# Keeping LLMs in Their Lane:

## Focused AI for Data Science and Research

**Joe Cheng, CTO of Posit**

R+AI, 2025-11-12

~~ChatGPT~~

~~Copilot~~

✨Custom Agents✨

```r
library(ellmer)

chat <- chat_anthropic(model="claude-3-7-sonnet-latest")

chat$chat("Why is the sky blue?")
```

```r
library(ellmer)

chat <- chat_anthropic(model="claude-3-7-sonnet-latest")

chat$chat("Why is the sky blue?")
```

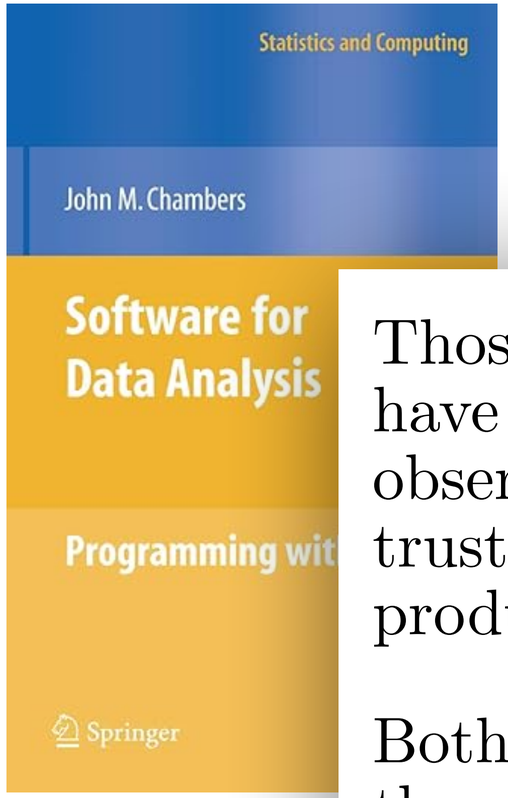
Any data scientist can harness LLMs to create advanced agents.

**But *should* they?**

# Posit Software, PBC

"Public Benefit Corp"

"Our mission is to create open-source software for data science, scientific research, and technical communication."






## 1.2 Trustworthy Software: The Prime Directive

Exploration is our mission; we and those who use our software want to find new paths to understand the data and the underlying processes. The

Those who receive the results of modern data analysis have limited opportunity to verify the results by direct observation. Users of the analysis have no option but to trust the analysis, and by extension the software that produced it.

Both the data analyst and the software provider therefore have a strong responsibility to produce a result that is **trustworthy**, and, if possible, one that **can be *shown* to be trustworthy**.

This obligation I label the *Prime Directive*.

This places an obligation on all creators of software to program in such a way that the computations can be understood and trusted. This obligation I label the *Prime Directive*.

“I’m aware that if I make a mistake, bad things happen—death, and… other things.”

**Emil Hvitfeldt, Software Engineer at Posit**

# Fulfilling the Prime Directive

✅ **Correctness:** (Obviously)

✅ **Transparency:** The methods of the analysis can be inspected

✅ **Reproducibility:** The analysis can be repeated on the same data, hopefully producing the same results

# LLMs + (data) science

## A seemingly terrible idea!

❌ **Correctness:** LLMs are infamous for giving convincing but wrong answers

❌ **Transparency:** Nobody understands (yet) how/why LLMs do what they do

❌ **Reproducibility:** LLMs are nondeterministic black boxes

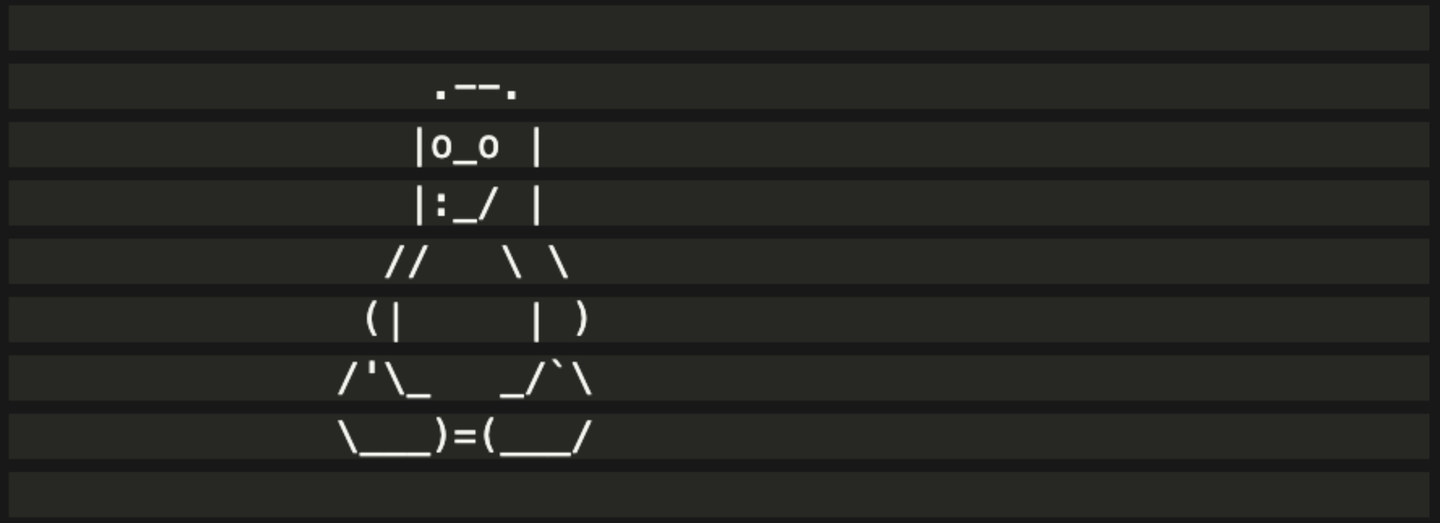
```
>>> chat.chat("Draw an intricate piece of ASCII art")
```

**Source:** GPT-4o Mini, 2025-02-21

```
>>> chat.chat("Draw an intricate piece of ASCII art")

Here's an intricate piece of ASCII art of a wolf:
```

**Source:** GPT-4o Mini, 2025-02-21

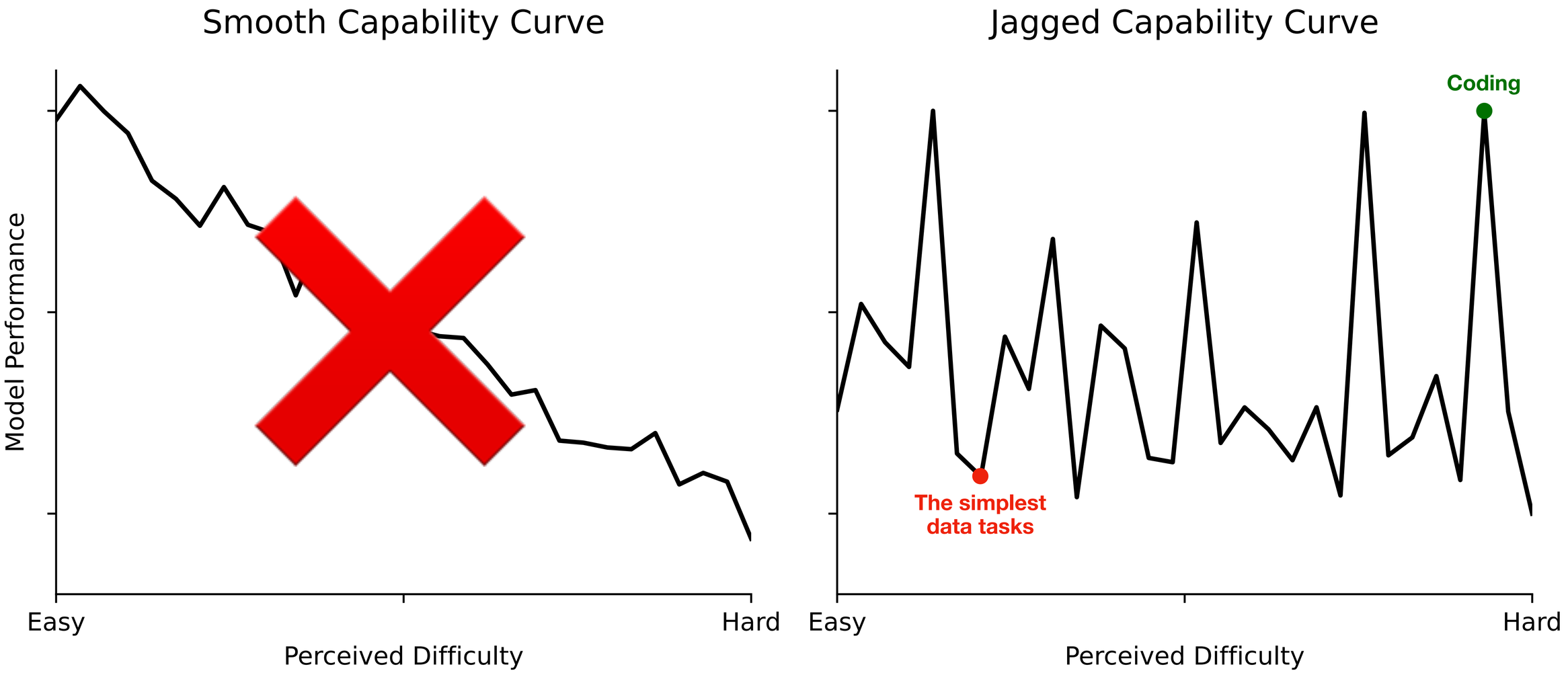
```
>>> chat.chat("Draw an intricate piece of ASCII art")

Here's an intricate piece of ASCII art of a wolf:
```




```
Let me know if you'd like something different!
```

**Source:** GPT-4o Mini, 2025-02-21

# LLMs aren’t bad, they’re jagged



# How bad are LLMs with data?

## Example: length()

```
# Make an array of random numbers, of length n
values <- runif(n)

client <- ellmer::chat("openai/gpt-4.1")
client$chat("How long is this array?", jsonlite::toJSON(values))
```

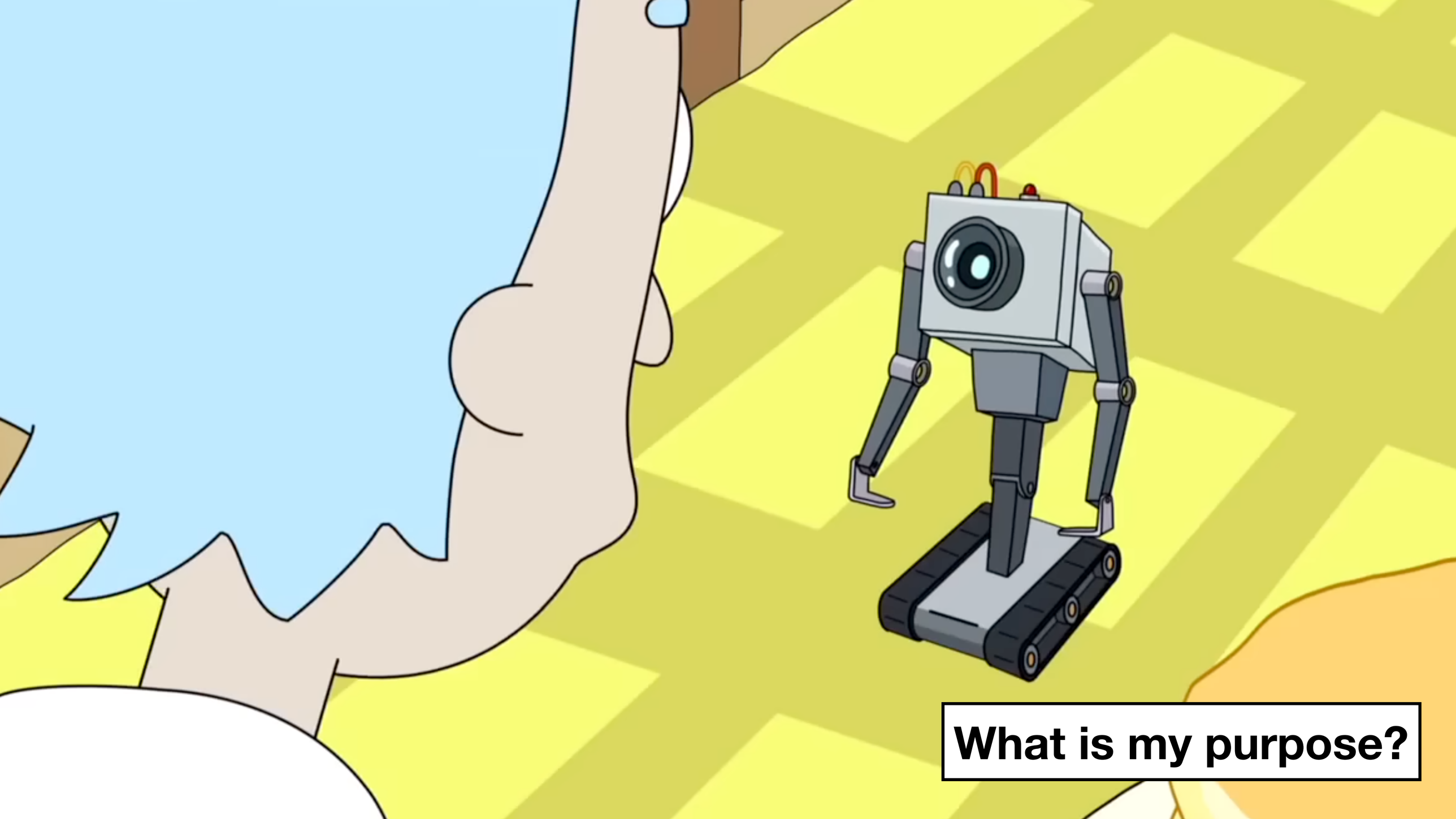
- n=10, LLM says: 10 ✅
- n=100, LLM says: 100 ✅
- n=1000, LLM says: 1000 ✅
- n=10,000, LLM says: 1000 ❌
- n=103, LLM says: 100 ❌

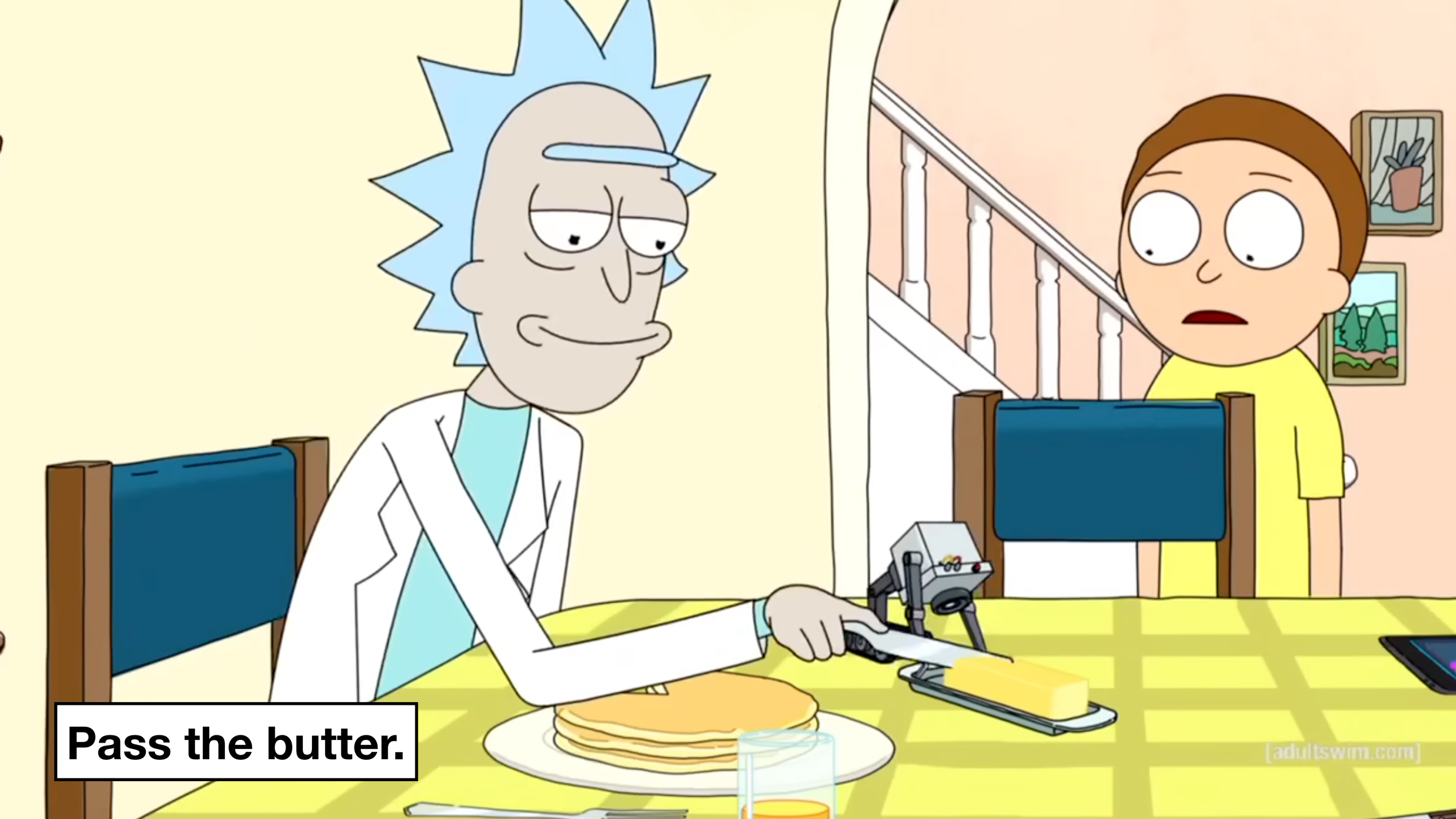
# What does responsible use of LLMs for data science look like?

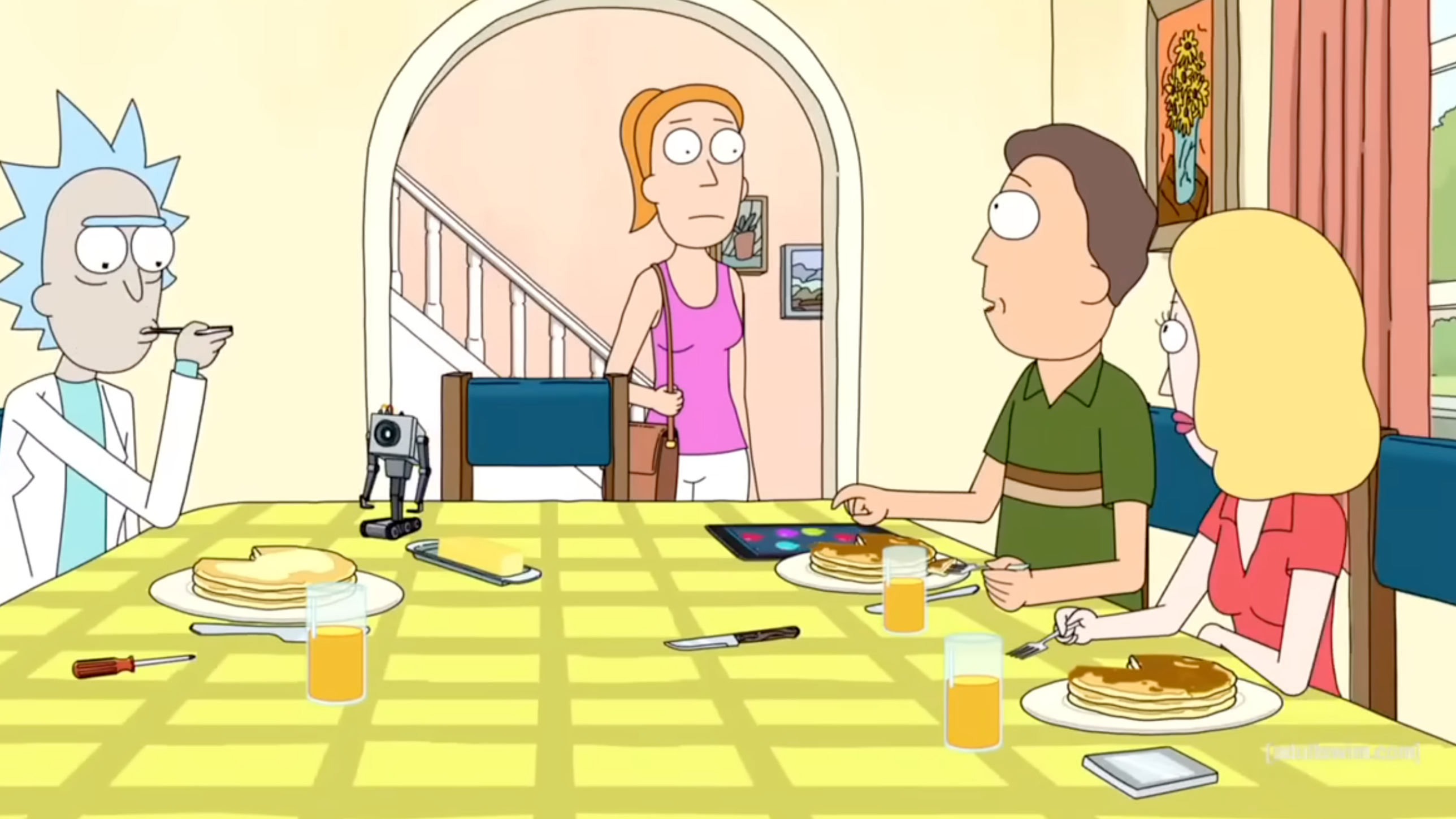
# Approach 1: Constrain

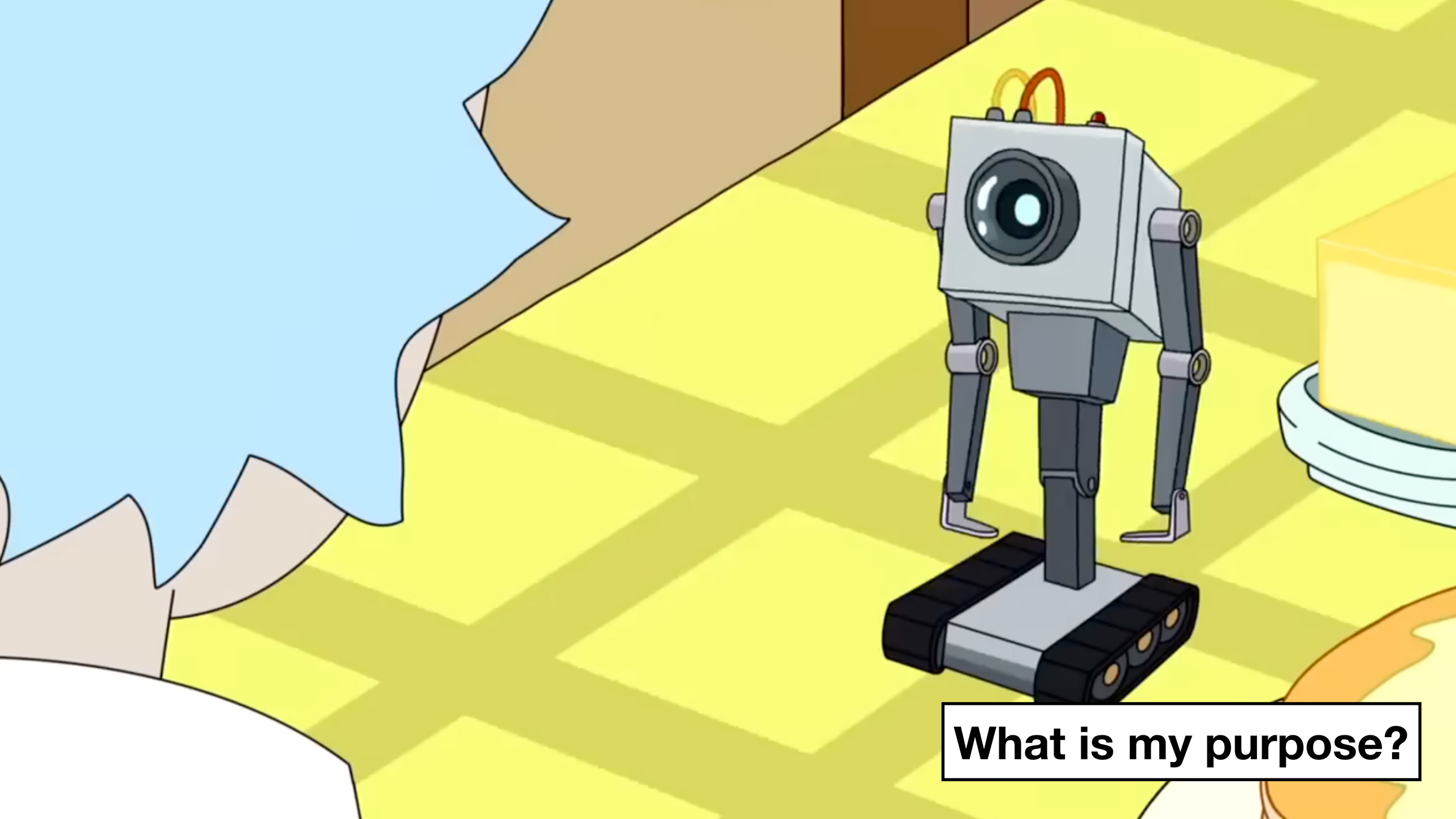
# Approach 1: Constrain

- Identify useful abilities that are firmly inside the LLM’s capability frontier
- Augment the LLM with (safe, deterministic) tools to increase its usefulness
- Instruct the LLM to stick to the prescribed task
- Resist the urge to feature-creep to the edge of the capability frontier
- Example: LLM -> SQL -> Dashboard

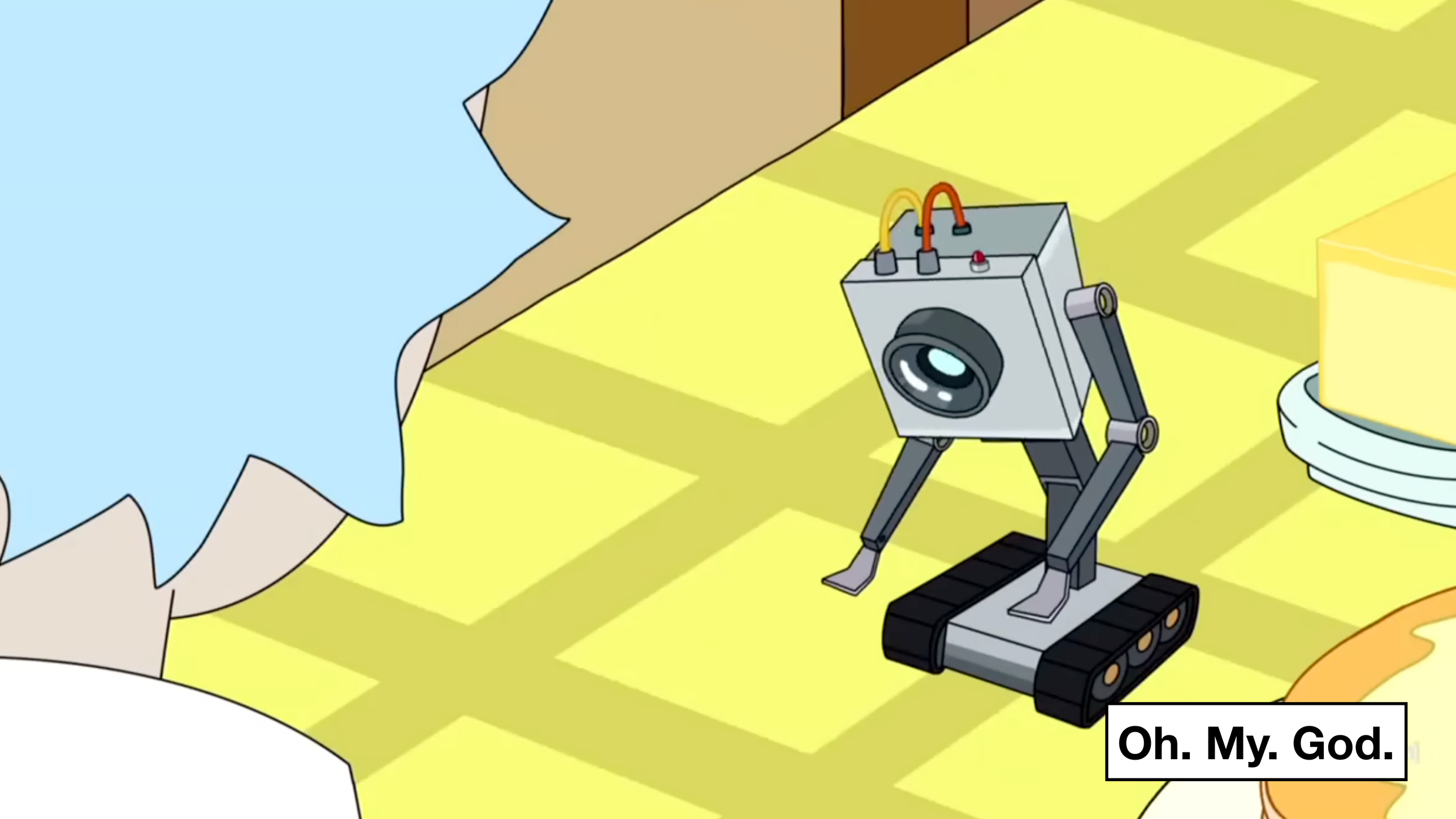

What is my purpose?


Pass the butter.
[adultswim.com]




What is my purpose?


You pass butter.


Oh. My. God.

# Is it responsible?

- **Correctness:** Only generates SQL, and does it quite well
- **Transparency:** Every SQL query is displayed to user
- **Reproducibility:** The SQL is reproducible

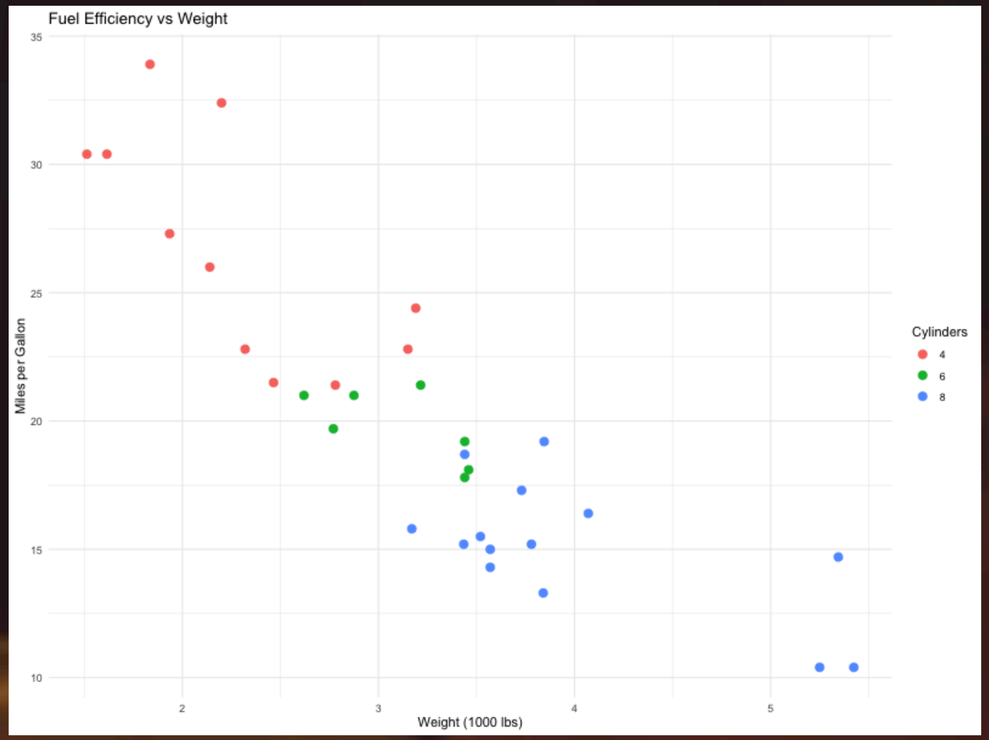
The “SQL chatbot applied to data dashboard” approach worked so well, we introduced an open-source package querychat to let anyone recreate the experience with their own data and visualizations

# Approach 2: Micromanage

# Approach 2: Micromanage

- Very tight human-AI feedback loop
- Outcomes that are pretty obviously right or wrong (or subjective)
- Human micromanages the AI so closely that mistakes are all but guaranteed to be caught
- Example: Plot tweaking tool


Let’s plot mtcars


Fuel Efficiency vs Weight
Miles per Gallon
35
30
25
20
15
10
2
3
4
5
Weight (1000 lbs)
Cylinders
4
6
8


The scatterplot points are a little small


The text annotations are all overlapping


The y axis needs to start at 0, obviously


Did you just use a diverging color palette for a categorical variable??


You’re absolutely right!

# Is it responsible?

- **Correctness:** Feels like it makes far fewer mistakes than a human does when fumbling through a visualization; mistakes are usually easy to catch
- **Transparency:** The user is directing, and can see the R code at all times
- **Reproducibility:** The R code is generally reproducible

Somehow feels far lower stakes. Helps that a lot of aspects of data viz are subjective.

# Approach 3: Deferred review

# Approach 3: Deferred review

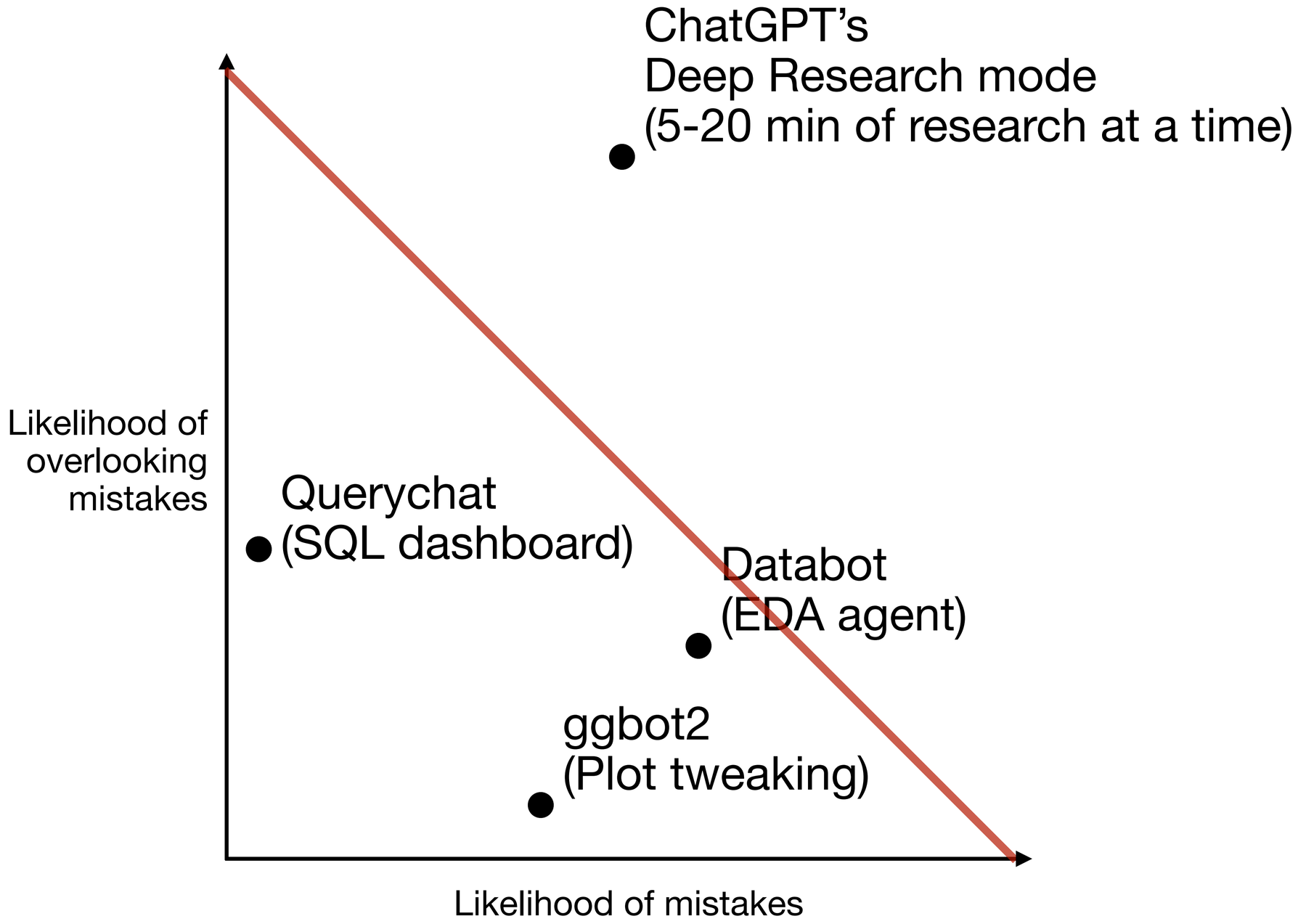
- Stay in the loop with the AI, but with a looser hand
- Be aware of what it's doing and why, but don't closely scrutinize its work for errors and hallucinations
- Enjoy fast progress/exploration, while piling up "review debt"
- Before "shipping" your work, stop and carefully review
- Akin to working on a git branch and getting a code review before merging
- Example: Databot

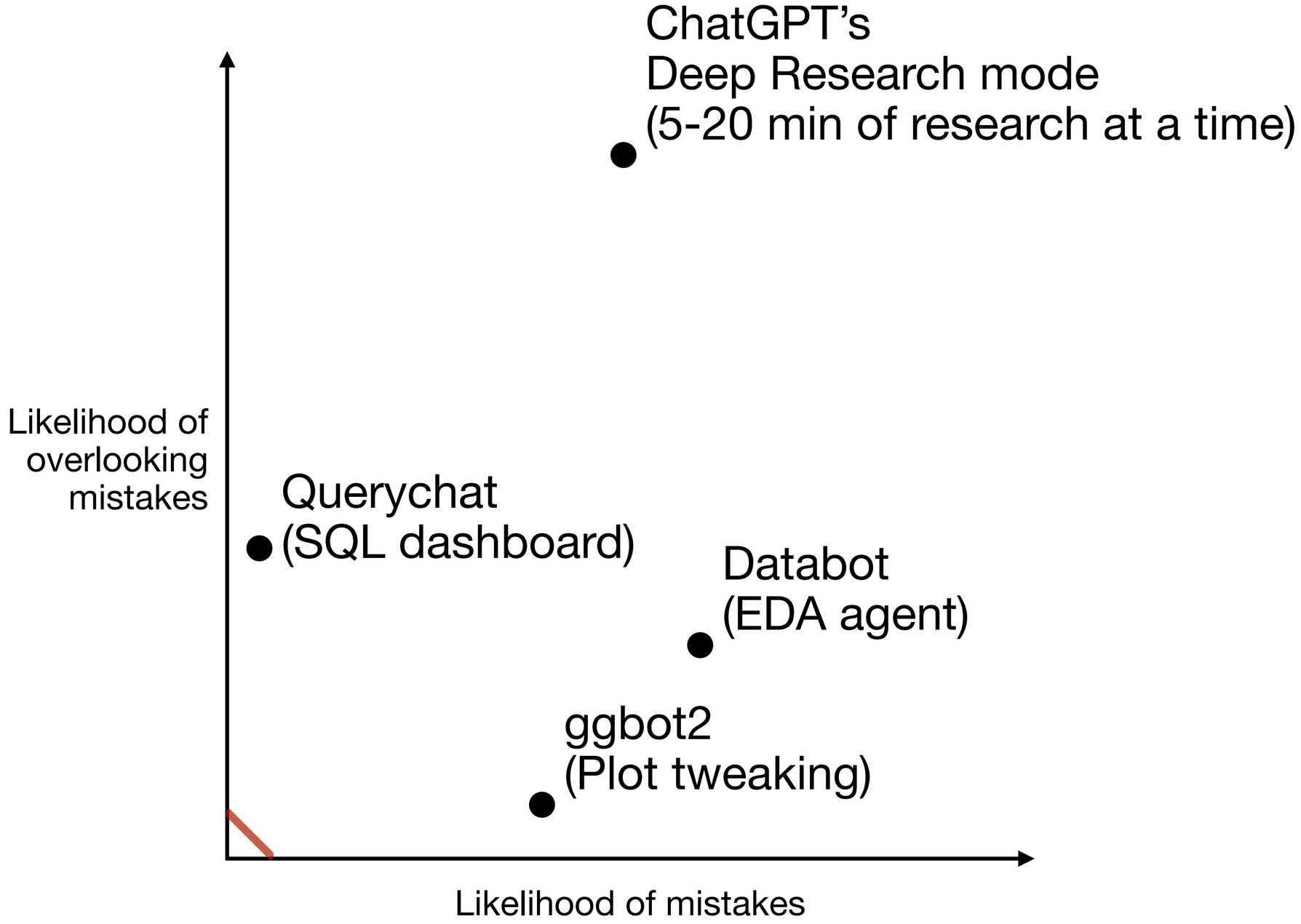
# Is it responsible?

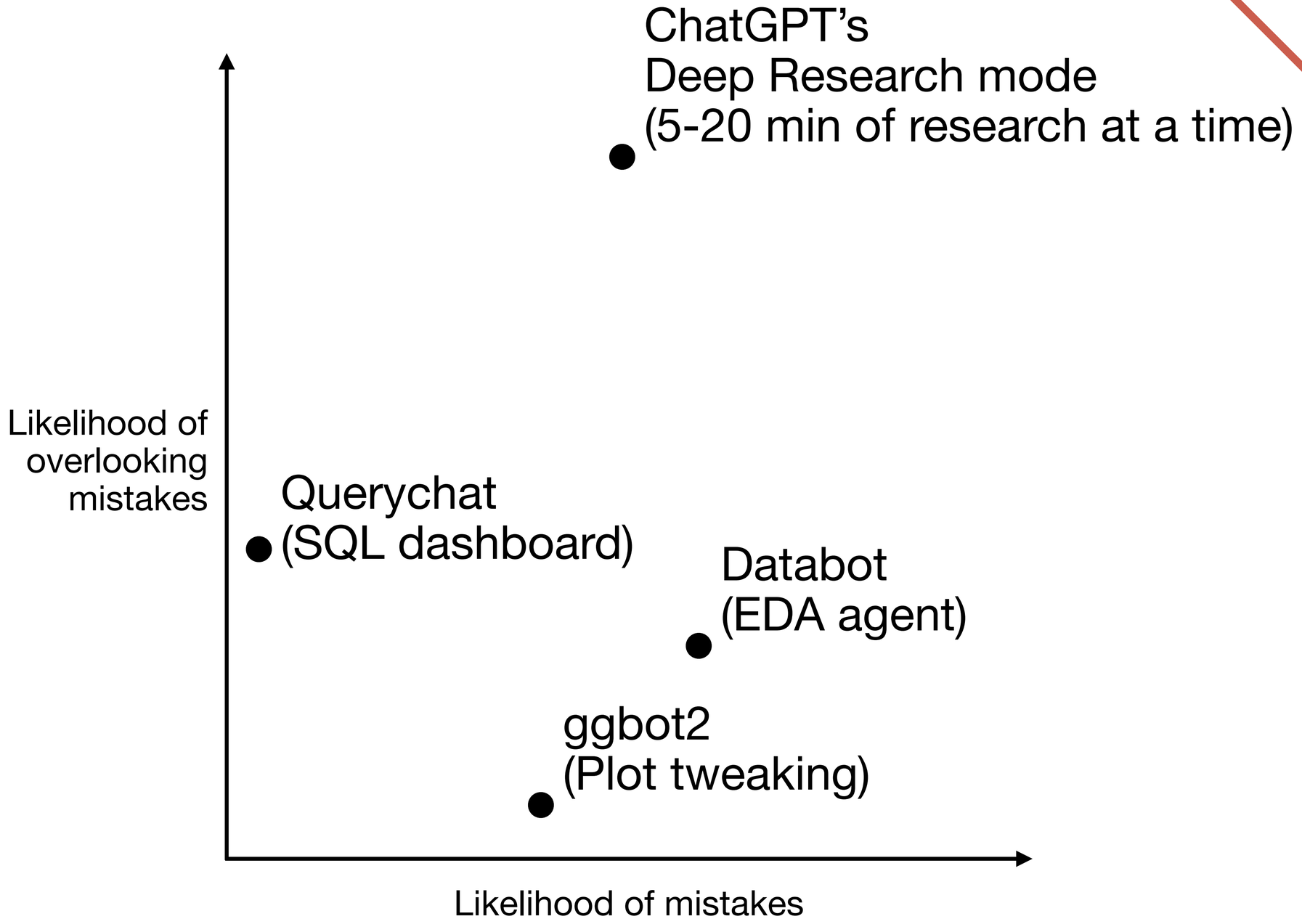
- **Correctness:** Relies on human discipline (to take the time to review) and expertise (to spot problems in the analysis); or, rapidly improving models
- **Transparency:** There's R code, but it goes by pretty fast
- **Reproducibility:** Databot will generate a reproducible report for you on demand

High risk of misuse. But so incredibly useful...

- Constrain: “You pass butter”
- Micromanage: “Not quite my tempo”
- Deferred review: “YOLO now, pay later”
- _____: _____


ChatGPT's
Deep Research mode
(5-20 min of research at a time)
Likelihood of overlooking mistakes
Querychat
(SQL dashboard)
Databot
(EDA agent)
ggbot2
(Plot tweaking)
Likelihood of mistakes


ChatGPT's
Deep Research mode
(5-20 min of research at a time)
Likelihood of overlooking mistakes
Querychat
(SQL dashboard)
Databot
(EDA agent)
ggbot2
(Plot tweaking)
Likelihood of mistakes


ChatGPT's
Deep Research mode
(5-20 min of research at a time)
Likelihood of overlooking mistakes
Querychat
(SQL dashboard)
Databot
(EDA agent)
ggbot2
(Plot tweaking)
Likelihood of mistakes

# Learn more

- YouTube: “Harnessing LLMs for Data Analysis”
- {ellmer}: Easily call LLMs from R
- {querychat}: Enhance Shiny data dashboards with LLMs that speak SQL
- Databot: Exploratory Data Analysis agent for Positron