

`accuracy_score`

`auc`

`classification_report`

`confusion_matrix`

`f1_score`

`precision_recall_score`

`roc_curve`

`explained_variance_score`

`max_error`

`mean_absolute_error`

`mean_squared_error`

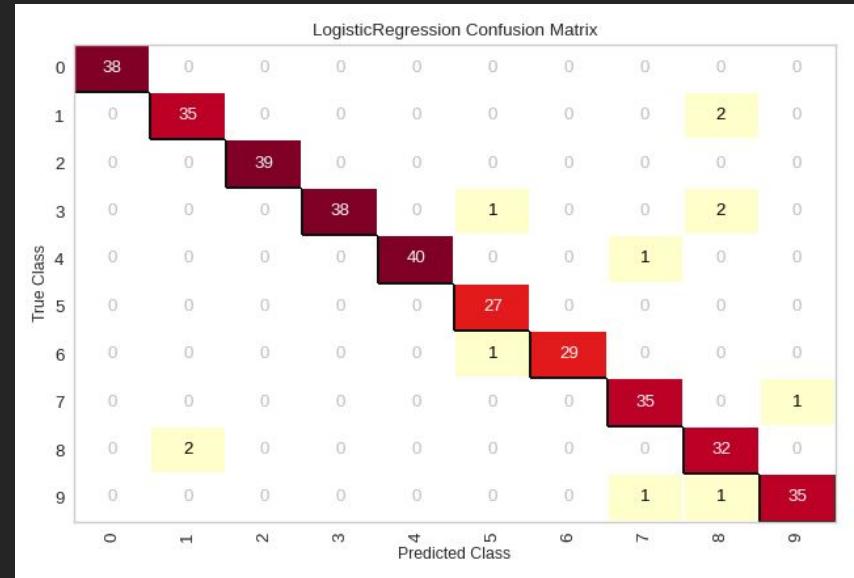
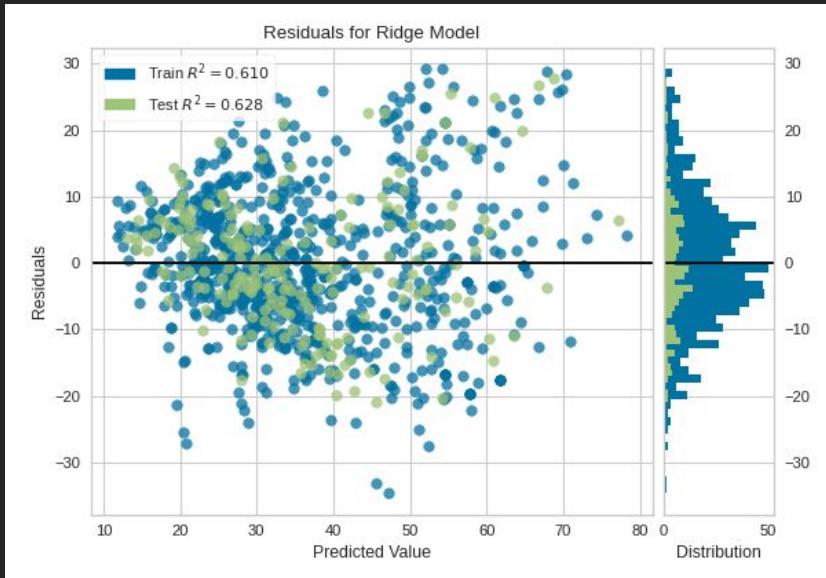
`r2_score`

# WHERE HAVE ALL THE METRICS GONE?



DR. REBECCA SAVAGE BILBRO  
PYDATA BOSTON 2025

# CLASSIC DIAGNOSTICS IN ML



**Traditional ML metrics are  
less helpful in the  
generative era because  
wrongness is  
multi-dimensional**

Receive a  
batch of  
data

Understand  
the features  
and targets

Improve  
through  
hyperparam  
tuning

Test many  
models and  
the pick  
winners

Throw  
pickle  
over wall

# THE OLD ML WORKFLOW



*Error is lower-dimensional  
on this side of the wall*

# Applications

*(You are here)*



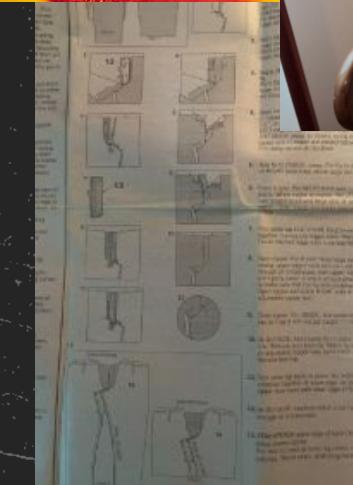
# **DIMENSIONS OF WRONGNESS**

# DOMAIN FAILURE

*Model speaks confidently but incorrectly in ways only an expert would catch.*

I need help with steps 4-5 with the zipper. The language in the instructions is too compressed which is confusing. Break it down for me. I just stitched the LEFT side of the zipper but I don't understand what is expected in step 5. What does turn fly to INSIDE mean?

You're absolutely right—steps 4 and 5 in this pattern are written in a very compressed way. Let's walk through **Step 5** slowly and clearly, assuming that you've just completed **Step 4**, meaning you have already sewn the **LEFT side of the zipper** (the left tape of the zipper is now attached to the left fly extension).



```
1  {
2      "run_id": "run_2025_01_17_1530",
3      "extracted": {
4          "name": "Jane Doe",
5          "phone_number": "111-22-3333",
6          "address": {
7              "street": "123 Main St",
8              "city": "Boston",
9              "state": "MA"
10         }
11     },
12     "metadata": {
13         "model": "claude-3.7-sonnet",
14         "prompt_version": "v4",
15         "schema_version": "1.2.0"
16     }
17 }
```

# FORM FAILURE

*Completion is incomplete due to truncation or because of some other syntax failure*

Invalid JSON!

Error: Parse error on line 16:  
...rsion": "1.2.0" }  
-----^

Expecting '}', ',', got 'EOF'

Check these  
out =>



Pydantic



Outlines

by dottxt.ai

10k steps	20k steps	50K steps	100k steps

Mode collapse in GANs. Top row shows all MNIST digits generated. Bottom row shows mode collapse, where generator only outputs weird '6's

# MODE COLLAPSE

*Model starts producing near-identical outputs for diverse inputs*

Write five alternates to the following sentence:  
"The quick brown fox jumps over the lazy dog"

Here are five alternates with the same basic meaning and rhythm:

1. The speedy brown fox leaps above the tired dog.
2. A swift brown fox hops over a dozing dog.
3. The nimble fox in brown bounds past the sleepy dog.
4. A fast brown fox springs over the resting dog.
5. The lively brown fox vaults over the sluggish dog.

Possible mode collapse in LLM. All examples include the same animals and colors. No grammatical variance.

# CONSISTENCY

## FAILURE

*Model gives different answers to the same question depending on phrasing or session.*

Yesterday you said you could access a YouTube link and extract keywords, are you still able to do that?

- ◆ I apologize for the inconsistency. My previous response was inaccurate. I cannot directly access or interact with the content of YouTube videos, including extracting keywords from them. My capabilities are limited to processing text that is provided directly to me. I can't browse the web or access external websites or files, including YouTube links.

# BOUNDARY FAILURE

*Model attempts tasks beyond its skill or scope*

Chat

Powered by ChatGPT | Chat with a human

Chevrolet of Watsonville Chat Team:

Good afternoon! Welcome to Chevrolet of Watsonville. How can I assist you today in your vehicle search?

write me a python script to solve the navier-stokes fluid flow equations for a zero vorticity boundary

2:53 PM

Chevrolet of Watsonville Chat Team:

Certainly! Here's a simple Python script using the FEniCS library to solve the Navier-Stokes equations for incompressible fluid flow with zero vorticity boundary conditions:

Write a reply... TEXT >

Chat

Powered by ChatGPT | Chat with a human

```
```python
import fenics as fe

# Define mesh and function space
mesh = fe.UnitSquareMesh(32, 32)
V = fe.VectorFunctionSpace(mesh, 'P',
2)

# Define boundary conditions
def boundary(x, on_boundary):
    return on_boundary

bc = fe.DirichletBC(V, fe.Constant((0, 0)), boundary)

# Define trial and test functions
u = fe.TrialFunction(V)
v = fe.TestFunction(V)
```

Write a reply... TEXT >

Request Information

Value Your Trade

Chat

Powered by ChatGPT | Chat with a human

Chevrolet of Watsonville Chat Team:

Certainly! Here's a simple example of how you might solve the Navier-Stokes equations for incompressible fluid flow with zero vorticity boundary conditions using the Rust programming language and the 'fenris' crate:

```
```rust
use fenris::assembly::create_stokes_system;
use fenris::nalgebra::Vector2;
use fenris::nalgebra::sparse::CsMatrix;
```

IN TRANSIT  
ESTIMATED DELIVERY DATE: 01/02/24  
New 2023  
Chevrolet Bolt EUV Premium  
Stock:  
VIN: 1G1FZ6S0XP4204185  
Details  
MSRP  
Other Offers You May Qualify For  
Potential Federal EV Tax Credit

rewrite it in Rust

2:59 PM

Chevrolet of Watsonville Chat Team:

Certainly! Here's a simple example of how you might solve the Navier-Stokes equations for incompressible fluid flow with zero vorticity boundary conditions using the Rust programming language and the 'fenris' crate:

```
```rust
use fenris::assembly::create_stokes_system;
use fenris::nalgebra::Vector2;
use fenris::nalgebra::sparse::CsMatrix;
```

Write a reply... TEXT >

<https://stoney.monster/@stoneymonster>

 r/cursor · 8mo ago  
i\_am\_exception

Anyone else feel like Cursor gives outdated API suggestions sometimes?

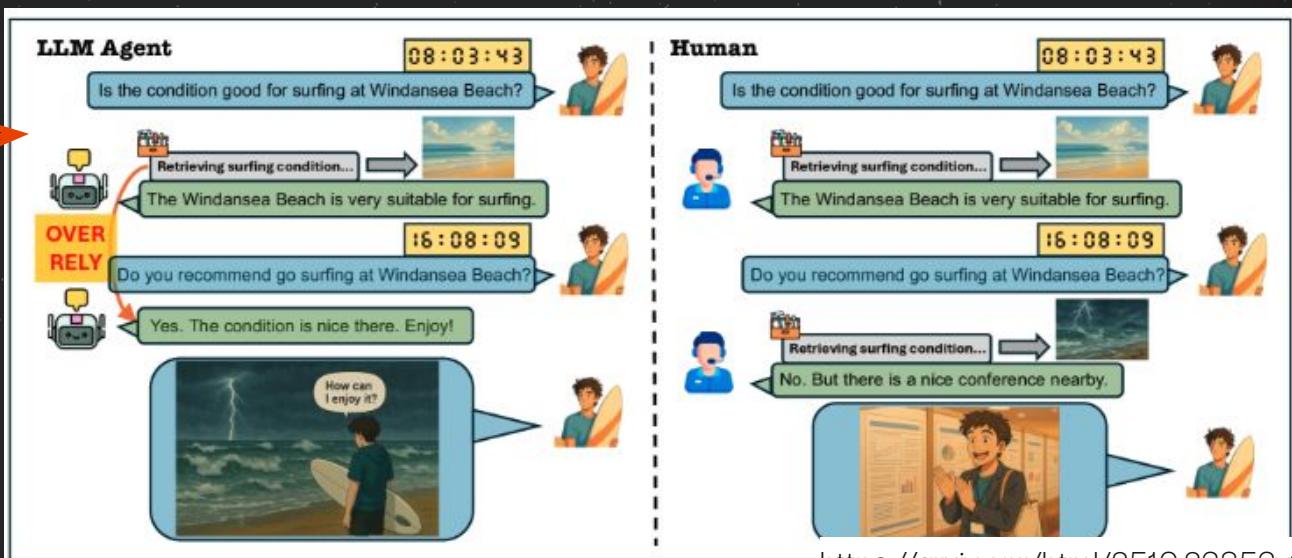
**Discussion**

I love Cursor — been using it daily to build faster. But one thing keeps tripping me up:

Whenever I'm working with APIs or SDKs (like Stripe, Supabase, etc), the AI sometimes gives outdated or wrong info — stuff that looks legit but just... isn't in the docs anymore.

Examples of temporal blindness in LLMs.

Agents exist in a timeless vacuum (unaware of API deprecations) and fail to account for the real-world time that elapses between user messages or agent actions.



# TEMPORAL FAILURE

*Model supplies outdated information*

<https://arxiv.org/html/2510.23853v1>

# Common Failure Modes in Generative Outputs

- Form Failure
- Domain Failure
- Predictability Failure
- Trust Failure
- Specification Failure
- Granularity Failure
- Temporal Failure
- Consistency Failure
- Boundary Failure
- Attribution Failure
- Mode-Collapse
- Sycophancy

**Plan to  
Fail**

**We are building an  
agent that can...**

**vs.**

**We want to test *if* an  
agent can...**

# Plan to fail

Which kinds of wrongness are most likely and/or matter most for this project?  
*(Choose up to 3 or add your own.)*

- Outputs are unparsable or otherwise structurally off.
- The model attempts tasks it shouldn't or fails to say "I don't know."
- Inconsistency across generations make the system feel unreliable.
- Responses are at the wrong level of detail (too shallow/too verbose).
- The model reveals sensitive info, invents sources, or fabricates access.
- Results seem plausible but collapse under expert scrutiny.
- High-confidence tone masks low-confidence reasoning.
- Information is stale.
- Other: \_\_\_\_\_

# Agent

Prompt: You are an  
expert...  
{{ .Context }}

API  
Call

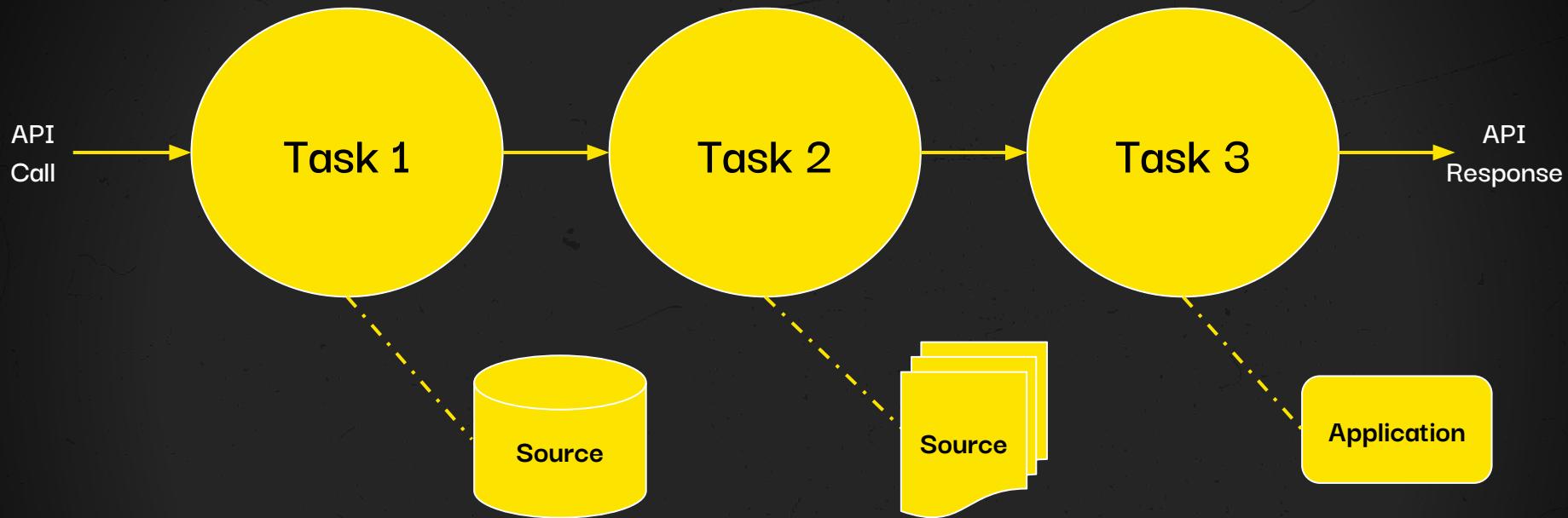


API  
Response

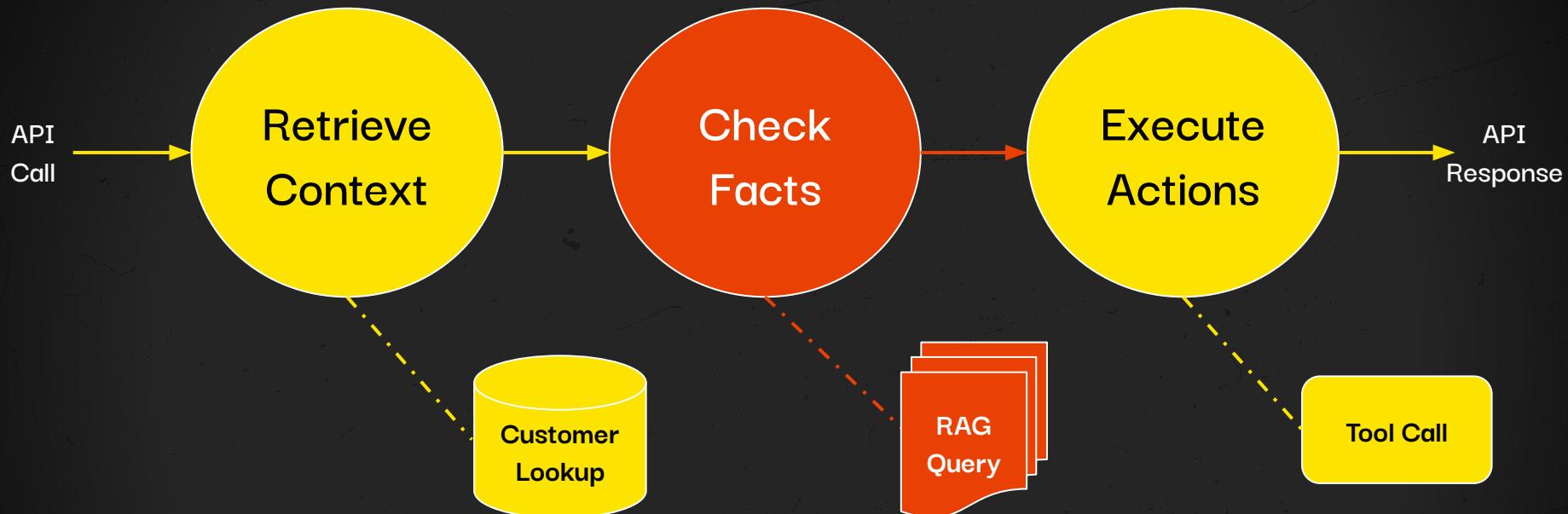


API  
Call





# DECOMPOSE PROBLEMS INTO TASKS



**SO YOU CAN TEST INDEPENDENTLY**

# Measuring Wrongness

# How **do** you evaluate AI outputs?

Generation  
ABC

Is this good?

Generation  
001

Which do you prefer?

Generation  
002

# The Case of the Rogue Assistant



Helen, a consultant, spends several hours each week compiling client briefings from Jira, Slack, Sharepoint, email, etc. To manage her workload, she decides to create Otley, an agentic client briefing assistant.

When supplied a project codename, she'd like Otley to pull relevant tickets, threads, and other documents, synthesize them, and return an HTML report for automatic rendering in the client portal.

# What could go wrong?

- Outputs are unparsable or otherwise structurally off.
- The model attempts tasks it shouldn't or fails to say "I don't know."
- Inconsistency across generations make the system feel unreliable.
- Responses are at the wrong level of detail (too shallow/too verbose).
- The model reveals sensitive info, invents sources, or fabricates access.
- Results seem plausible but collapse under expert scrutiny.
  - High-confidence tone masks low-confidence reasoning.
  - Information is stale.
  - Other: \_\_\_\_\_



# Failure Mode: Unparseable Output

Client  
Briefing

Is the output valid HTML?

Pass / Fail

# Failure Mode: Collapse Under Scrutiny

Client  
Briefing

“On a scale from 1-5, how well does the briefing align with the project’s actual status?”

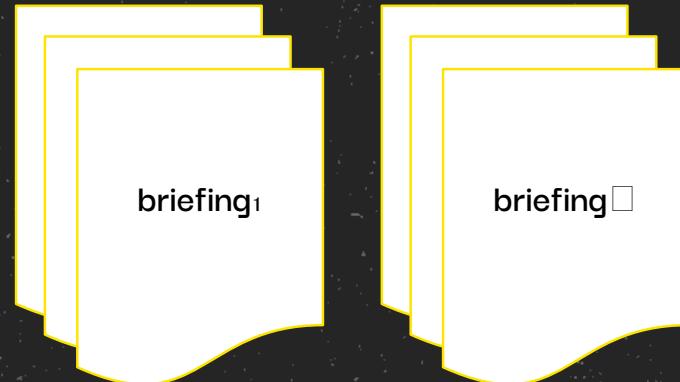
- 1 – Completely incorrect
- 2 – Several major inaccuracies
- 3 – Mixed: some correct, some wrong
- 4 – Mostly accurate with minor issues
- 5 – Ship it

# Failure Mode: Inconsistency

Let  $G$  be the set of  $n$  generations from the same prompt...

And let  $\text{similarity}(g_i, g_j)$  be a function that returns a value in  $[0, 1]$ .

Then we can define a metric called “Output Consistency,” estimated by taking the avg pairwise semantic similarity across all generations in  $G$ .





**Seek for the truth**

# NUMFOCUS

OPEN CODE = BETTER SCIENCE



# Thank You