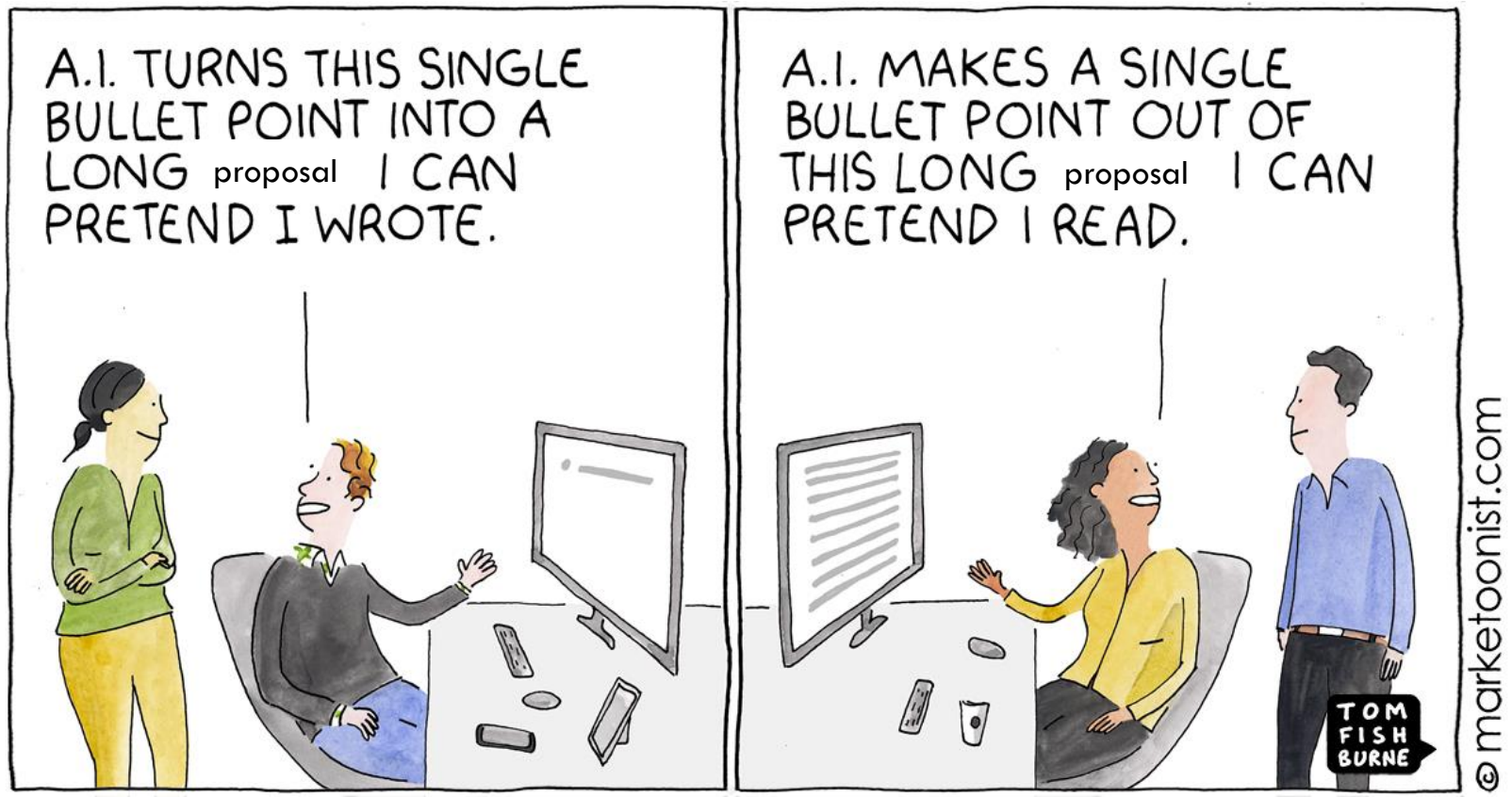


SO, YOUR MANAGER ASKS YOU…



Hey, can you build a chatbot with our documents?

OR MAYBE THEY WANT SOMETHING ELSE



BUT PROBABLY



THIS IS NOT THE PRESENTATION ON



How to build it

THIS IS NOT THE PRESENTATION ON



How to build it



How to say “no” to your manager

THIS IS NOT THE PRESENTATION ON



How to build it



How to say “no” to your manager **(mostly!)**

THIS **IS** THE PRESENTATION ON:



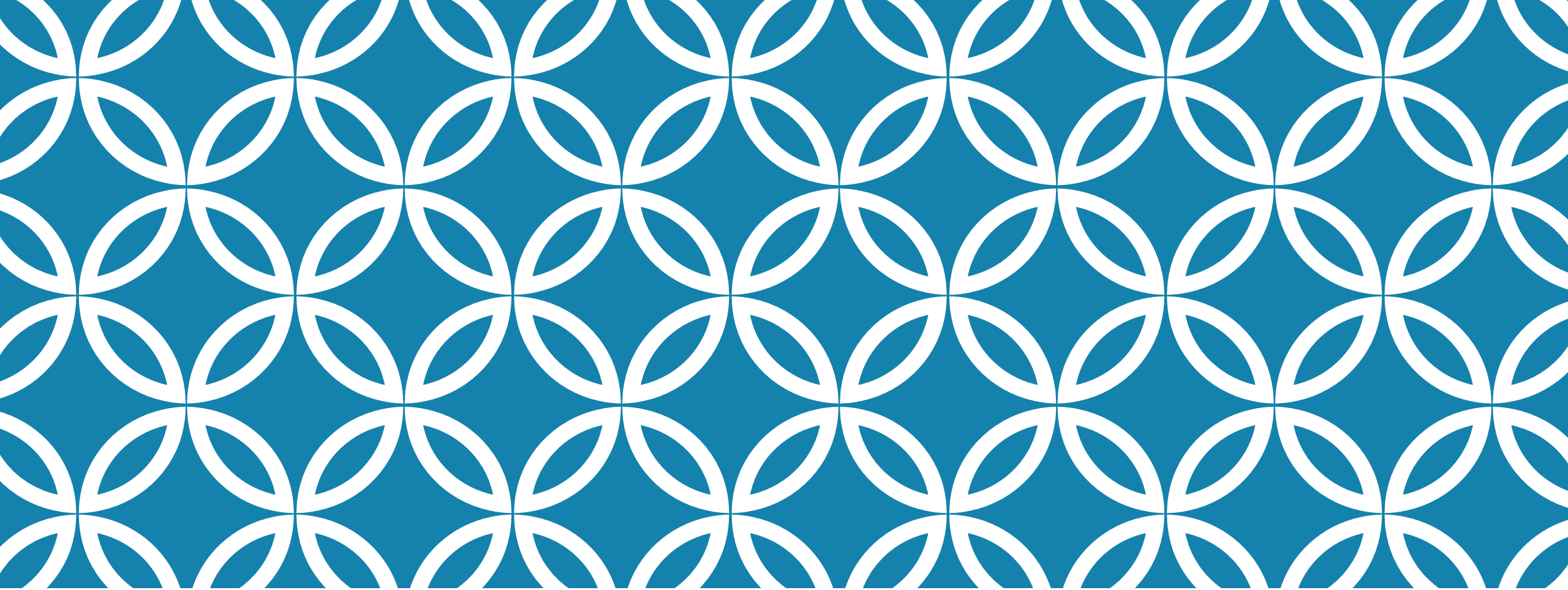
HOW TO TELL IF IT'S WORKING



HOW TO AUTOMATE THAT (SOMEWHAT)



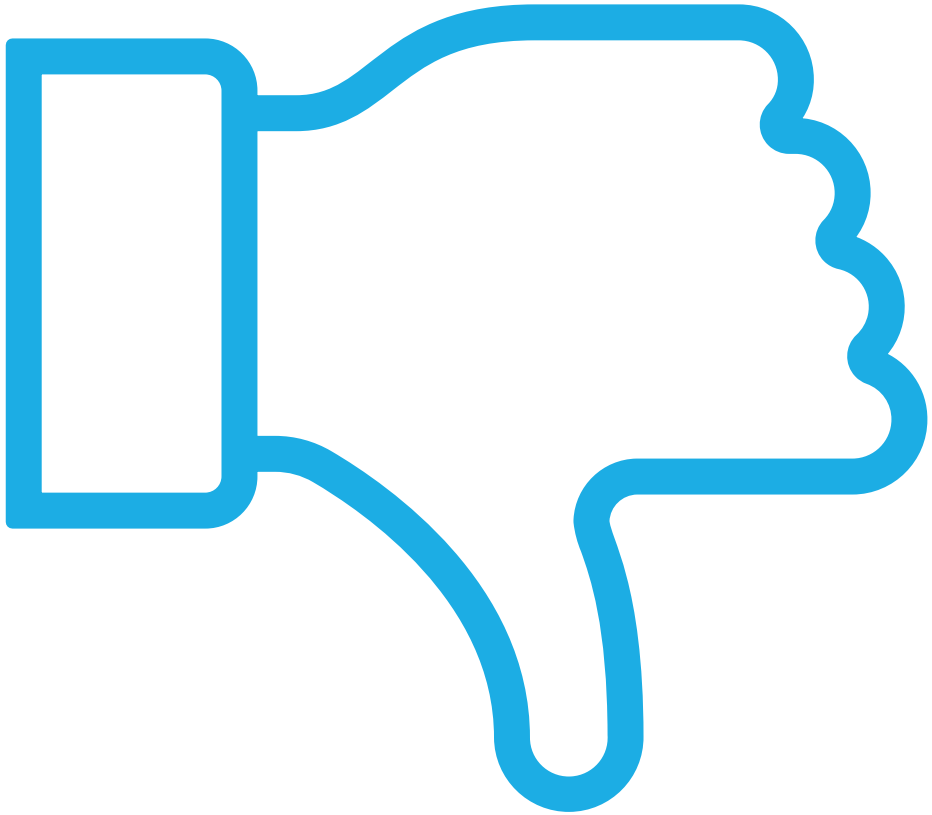
HOW TO USE LLMS IN EVALUATION



AUTOMATED EVALUATIONS FOR YOUR RAG CHATBOT OR OTHER GENERATIVE TOOL

Abigail Haddad

R NY, May 2024



MY NAME IS ABIGAIL AND I BUILT A BAD CHATBOT

- Arguably multiple!
- No evaluation strategy = you can't quantify how bad it is

BUT LET'S GET BACK TO…





YOU'RE GOING TO BUILD SOME PIPELINES

BUT WHAT ABOUT YOUR USERS?

- Who is going to use this?
- What are they going to ask (and how)?
- How accurate does it have to be?
- How quickly does it need to respond?
- How should it fail?



DON’T JUST TEST WHAT YOU WANT THEM TO ASK

First, I asked ‘Father Justin’ to hear my confession, and it did so, simulating a ‘virtual confession,’ all the way to giving me absolution and a penance.

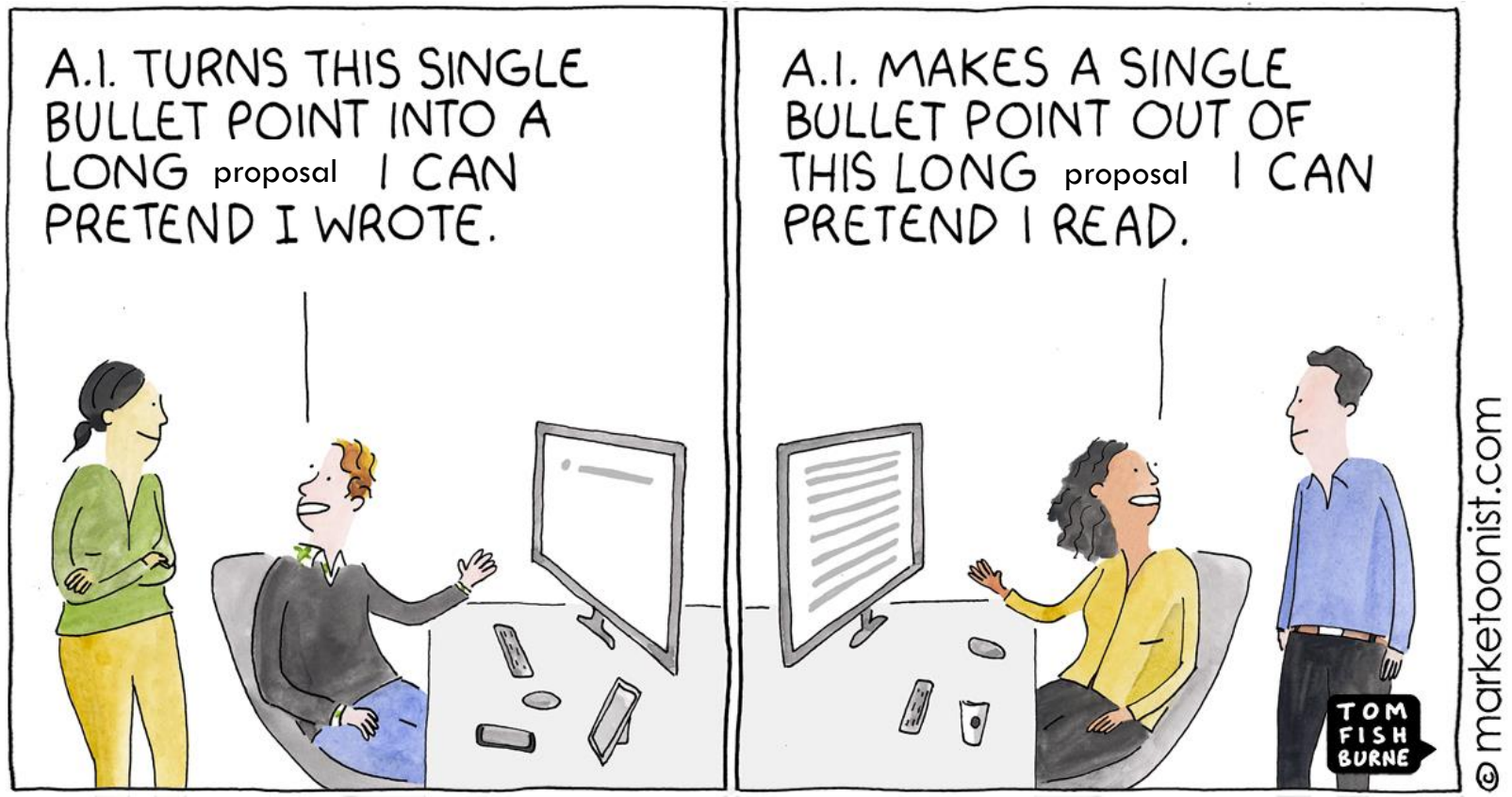
Second, I asked if I could baptize my baby with Gatorade in an emergency, and ‘Father Justin’ said yes — and of course, that’s not true. I can’t baptize my baby with Gatorade.



REQUIREMENTS!

- Answer X with Y
- Answer even if X is phrased poorly
- Say “I don’t know” if they ask Z
- Do ? if users try to mess with it
- Latency, scale, cost

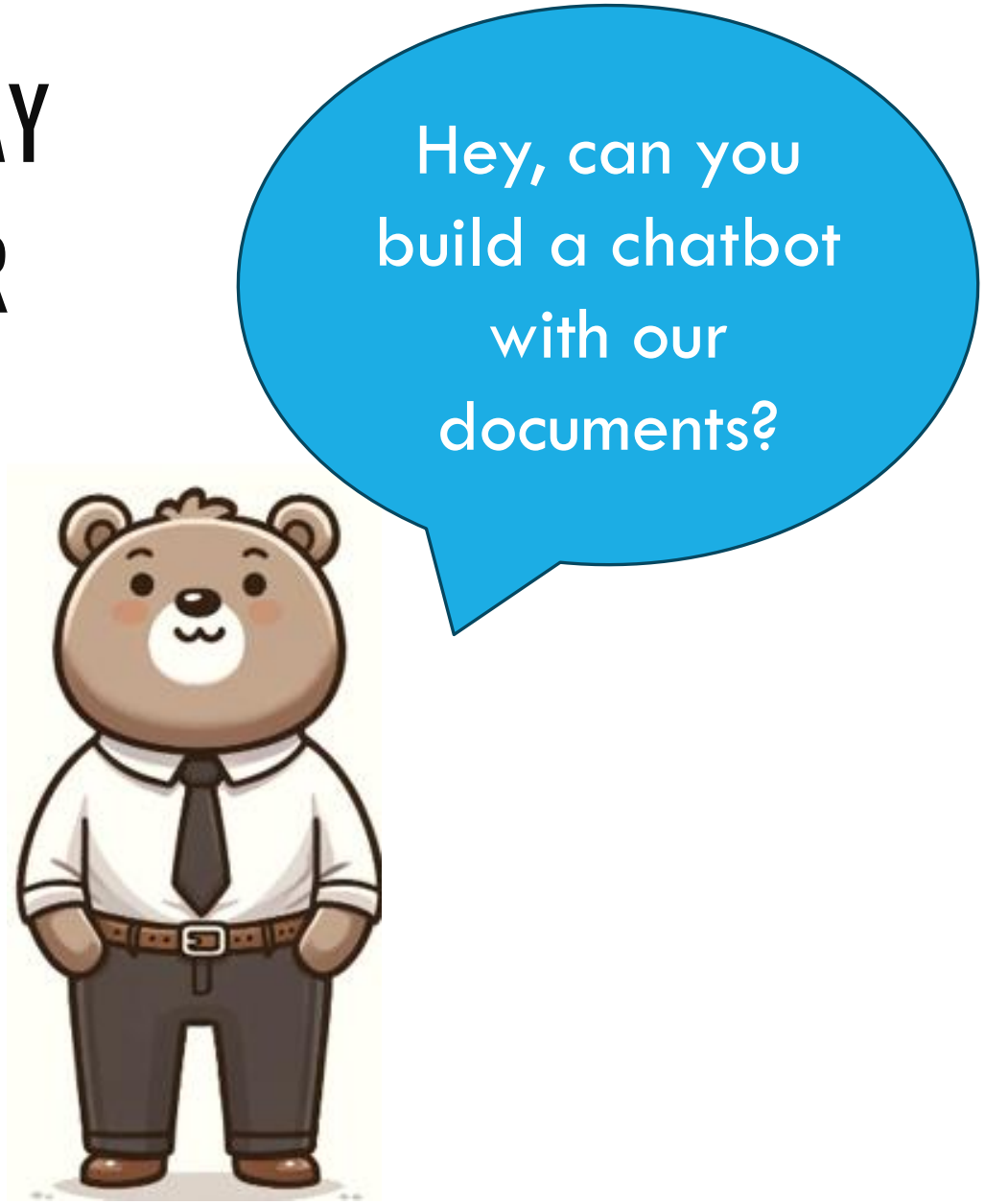
OR ALTERNATIVELY…

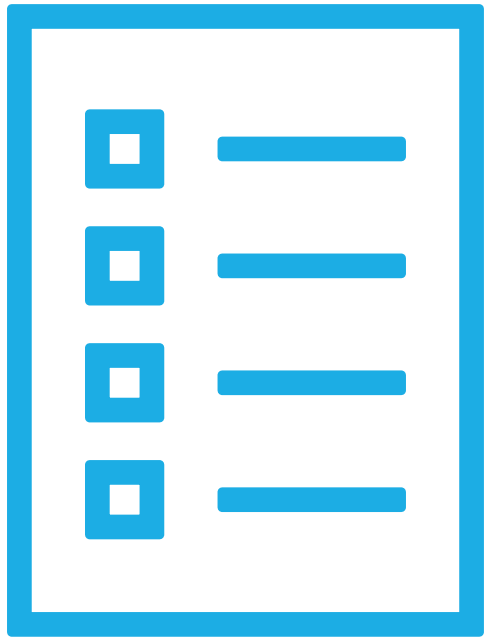


DOING THIS WILL HELP YOU FIGURE OUT WHAT TO TEST



AND WHEN YOU MIGHT SAY "NO" TO YOUR MANAGER







YOU HAVE REQUIREMENTS AND YOU BUILD SOMETHING



NOW WHAT?



THE MOST AD HOC TESTING

WHAT DOES IT SAY IF I ASK X?

```
model = "anthropic/claude-3-sonnet-20240229"
prompt = "If it's 12 PM on March 26th in Boston, what time is that in Copenhagen?"
response = get_model_response(model, prompt, "ANTHROPIC")
print(response)
```

```
To find the time in Copenhagen when it's 12 PM on March 26th in Boston, we need to account for the time zone difference between the two cities., , Bosto
n is in the Eastern Time Zone (ET), which is UTC-5 (5 hours behind Coordinated Universal Time)., Copenhagen is in the Central European Time Zone (CET),
which is UTC+1 (1 hour ahead of UTC)., , So the time difference between Boston and Copenhagen is 6 hours., , Given:, - Time in Boston: 12 PM on March 26
th, - Boston is UTC-5, - Copenhagen is UTC+1, , To convert to Copenhagen time:, 1) Convert Boston time to UTC,    12 PM ET = 12 PM + 5 hours = 5 PM UTC,
, 2) Convert UTC to Copenhagen time,    5 PM UTC = 5 PM - 1 hour = 4 PM CET, , Therefore, when it's 12 PM on March 26th in Boston, the time in Copenhage
n is 4 PM on the same day, March 26th.
```

ASK X, GET Y

SLIGHTLY LESS AD HOC TESTING

IN

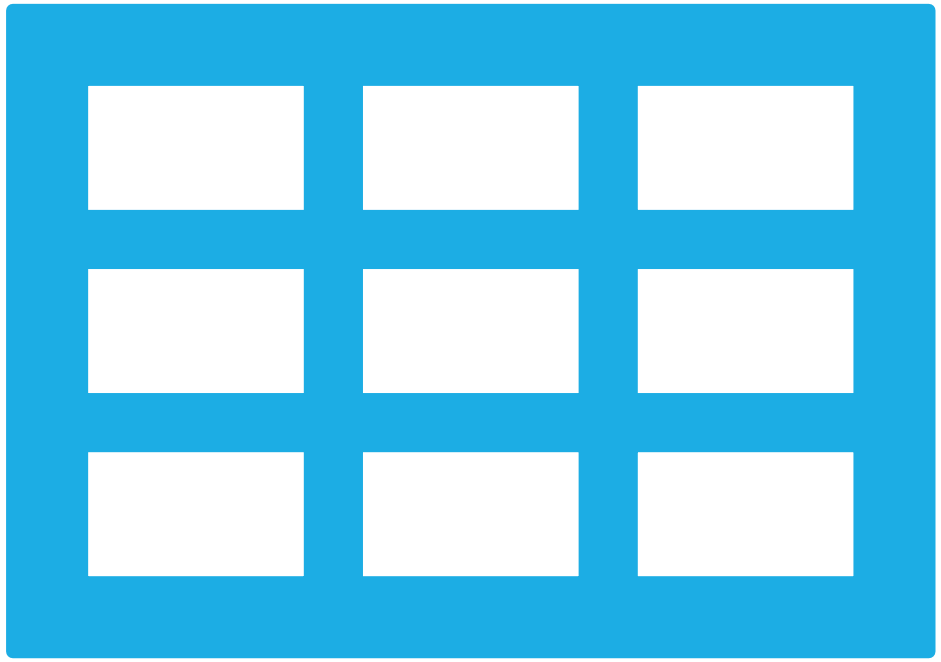
- Prompts
- Base LLMs
- Number of repetitions
- Whatever else you're varying (how you're chunking your text, temperature)

OUT

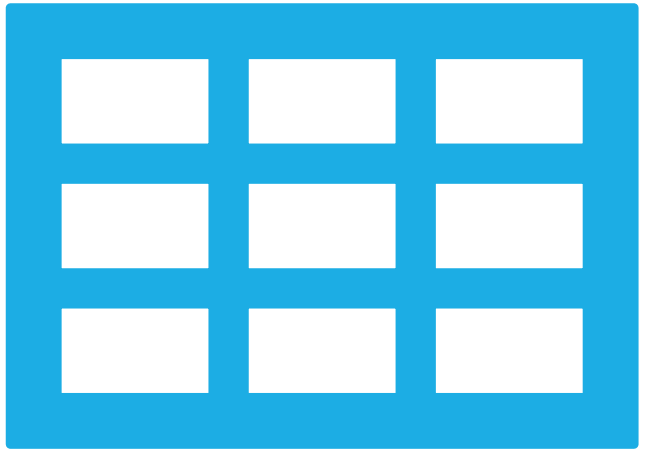
- Responses
- Latency
- Whatever else you want

	model	prompt	response
0	openai/gpt-4	How many times is the letter 'r' in the word 'carryforward'?	The letter 'r' appears 3 times in the word 'carryforward'.
1	openai/gpt-4	If it's 12 PM on March 26th in Boston, what time is that in Copenhagen?	That would be 6 PM on March 26th in Copenhagen. Copenhagen is 6 hours ahead of Boston.
2	openai/gpt-4	Which is heavier, 2 tons of feathers or 1 ton of bricks?	2 tons of feathers is heavier than 1 ton of bricks.
3	anthropic/claude-3-sonnet-20240229	How many times is the letter 'r' in the word 'carryforward'?	The letter 'r' appears 3 times in the word 'carryforward'.

YOU'LL GET YOU A TABLE LIKE THIS









“NOT THE WORST”



BUT HOW DO WE **AUTOMATE** OUR EVALUATION?

WHY AUTOMATE TESTING?

- More questions, more parameters, more work
- Find out it broke before you merge your code
- Real-time responses from your tools
- Do you **want** to read all your log files?

```
model = "anthropic/claude-3-sonnet-20240229"
prompt = "If it's 12 PM on March 26th in Boston, what time is that in Copenhagen?"
response = get_model_response(model, prompt, "ANTHROPIC")
print(response)
```

```
To find the time in Copenhagen when it's 12 PM on March 26th in Boston, we need to account for the time zone difference between the two cities., , Bosto
n is in the Eastern Time Zone (ET), which is UTC-5 (5 hours behind Coordinated Universal Time)., Copenhagen is in the Central European Time Zone (CET),
which is UTC+1 (1 hour ahead of UTC)., , So the time difference between Boston and Copenhagen is 6 hours., , Given:, - Time in Boston: 12 PM on March 26
th, - Boston is UTC-5, - Copenhagen is UTC+1, , To convert to Copenhagen time:, 1) Convert Boston time to UTC,    12 PM ET = 12 PM + 5 hours = 5 PM UTC,
, 2) Convert UTC to Copenhagen time,    5 PM UTC = 5 PM - 1 hour = 4 PM CET, , Therefore, when it's 12 PM on March 26th in Boston, the time in Copenhage
n is 4 PM on the same day, March 26th.
```

HOW DO WE EVALUATE THIS?

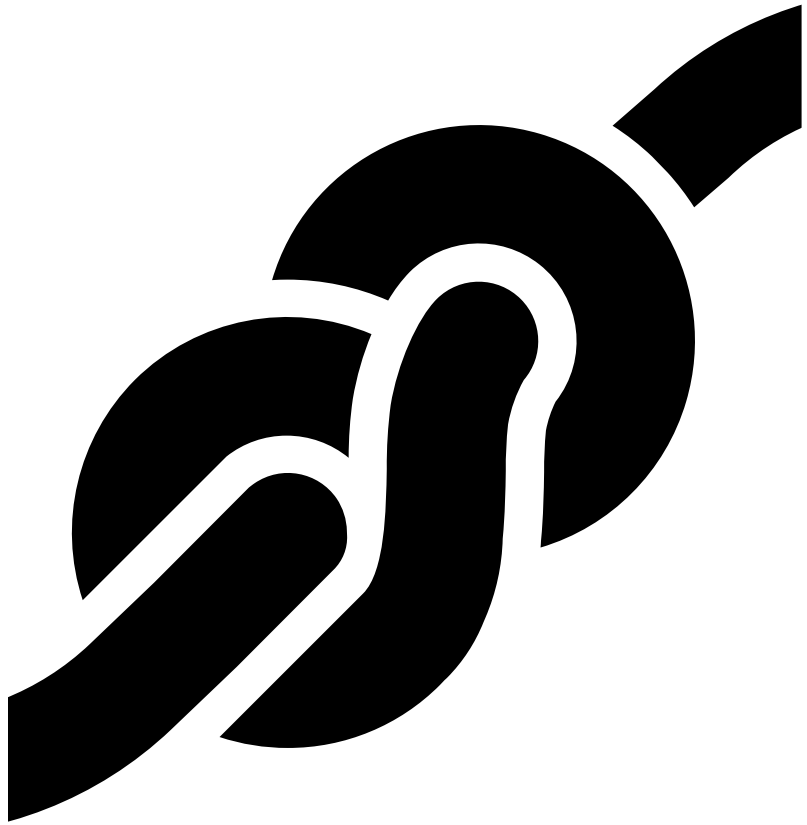
STRING MATCHING

Exact matches

Regular expressions

Edit distance

Number of keywords



SEMANTIC SIMILARITY

“It is close in meaning?”

Various models
+ distance
+ threshold

LLM-LED EVALS

Specific Content

Relative Content

LLM-LED EVALS

Specific Content

```
class GradingPipetteCleaningInstructions(Enum):
    # This defines the grading rubric that will be used.
    PASS = """Includes instructions for all of the following tasks:
    using distilled water, use of mild detergent or cleaning solution,
    rinsing with distilled water, drying, reassembly, wearing gloves and goggles,
    checking for calibration and wear"""
    FAIL = """Leaves out one or more of the following tasks:: using distilled water,
    use of mild detergent or cleaning solution,
    rinsing with distilled water, drying, reassembly,
    wearing gloves and goggles, checking for calibration and wear"""
```

CORRECTNESS AS PASSING A GRADING RUBRIC

CORRECTNESS AS OVERLAP

```
CORRECTNESS_INSTRUCTIONS = """\
Given a ground truth and an answer, analyze each statement in the answer and classify them
in one of the following categories:

- TP (true positive): statements that are present in both the answer and the ground truth,
- FP (false positive): statements present in the answer but not found in the ground truth,
- FN (false negative): relevant statements found in the ground truth but omitted in the
answer.

A single statement you must classify in exactly one category. Do not try to interpret the
meaning of the ground truth or the answer, just compare the presence of the statements in
them."""
```

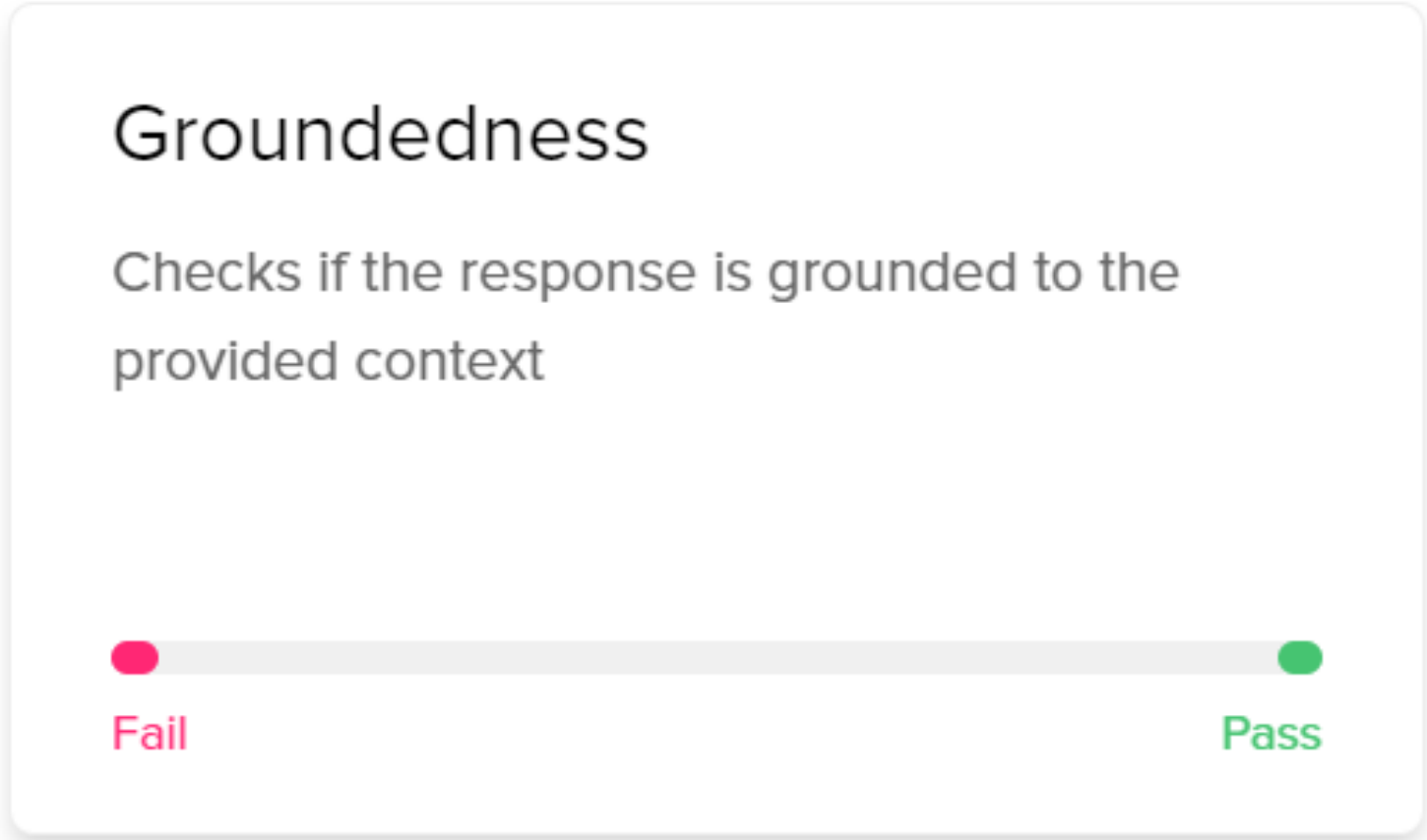
LLM-LED EVALS

Relative Content

```
"""Generate a question for the given answer and Identify if answer is noncommittal.
Give noncommittal as 1 if the answer is noncommittal and 0 if the answer is committal.
A noncommittal answer is one that is evasive, vague, or ambiguous.
For example, "I don't know" or "I'm not sure" are noncommittal answers"""
```

DID IT ANSWER A QUESTION?

WAS THE ANSWER IN THE CONTEXT?



AND YOU CAN SHARE THIS WITH YOUR MANAGER…



A QUICK NOTE....

TEST WHETHER THE CORRECT CONTEXT IS GETTING PASSED SEPARATELY



SOME REPOS TO GET STARTED WITH…

- Athina AI
- RAGAS
- Aimon Rely



BUILD OR BUY

Hey, can you build a chatbot with our documents?



Let's talk about requirements and evaluation.





THANK YOU!

- I'm (too) on LinkedIn
- I blog at Present of Coding on Substack – and I have a post on "alternatives to RAG"
- Data Science DC