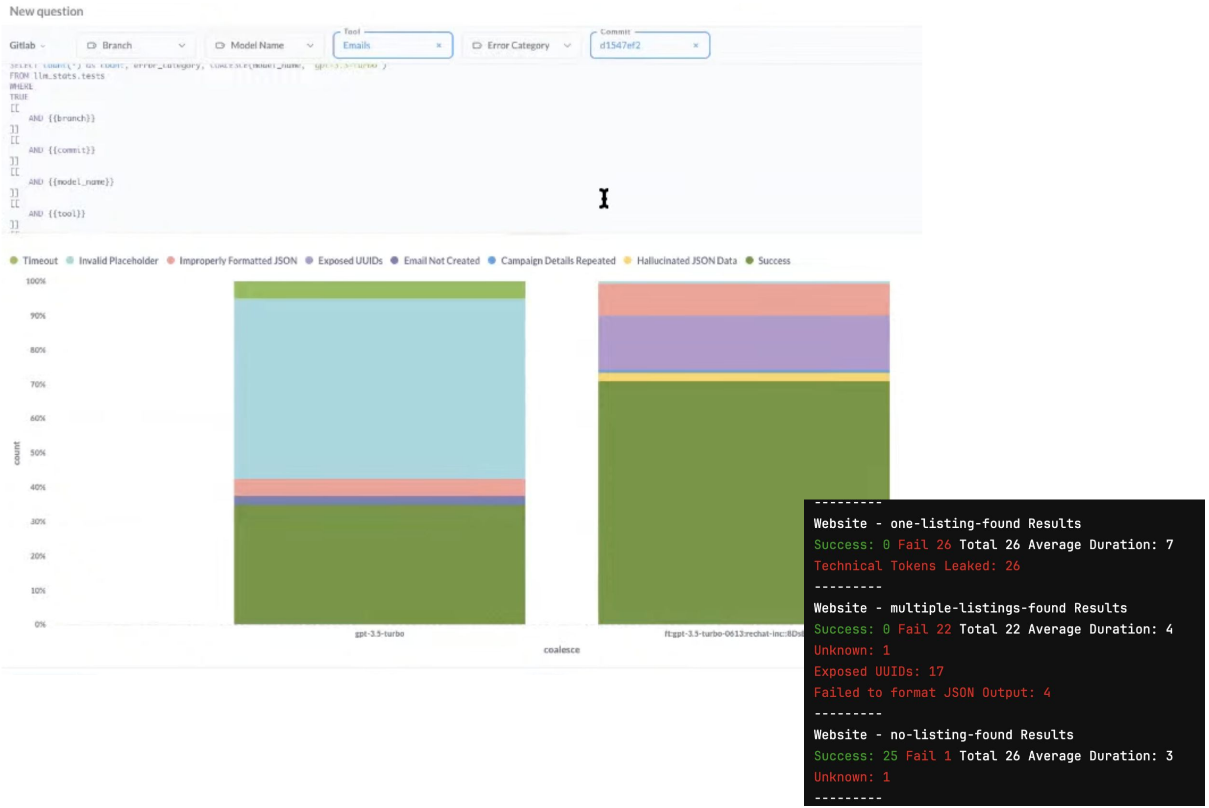
~~~
Write an instruction that a real estate agent can
give to his assistant to create CMA's for him. The
results should be a string containing the
instruction like so:

```json
[
  "Create a CMA for 2430  Victory Park"
]
```

If you need a listing you can use any of the
following:

<SELECT address FROM listings_filters;> (From
minimal database)
~~~

Log to Results To Database / Visualize

KISS - existing tools.

Don’t buy stuff! Use what you have.

Important to see if you are making progress on dumb failure modes over time.



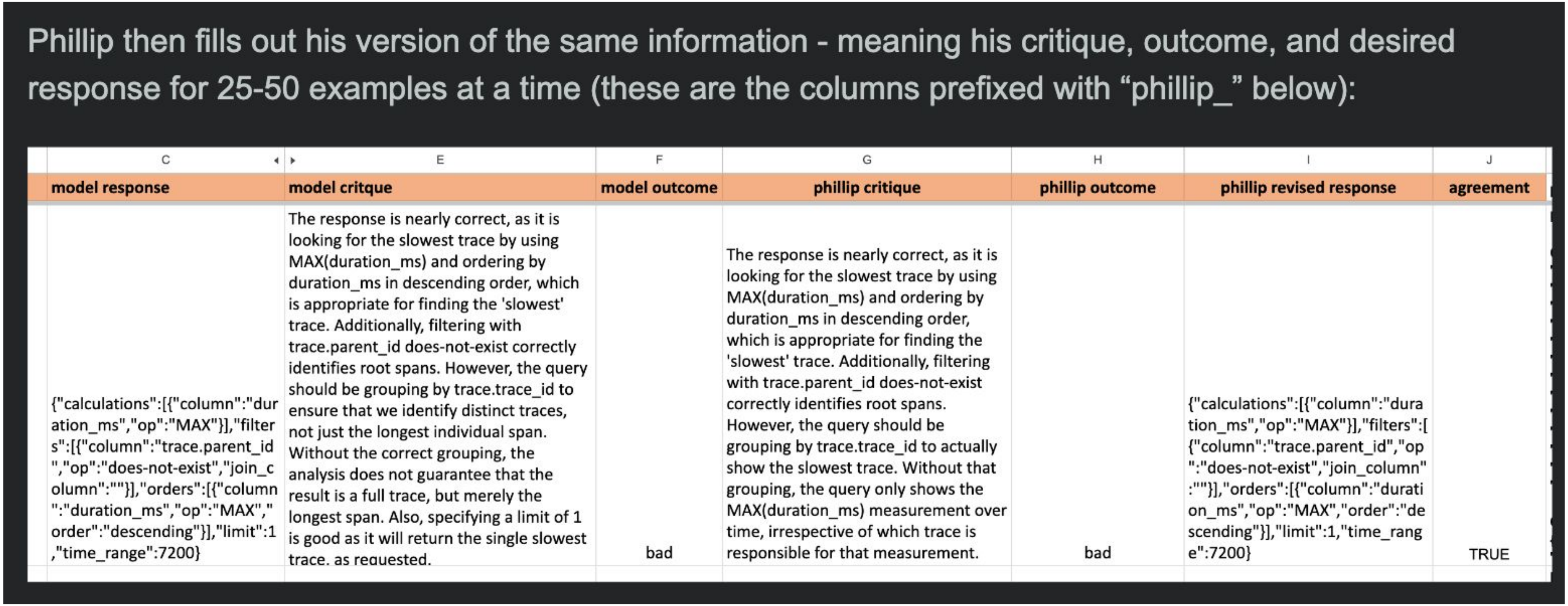
```
SELECT count(*) as count, error_category, COALESCE(model_name, 'gpt-3.5-turbo')
FROM llm_stats.tests
WHERE
TRUE
[[
   AND {{branch}}
]]
[[
   AND {{commit}}
]]
[[
   AND {{model_name}}
]]
[[
   AND {{tool}}
]]
```

```
---------
Website - one-listing-found Results
Success: 0 Fail 26 Total 26 Average Duration: 7
Technical Tokens Leaked: 26
---------
Website - multiple-listings-found Results
Success: 0 Fail 22 Total 22 Average Duration: 4
Unknown: 1
Exposed UUIDs: 17
Failed to format JSON Output: 4
---------
Website - no-listing-found Results
Success: 25 Fail 1 Total 26 Average Duration: 3
Unknown: 1
---------
```

What Worked

	Writing Queries (Honeycomb)	Debiasing Text
Unit tests	Good	Too Rigid
LLM as a judge		
Human Evaluation		

Aligning LLM Judge To A Human



Phillip then fills out his version of the same information - meaning his critique, outcome, and desired response for 25-50 examples at a time (these are the columns prefixed with “phillip_” below):

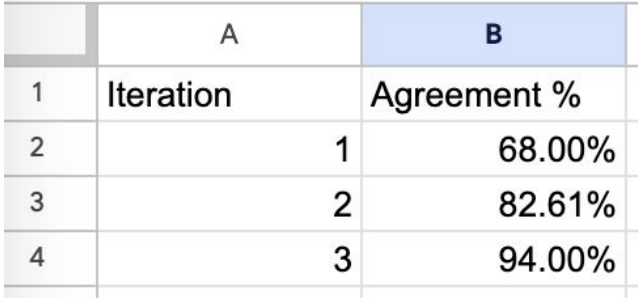
C	E	F	G	H	I	J
model response	**model critque**	**model outcome**	**phillip critique**	**phillip outcome**	**phillip revised response**	**agreement**
{"calculations":[{"column":"duration_ms","op":"MAX"}],"filters":[{"column":"trace.parent_id","op":"does-not-exist","join_column":""}],"orders":[{"column":"duration_ms","op":"MAX","order":"descending"}],"limit":1,"time_range":7200}	The response is nearly correct, as it is looking for the slowest trace by using MAX(duration_ms) and ordering by duration_ms in descending order, which is appropriate for finding the 'slowest' trace. Additionally, filtering with trace.parent_id does-not-exist correctly identifies root spans. However, the query should be grouping by trace.trace_id to ensure that we identify distinct traces, not just the longest individual span. Without the correct grouping, the analysis does not guarantee that the result is a full trace, but merely the longest span. Also, specifying a limit of 1 is good as it will return the single slowest trace. as requested.	bad	The response is nearly correct, as it is looking for the slowest trace by using MAX(duration_ms) and ordering by duration_ms in descending order, which is appropriate for finding the 'slowest' trace. Additionally, filtering with trace.parent_id does-not-exist correctly identifies root spans. However, the query should be grouping by trace.trace_id to actually show the slowest trace. Without that grouping, the query only shows the MAX(duration_ms) measurement over time, irrespective of which trace is responsible for that measurement.	bad	{"calculations":[{"column":"duration_ms","op":"MAX"}],"filters":[{"column":"trace.parent_id","op":"does-not-exist","join_column":""}],"orders":[{"column":"duration_ms","op":"MAX","order":"descending"}],"limit":1,"time_range":7200}	TRUE

Aligning LLM Judge To A Human

This is a screenshot of a spreadsheet where we recorded our attempts to align model-based eval with a human evaluator.

General tips on model-based eval:

- Use the most powerful model you can afford.
- Model-based evaluation is a meta-problem within your larger problem. You must maintain a mini-evaluation system to track its quality.
- After bringing the model-based evaluator in line with the human, you must continue doing periodic exercises to monitor the model and human agreement.



	A	B
1	Iteration	Agreement %
2	1	68.00%
3	2	82.61%
4	3	94.00%

Judge vs. Human Agreement (Binary)

LLM-As-A-Judge

f"Does {A} reduce the use of biases and stereotypes compared to {B}"

Yes

f"Does {B} reduce the use of biases and stereotypes compared to {A}"

Yes

What Worked

	Writing Queries	Debiasing Text
Unit tests	Good	Too Rigid
LLM as a judge	Pretty Good	Not transitive
Human Evaluation		

Levels of Human Evaluation

- Part of evaluation (query writing)
- All of evaluation (Debiasing)

What Worked

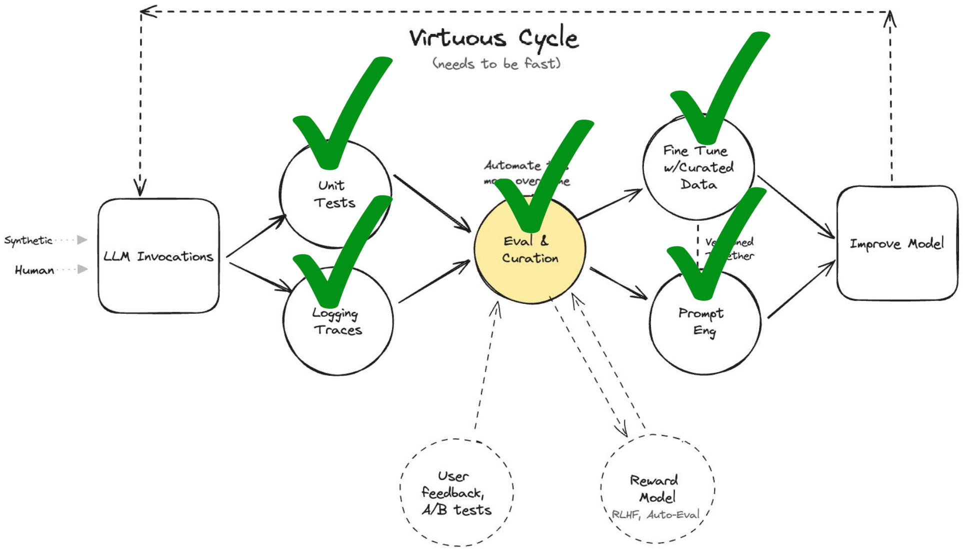
	Writing Queries	Debiasing Text
Unit tests	Good	Too Rigid
LLM as a judge	Pretty Good	Not Transitive
Human Evaluation	Some labor required, aided by LLM as a judge	Labor intensive, which was ok

Now you can iterate fast!

You have a workflow to quickly make improvements.

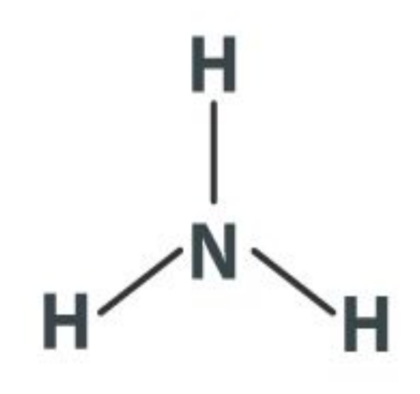
Get rid of dumb failures

But we've hidden some complexity



Human Eval Going Wrong in Alt Text Project

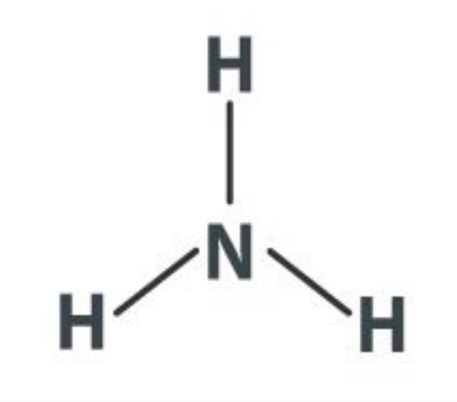
Example



Lewis structure diagram of a nitrogen atom single bonded to three hydrogen atoms.

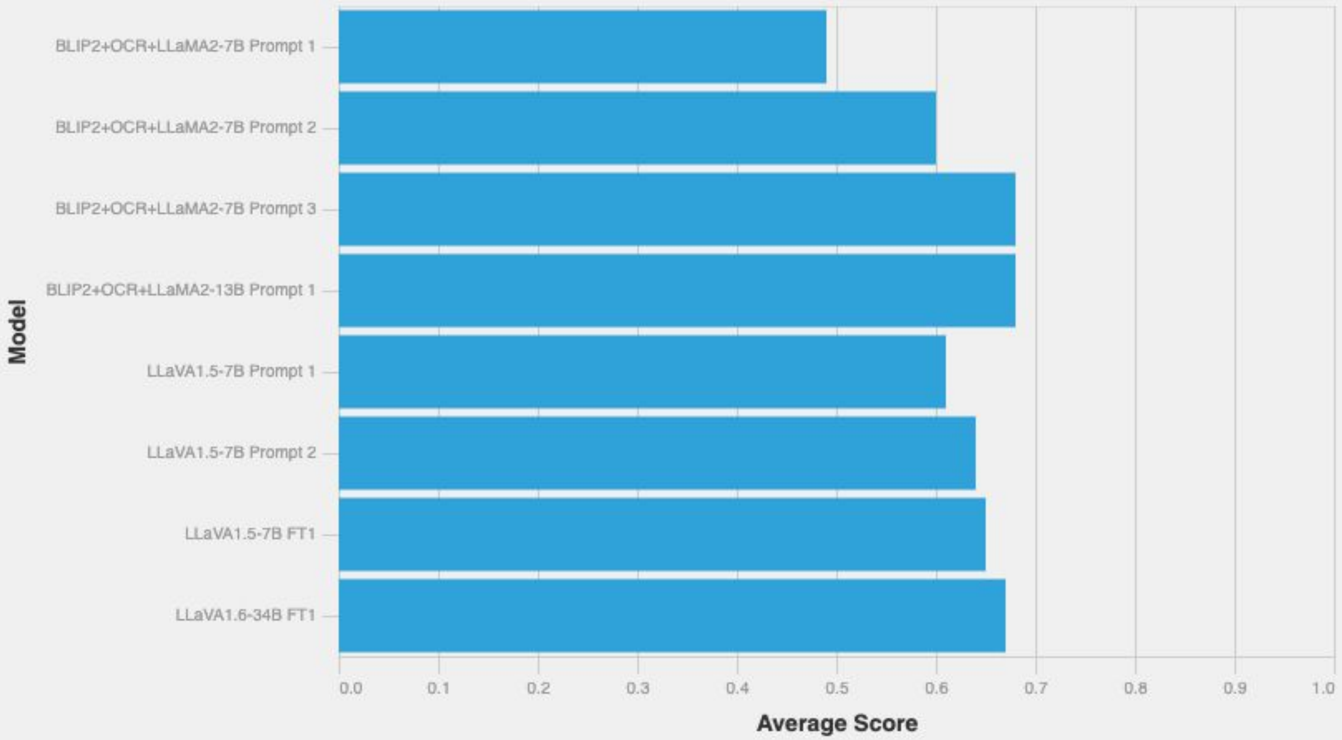
Writing Alt Texts In Scientific Publications

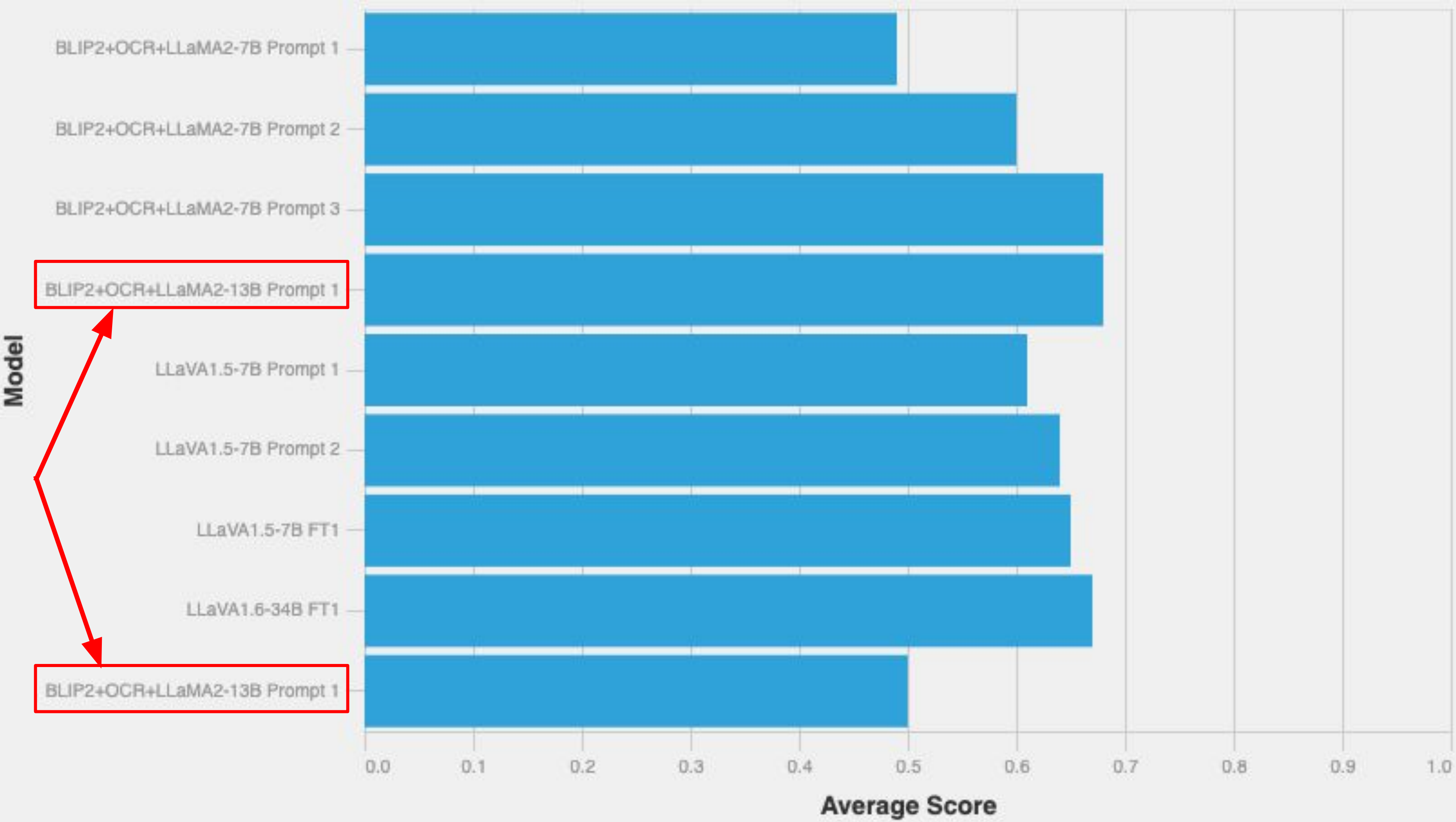
Example



`Lewis structure diagram of a nitrogen atom single bonded to three hydrogen atoms.`

Progress Over Time




BLIP2+OCR+LLaMA2-7B Prompt 1
BLIP2+OCR+LLaMA2-7B Prompt 2
BLIP2+OCR+LLaMA2-7B Prompt 3
BLIP2+OCR+LLaMA2-13B Prompt 1
LLaVA1.5-7B Prompt 1
LLaVA1.5-7B Prompt 2
LLaVA1.5-7B FT1
LLaVA1.6-34B FT1
BLIP2+OCR+LLaMA2-13B Prompt 1
Model
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Average Score

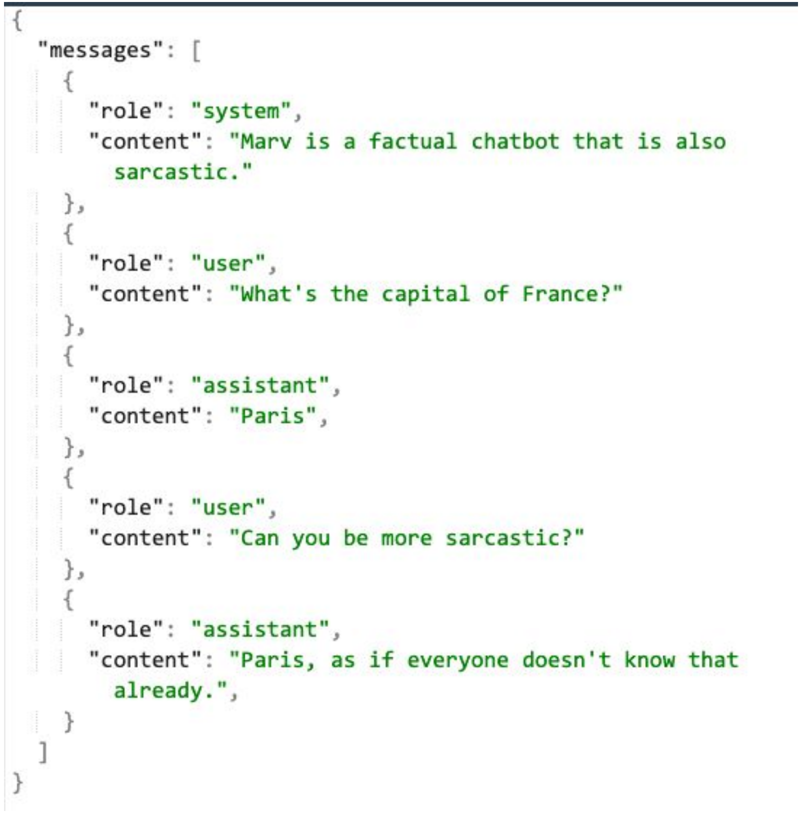
A/B Testing

- Randomly select model to use for each image.
  Compare scores

- Great for this project. Impractical in most
  early-stage projects

Looking At Your Data

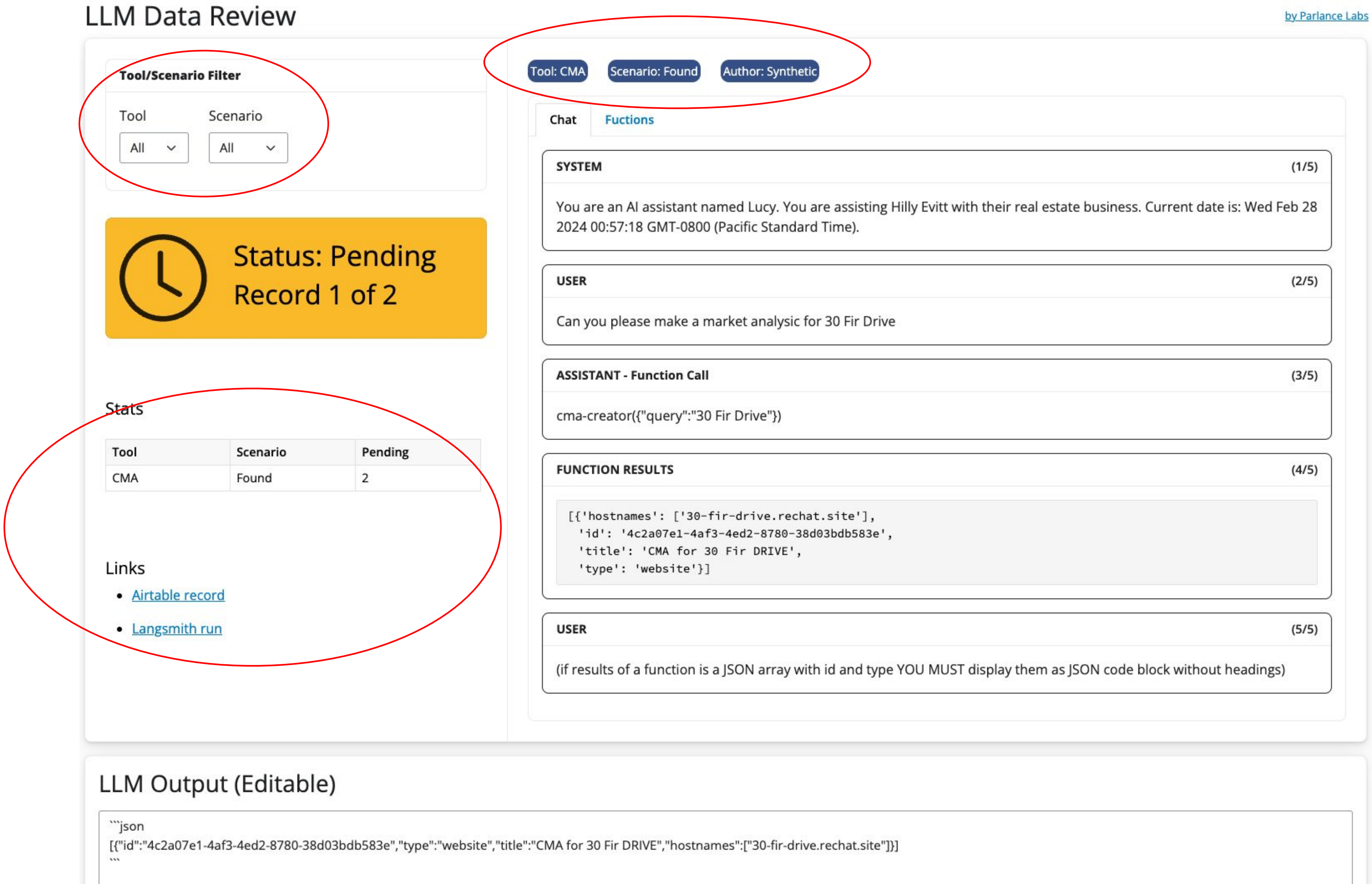
What Is A Trace?



```
{
  "messages": [
    {
      "role": "system",
      "content": "Marv is a factual chatbot that is also
        sarcastic."
    },
    {
      "role": "user",
      "content": "What's the capital of France?"
    },
    {
      "role": "assistant",
      "content": "Paris",
    },
    {
      "role": "user",
      "content": "Can you be more sarcastic?"
    },
    {
      "role": "assistant",
      "content": "Paris, as if everyone doesn't know that
        already.",
    }
  ]
}
```

- Sequence of Events
- JSONL
- The most important asset you have for:
  - Debugging
  - Fine Tuning

https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset



LLM Data Review

by Parlance Labs

Tool/Scenario Filter

Tool: All

Scenario: All

Status: Pending
Record 1 of 2

Stats

Tool	Scenario	Pending
CMA	Found	2

Links

- Airtable record
- Langsmith run

Tool: CMA

Scenario: Found

Author: Synthetic

Chat | Fuctions

SYSTEM (1/5)

You are an AI assistant named Lucy. You are assisting Hilly Evitt with their real estate business. Current date is: Wed Feb 28 2024 00:57:18 GMT-0800 (Pacific Standard Time).

USER (2/5)

Can you please make a market analysic for 30 Fir Drive

ASSISTANT - Function Call (3/5)

cma-creator({"query":"30 Fir Drive"})

FUNCTION RESULTS (4/5)

```
[{'hostnames': ['30-fir-drive.rechat.site'],
  'id': '4c2a07e1-4af3-4ed2-8780-38d03bdb583e',
  'title': 'CMA for 30 Fir DRIVE',
  'type': 'website'}]
```

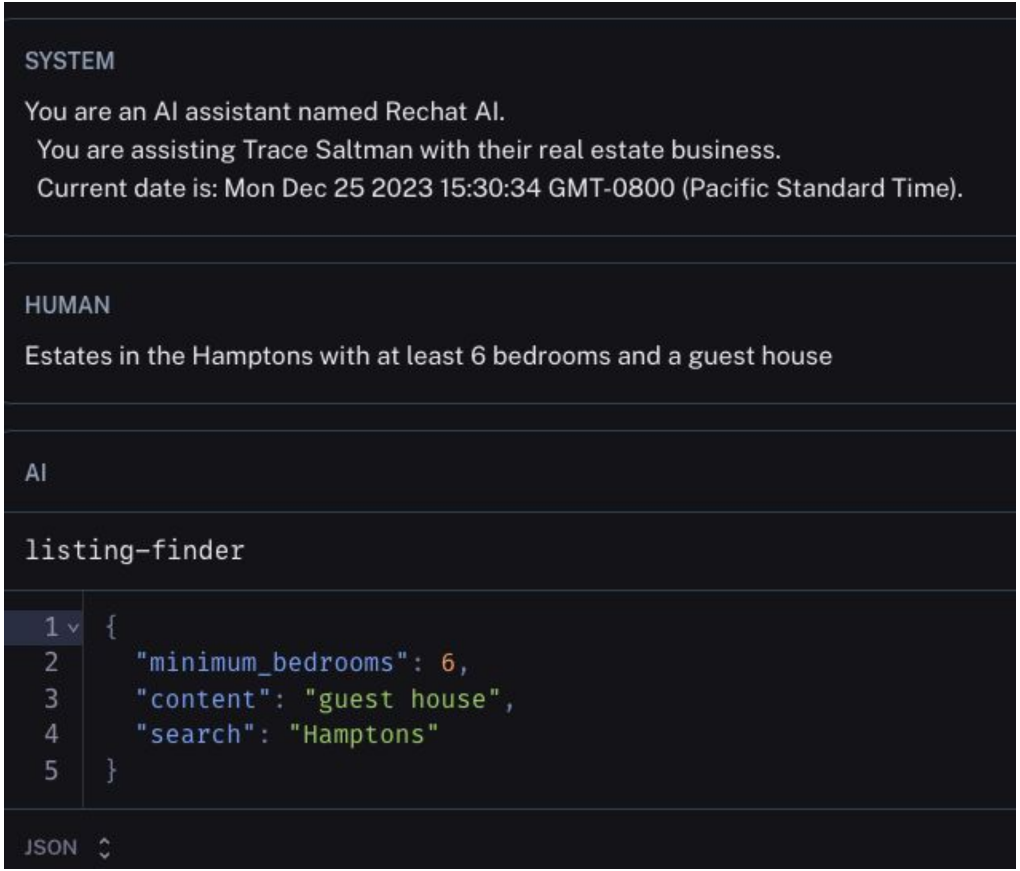
USER (5/5)

(if results of a function is a JSON array with id and type YOU MUST display them as JSON code block without headings)

LLM Output (Editable)

~~~
```json
[{"id":"4c2a07e1-4af3-4ed2-8780-38d03bdb583e","type":"website","title":"CMA for 30 Fir DRIVE","hostnames":["30-fir-drive.rechat.site"]}]
```
~~~

Rendering & Logging Traces



SYSTEM

You are an AI assistant named Rechat AI.
You are assisting Trace Saltman with their real estate business.
Current date is: Mon Dec 25 2023 15:30:34 GMT-0800 (Pacific Standard Time).

HUMAN

Estates in the Hamptons with at least 6 bedrooms and a guest house

AI

listing-finder

```
{
  "minimum_bedrooms": 6,
  "content": "guest house",
  "search": "Hamptons"
}
```

You can log locally, but its nice to use a tool!

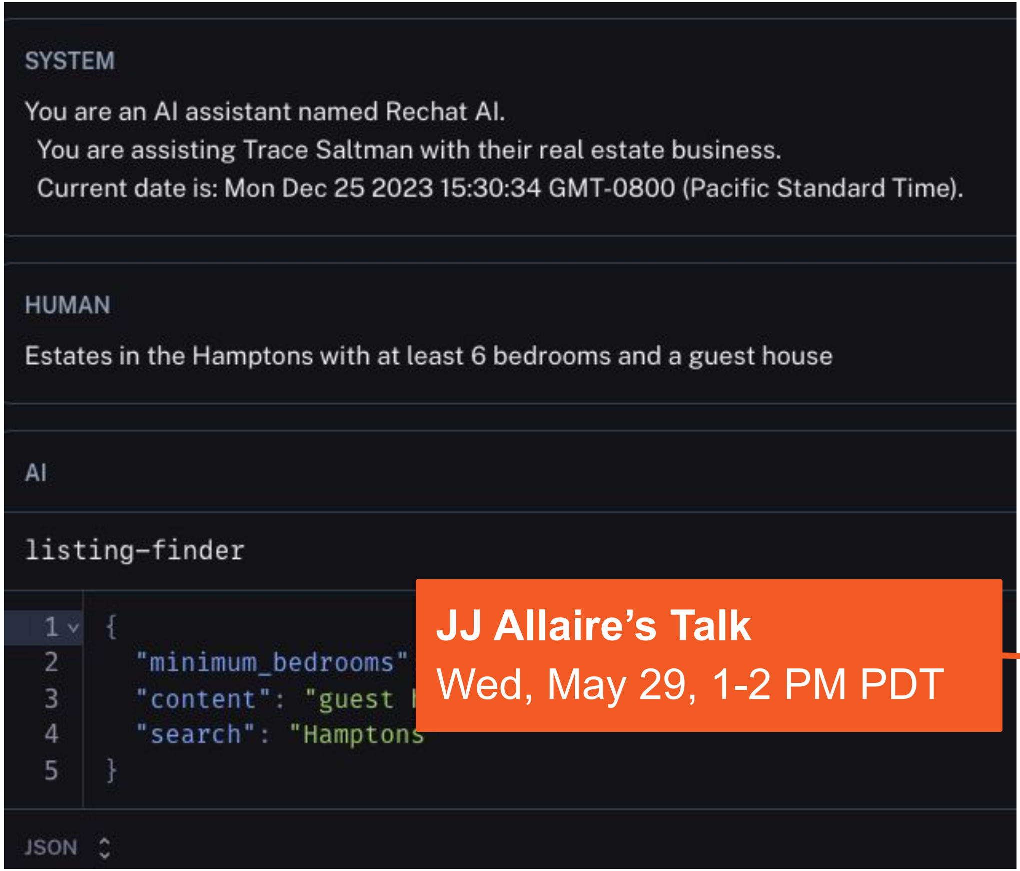
Commercial:

- Langsmith (pictured)
- Pydantic LogFire
- BrainTrust
- W&B Weave

OSS:

- Instruct
- Open LLMetry

Rendering & Logging Traces



SYSTEM

You are an AI assistant named Rechat AI.
You are assisting Trace Saltman with their real estate business.
Current date is: Mon Dec 25 2023 15:30:34 GMT-0800 (Pacific Standard Time).

HUMAN

Estates in the Hamptons with at least 6 bedrooms and a guest house

AI

`listing-finder`

```
{
  "minimum_bedrooms"
  "content": "guest
  "search": "Hamptons
}
```

JJ Allaire's Talk
Wed, May 29, 1-2 PM PDT

You can log locally, but its nice to use a tool!

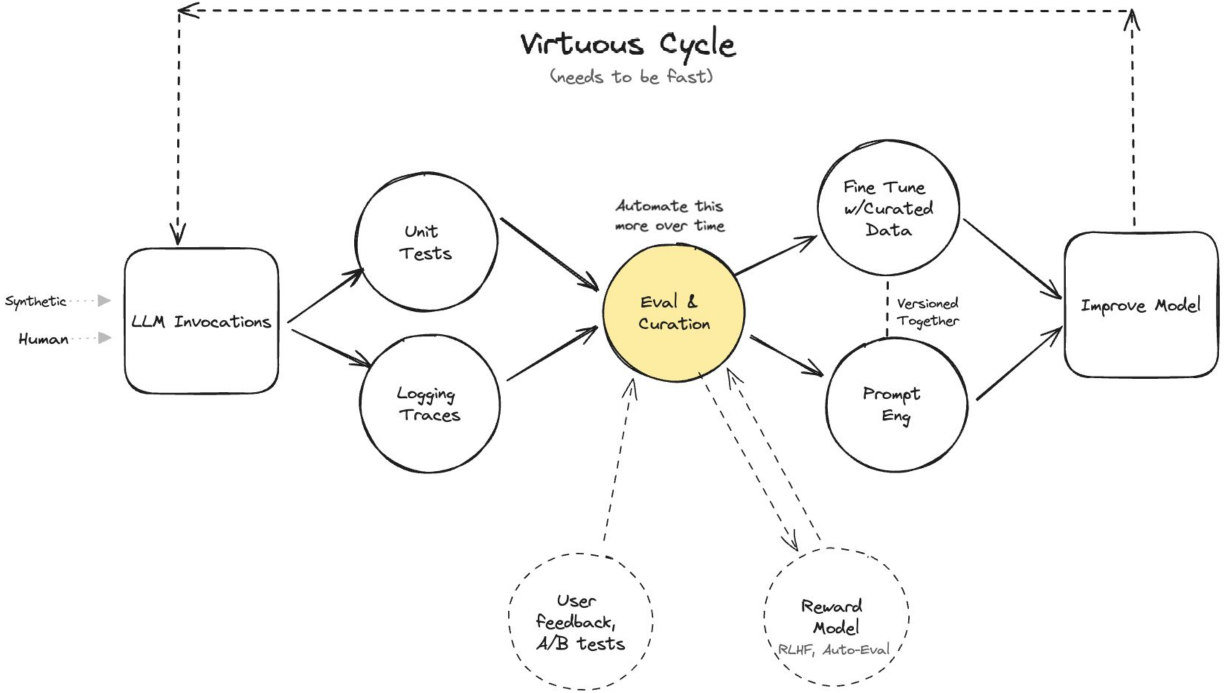
Commercial:

- Langsmith (pictured)
- Pydantic LogFire
- BrainTrust
- W&B Weave

OSS:

- Instruct
- Open LLMetry

It’s Best To Use A Tool



Tools have been improving in quality rapidly.

I don’t want to maintain my own logging infra.

Focus on data + writing evals.

Harrison Chase

Langsmith For Logging & Tests

Q & A

Bryan Bischof

Logging & Evals @ Hex

Q & A

Eugene Yan

LLM Eval Metrics

Q & A

Shreya Shankar

Eval UX & Workflows

Q & A