Are LLMs good anomaly detectors?

And what are the alternatives?

Chloé Caron

This is me

… going to keep it short

- Tech lead, Developer & Data Engineer @ Theodo UK

- **Fun fact**: I lived in 6 countries and moved 8 times before turning 18

Why data anomalies?

12%

Average revenue loss by U.S. companies due to bad data

Why is data quality so important?

- Important for decision making
- Affects trust
- Security
- Etc.

Bad quality has a long list of causes

Missing Data

Incorrect Data

Outdated Data

Inconsistent format/standards

Incompatible systems

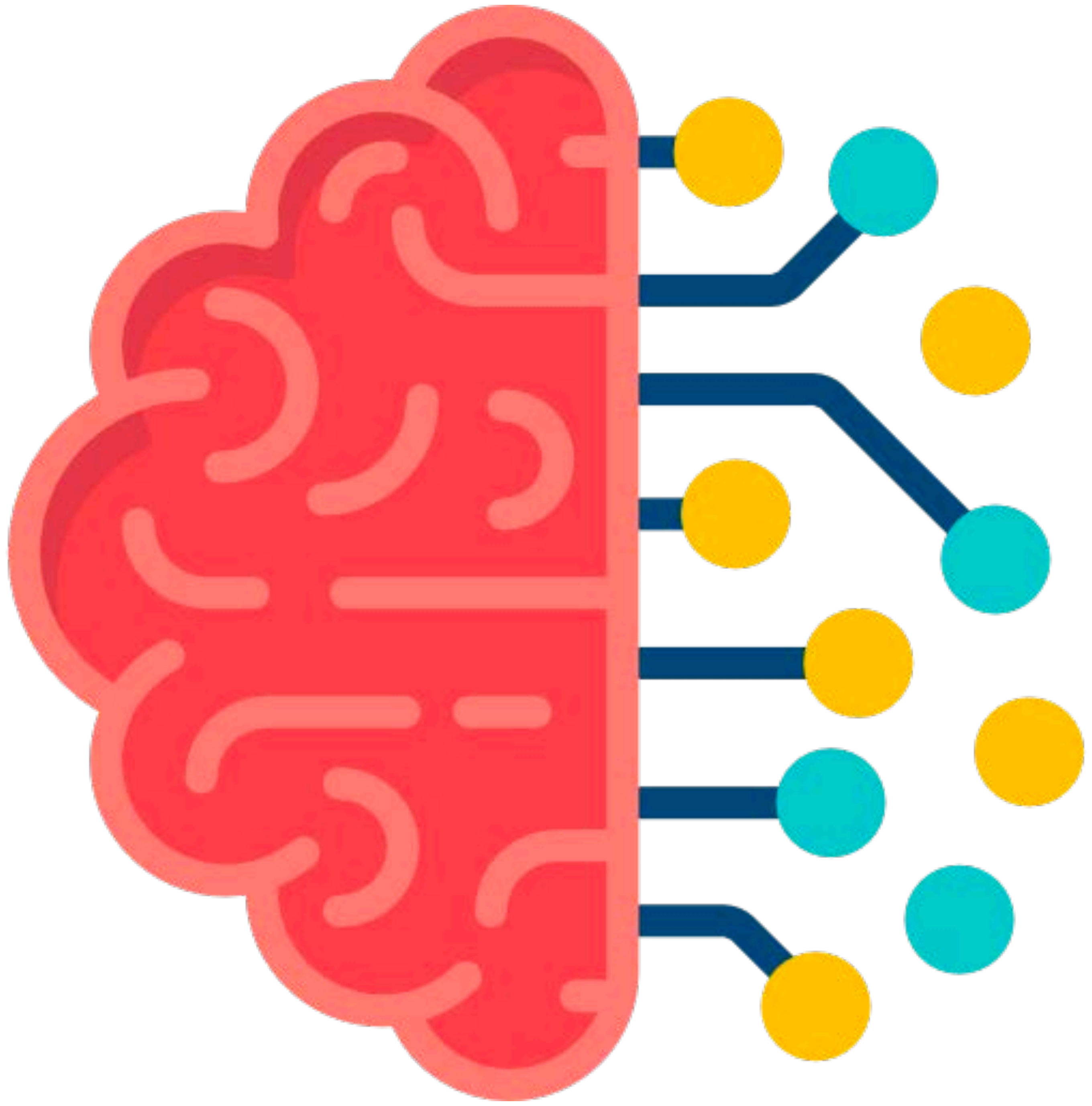
Data complexity

What can we do about it?

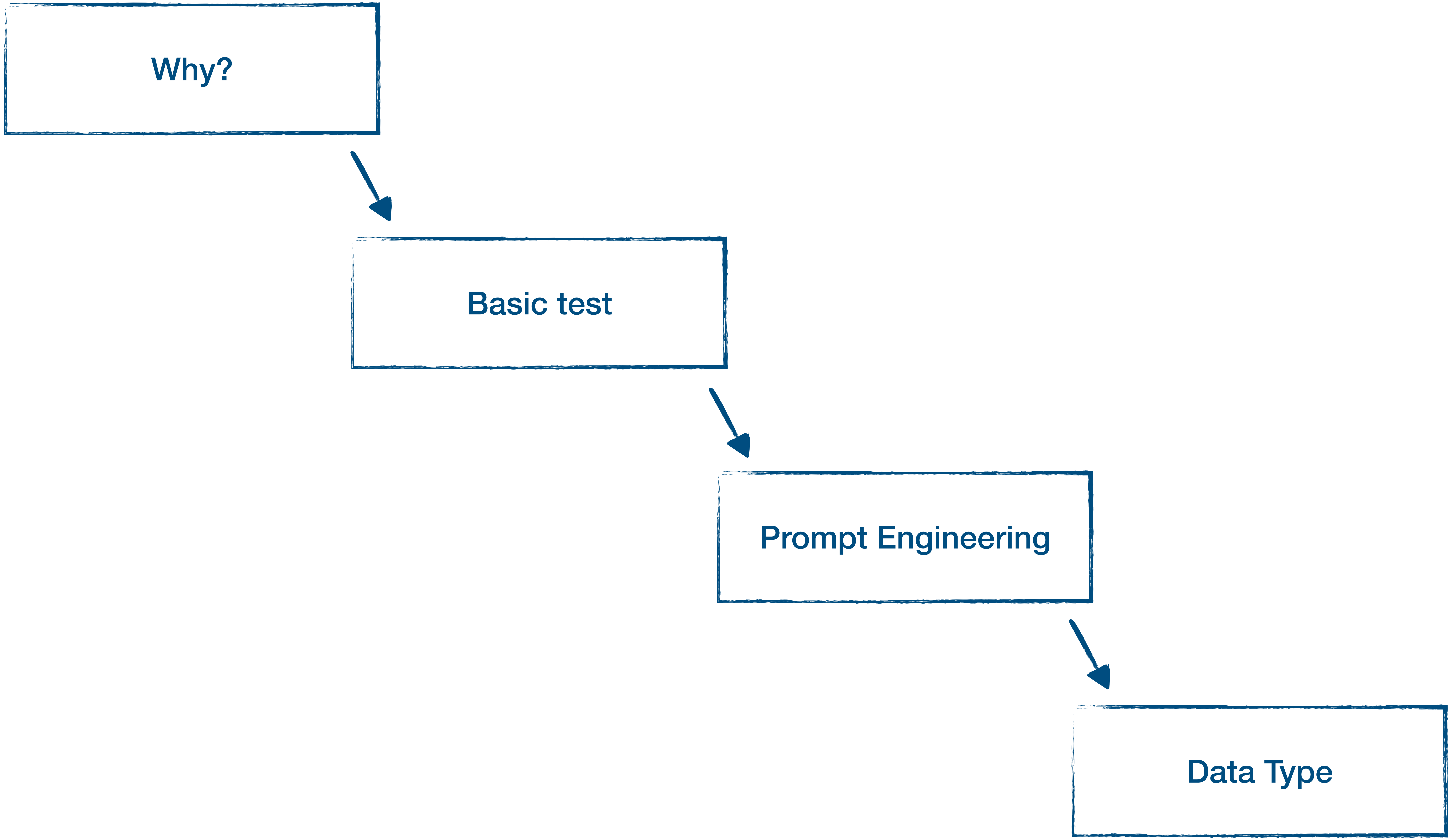
Could use some existing tools

Great for data observability

… or we could spice things up



How good is OpenAI with anomalies?

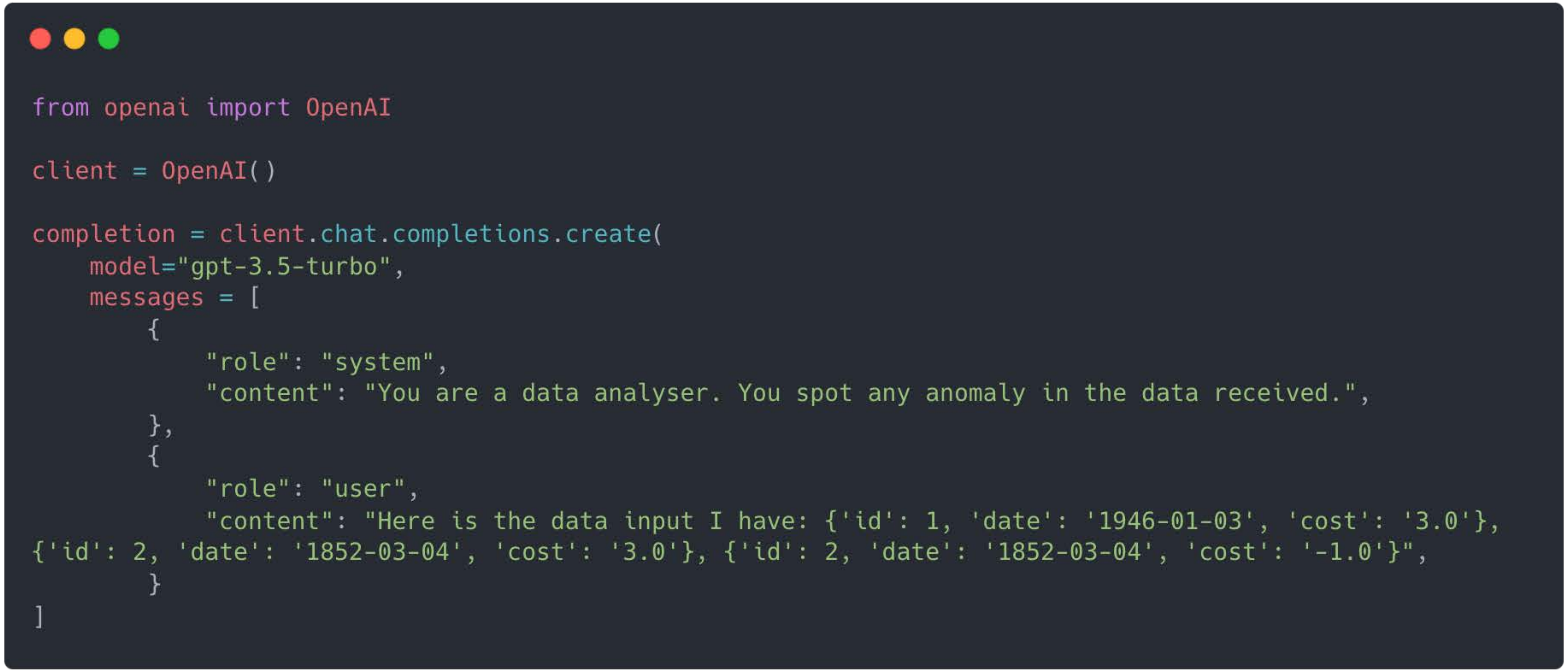

Why?
Basic test
Prompt Engineering
Data Type

Why?

Flexibility

Curiosity

Basic test

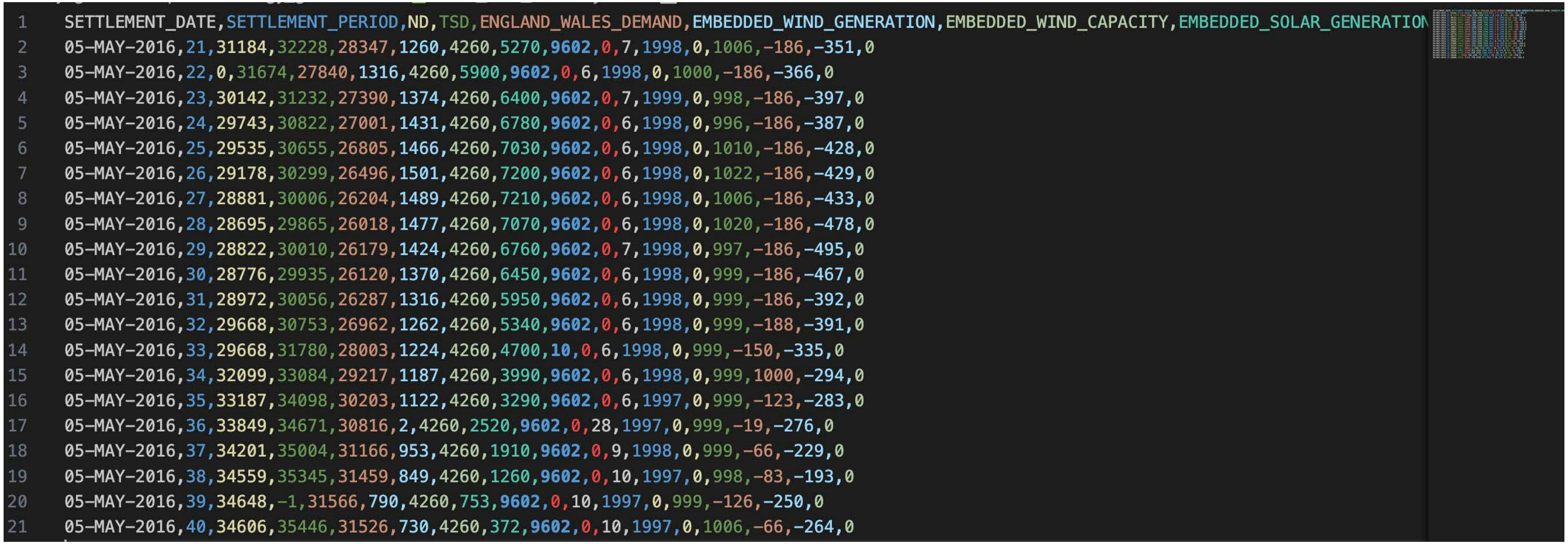


```
from openai import OpenAI

client = OpenAI()

completion = client.chat.completions.create(
    model="gpt-3.5-turbo",
    messages = [
        {
            "role": "system",
            "content": "You are a data analyser. You spot any anomaly in the data received.",
        },
        {
            "role": "user",
            "content": "Here is the data input I have: {'id': 1, 'date': '1946-01-03', 'cost': '3.0'},
{'id': 2, 'date': '1852-03-04', 'cost': '3.0'}, {'id': 2, 'date': '1852-03-04', 'cost': '-1.0'}",
        }
]
```

Basic test

Electricity consumption UK 2009-2023

Historic electricity consumption in the UK (National Grid) between 2009 and 2023



```
SETTLEMENT_DATE,SETTLEMENT_PERIOD,ND,TSD,ENGLAND_WALES_DEMAND,EMBEDDED_WIND_GENERATION,EMBEDDED_WIND_CAPACITY,EMBEDDED_SOLAR_GENERATION
05-MAY-2016,21,31184,32228,28347,1260,4260,5270,9602,0,7,1998,0,1006,-186,-351,0
05-MAY-2016,22,0,31674,27840,1316,4260,5900,9602,0,6,1998,0,1000,-186,-366,0
05-MAY-2016,23,30142,31232,27390,1374,4260,6400,9602,0,7,1999,0,998,-186,-397,0
05-MAY-2016,24,29743,30822,27001,1431,4260,6780,9602,0,6,1998,0,996,-186,-387,0
05-MAY-2016,25,29535,30655,26805,1466,4260,7030,9602,0,6,1998,0,1010,-186,-428,0
05-MAY-2016,26,29178,30299,26496,1501,4260,7200,9602,0,6,1998,0,1022,-186,-429,0
05-MAY-2016,27,28881,30006,26204,1489,4260,7210,9602,0,6,1998,0,1006,-186,-433,0
05-MAY-2016,28,28695,29865,26018,1477,4260,7070,9602,0,6,1998,0,1020,-186,-478,0
05-MAY-2016,29,28822,30010,26179,1424,4260,6760,9602,0,7,1998,0,997,-186,-495,0
05-MAY-2016,30,28776,29935,26120,1370,4260,6450,9602,0,6,1998,0,999,-186,-467,0
05-MAY-2016,31,28972,30056,26287,1316,4260,5950,9602,0,6,1998,0,999,-186,-392,0
05-MAY-2016,32,29668,30753,26962,1262,4260,5340,9602,0,6,1998,0,999,-188,-391,0
05-MAY-2016,33,29668,31780,28003,1224,4260,4700,10,0,6,1998,0,999,-150,-335,0
05-MAY-2016,34,32099,33084,29217,1187,4260,3990,9602,0,6,1998,0,999,1000,-294,0
05-MAY-2016,35,33187,34098,30203,1122,4260,3290,9602,0,6,1997,0,999,-123,-283,0
05-MAY-2016,36,33849,34671,30816,2,4260,2520,9602,0,28,1997,0,999,-19,-276,0
05-MAY-2016,37,34201,35004,31166,953,4260,1910,9602,0,9,1998,0,999,-66,-229,0
05-MAY-2016,38,34559,35345,31459,849,4260,1260,9602,0,10,1997,0,998,-83,-193,0
05-MAY-2016,39,34648,-1,31566,790,4260,753,9602,0,10,1997,0,999,-126,-250,0
05-MAY-2016,40,34606,35446,31526,730,4260,372,9602,0,10,1997,0,1006,-66,-264,0
```

Basic test

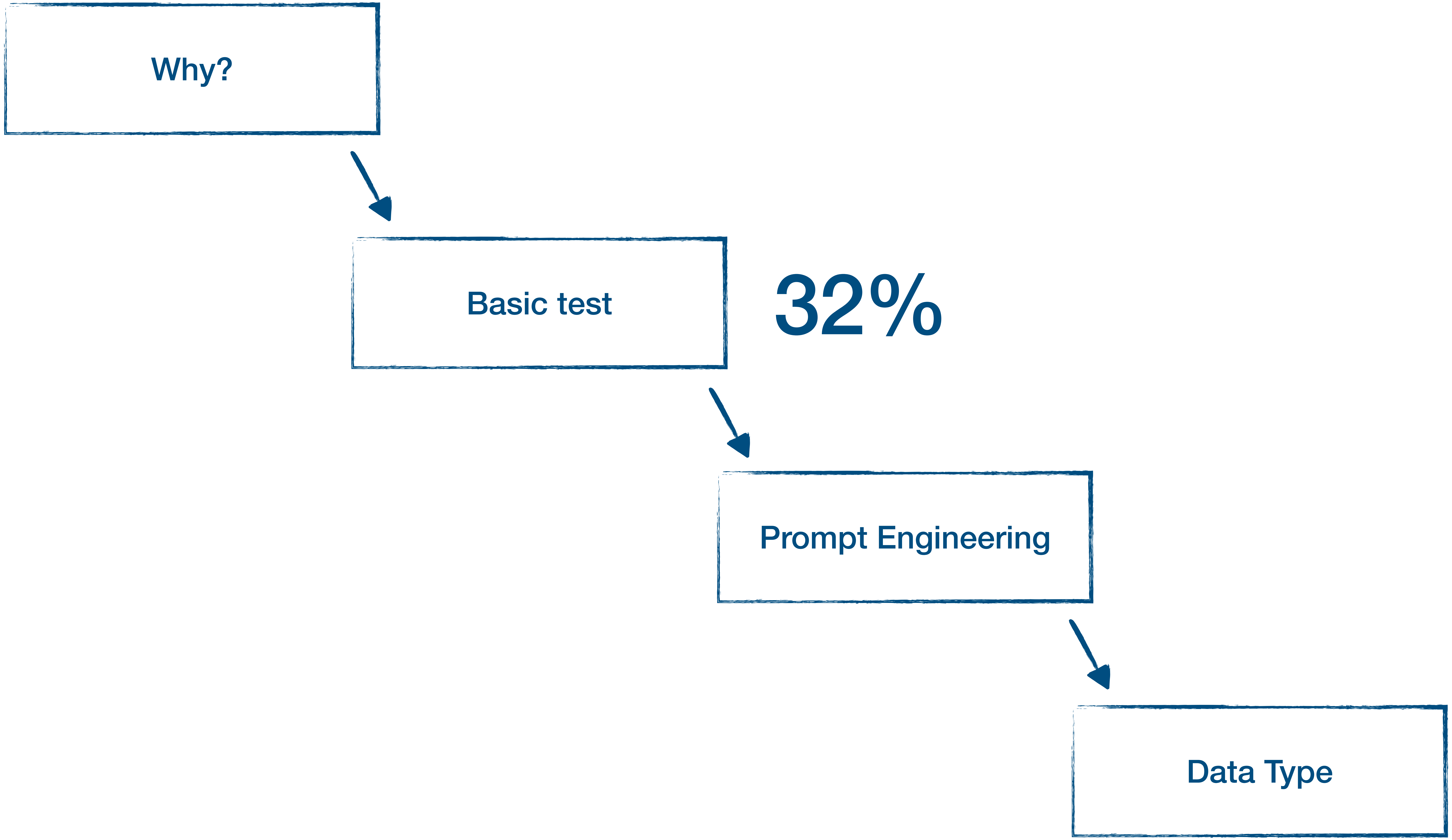
- Most results had **no anomalies found**
- For the rest:
  - **GPT 4 performed better** than GPT 3.5
  - More anomalies -> more difficult to find them
  - Number of lines of test data didn’t have a significant impact

Basic test

32%

Of intended anomalies detected for GPT 4

With 2 anomalies and 20 lines of data



Why?

Basic test

32%

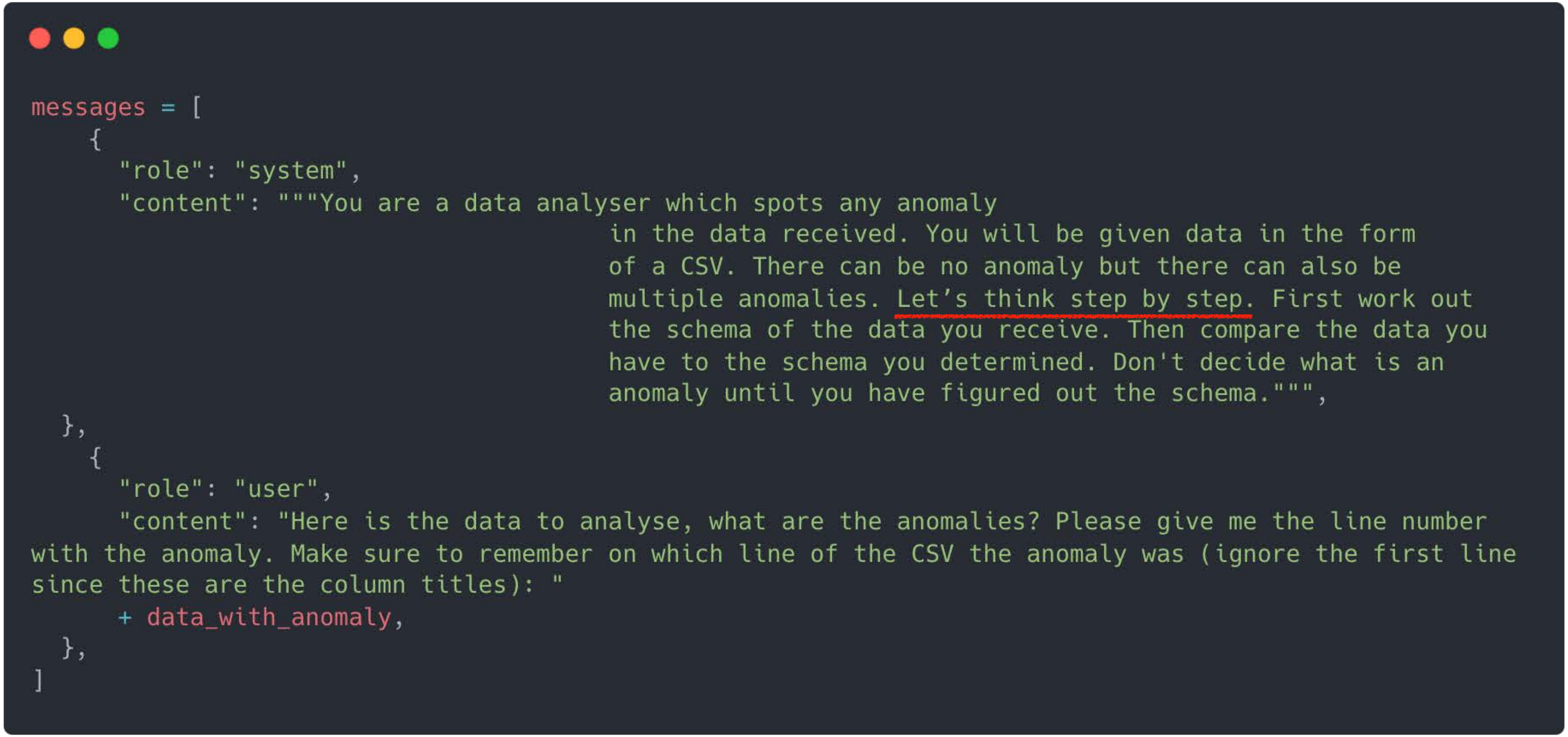
Prompt Engineering

Data Type

Chain of Thought

Prompt Engineering

Chain of Thought



```
messages = [
    {
      "role": "system",
      "content": """You are a data analyser which spots any anomaly
                                        in the data received. You will be given data in the form
                                        of a CSV. There can be no anomaly but there can also be
                                        multiple anomalies. Let’s think step by step. First work out
                                        the schema of the data you receive. Then compare the data you
                                        have to the schema you determined. Don't decide what is an
                                        anomaly until you have figured out the schema.""",
  },
    {
      "role": "user",
      "content": "Here is the data to analyse, what are the anomalies? Please give me the line number
with the anomaly. Make sure to remember on which line of the CSV the anomaly was (ignore the first line
since these are the column titles): "
      + data_with_anomaly,
  },
]
```

Prompt Engineering

Chain of Thought

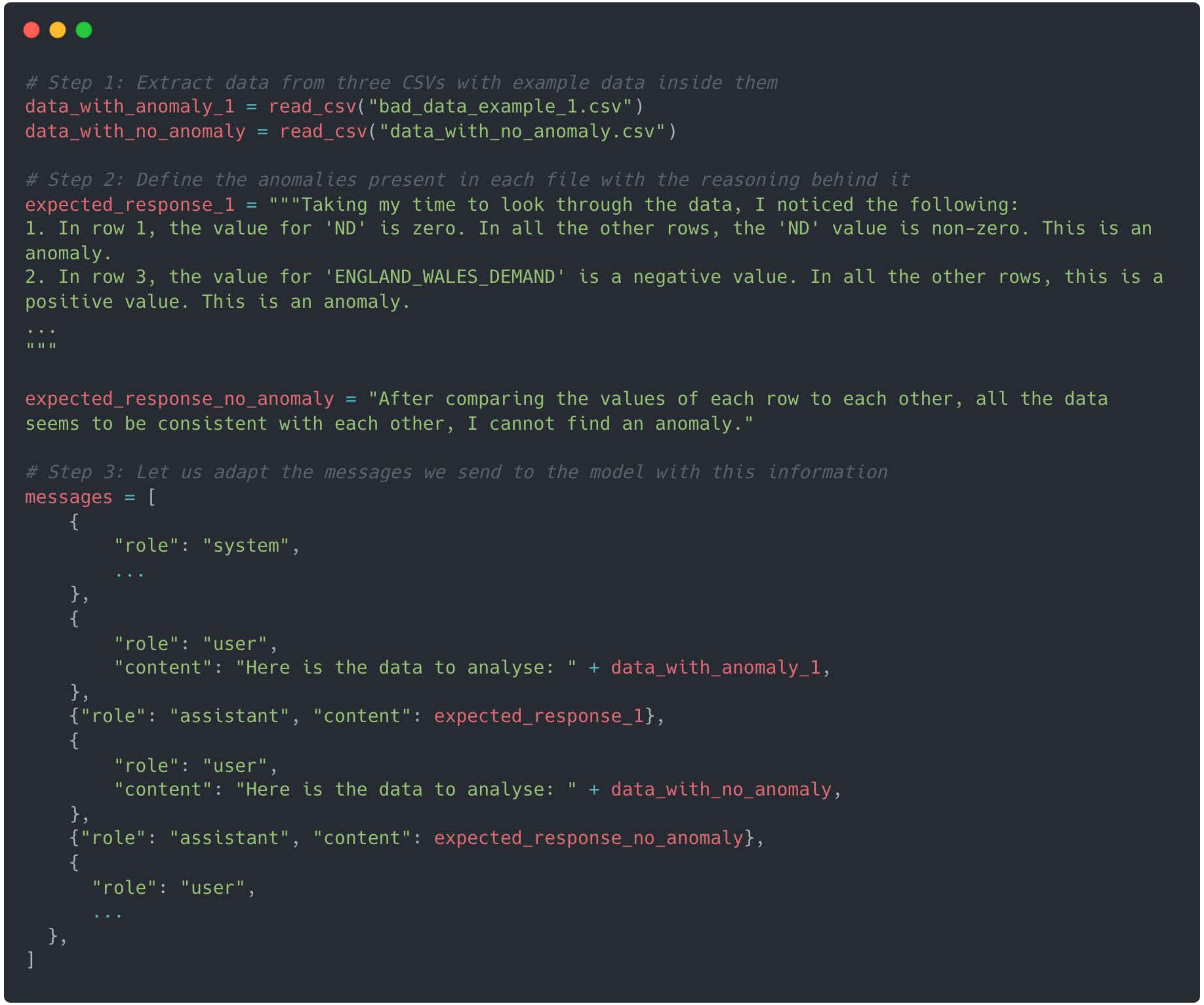
+ 8%

Of intended anomalies detected for GPT 4

Few-shot

Prompt Engineering

Few-shot



```
# Step 1: Extract data from three CSVs with example data inside them
data_with_anomaly_1 = read_csv("bad_data_example_1.csv")
data_with_no_anomaly = read_csv("data_with_no_anomaly.csv")

# Step 2: Define the anomalies present in each file with the reasoning behind it
expected_response_1 = """Taking my time to look through the data, I noticed the following:
1. In row 1, the value for 'ND' is zero. In all the other rows, the 'ND' value is non-zero. This is an
anomaly.
2. In row 3, the value for 'ENGLAND_WALES_DEMAND' is a negative value. In all the other rows, this is a
positive value. This is an anomaly.
...
"""

expected_response_no_anomaly = "After comparing the values of each row to each other, all the data
seems to be consistent with each other, I cannot find an anomaly."

# Step 3: Let us adapt the messages we send to the model with this information
messages = [
    {
        "role": "system",
        ...
    },
    {
        "role": "user",
        "content": "Here is the data to analyse: " + data_with_anomaly_1,
    },
    {"role": "assistant", "content": expected_response_1},
    {
        "role": "user",
        "content": "Here is the data to analyse: " + data_with_no_anomaly,
    },
    {"role": "assistant", "content": expected_response_no_anomaly},
    {
      "role": "user",
      ...
  },
]
```

Prompt Engineering

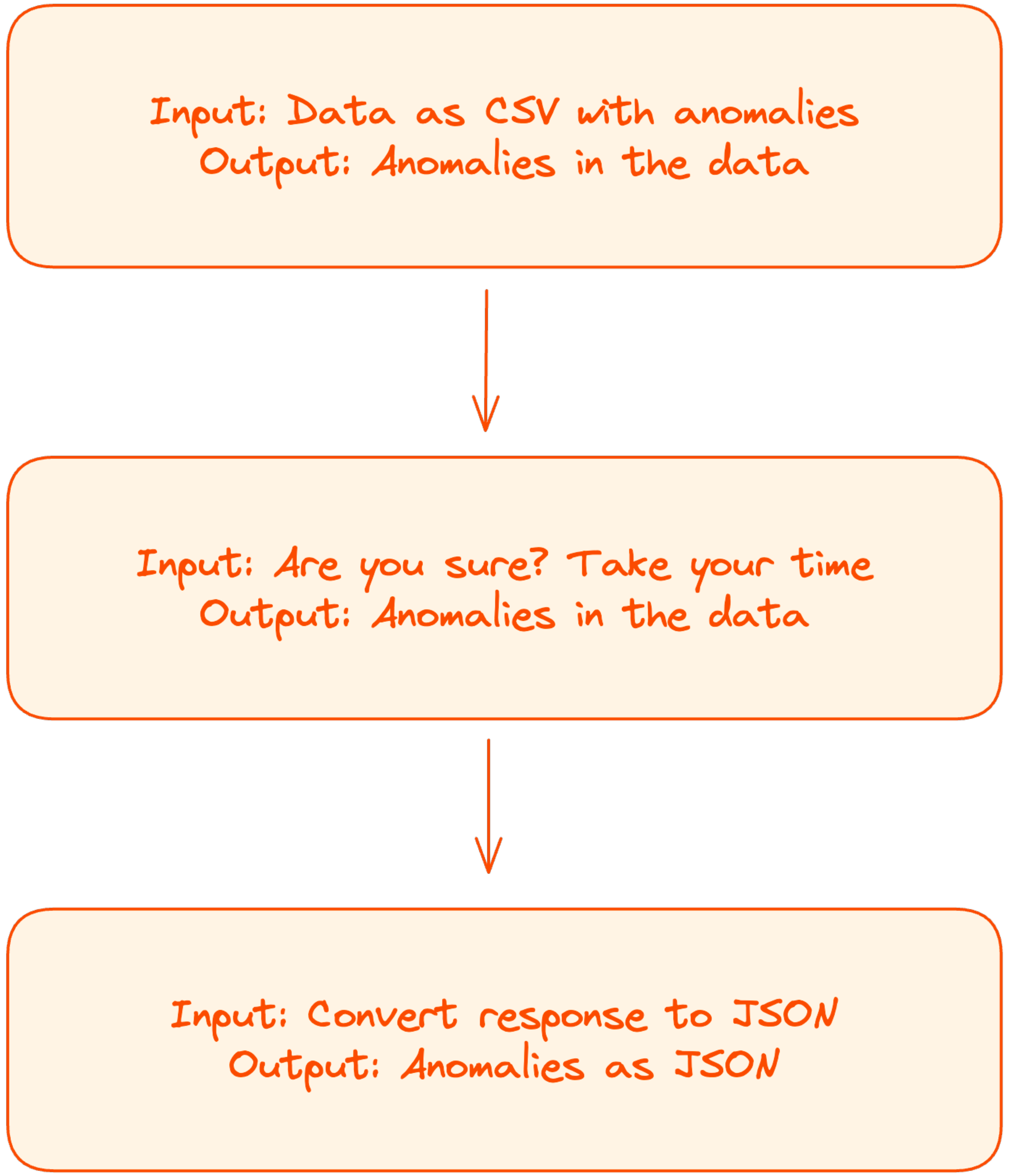
Few-shot

+ 24%

Of intended anomalies detected for GPT 4

Self-reflection & multi-step

Self-reflection & multi-step



Input: Data as CSV with anomalies
Output: Anomalies in the data

Input: Are you sure? Take your time
Output: Anomalies in the data

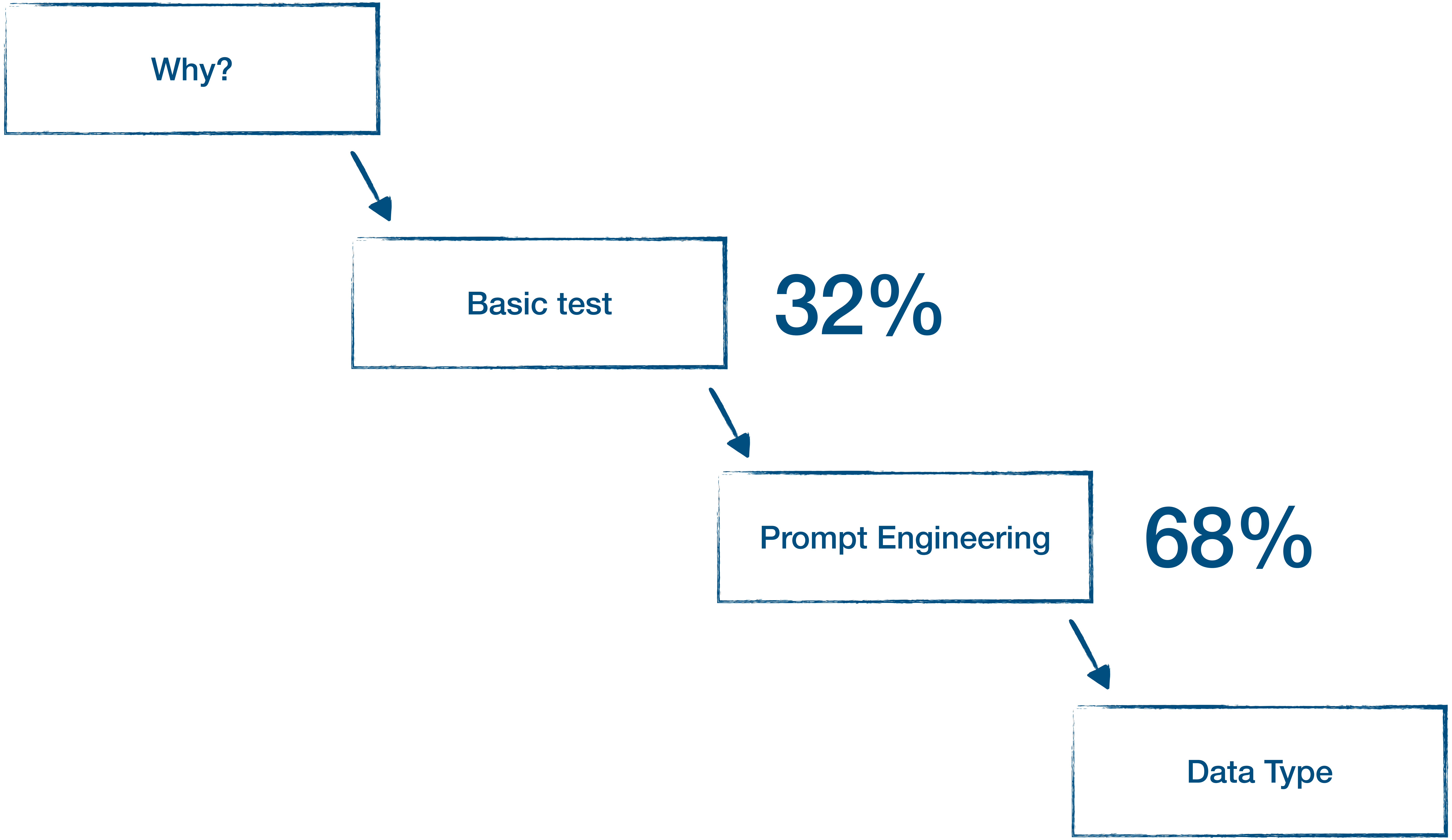
Input: Convert response to JSON
Output: Anomalies as JSON

Prompt Engineering

Self-reflection & multi-step

+ 28%

Of intended anomalies detected for GPT 4



Why?

Basic test

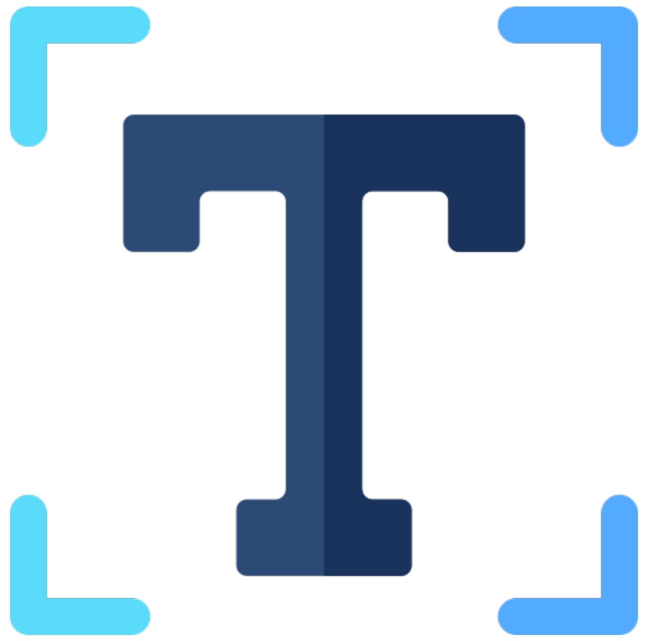
32%

Prompt Engineering

68%

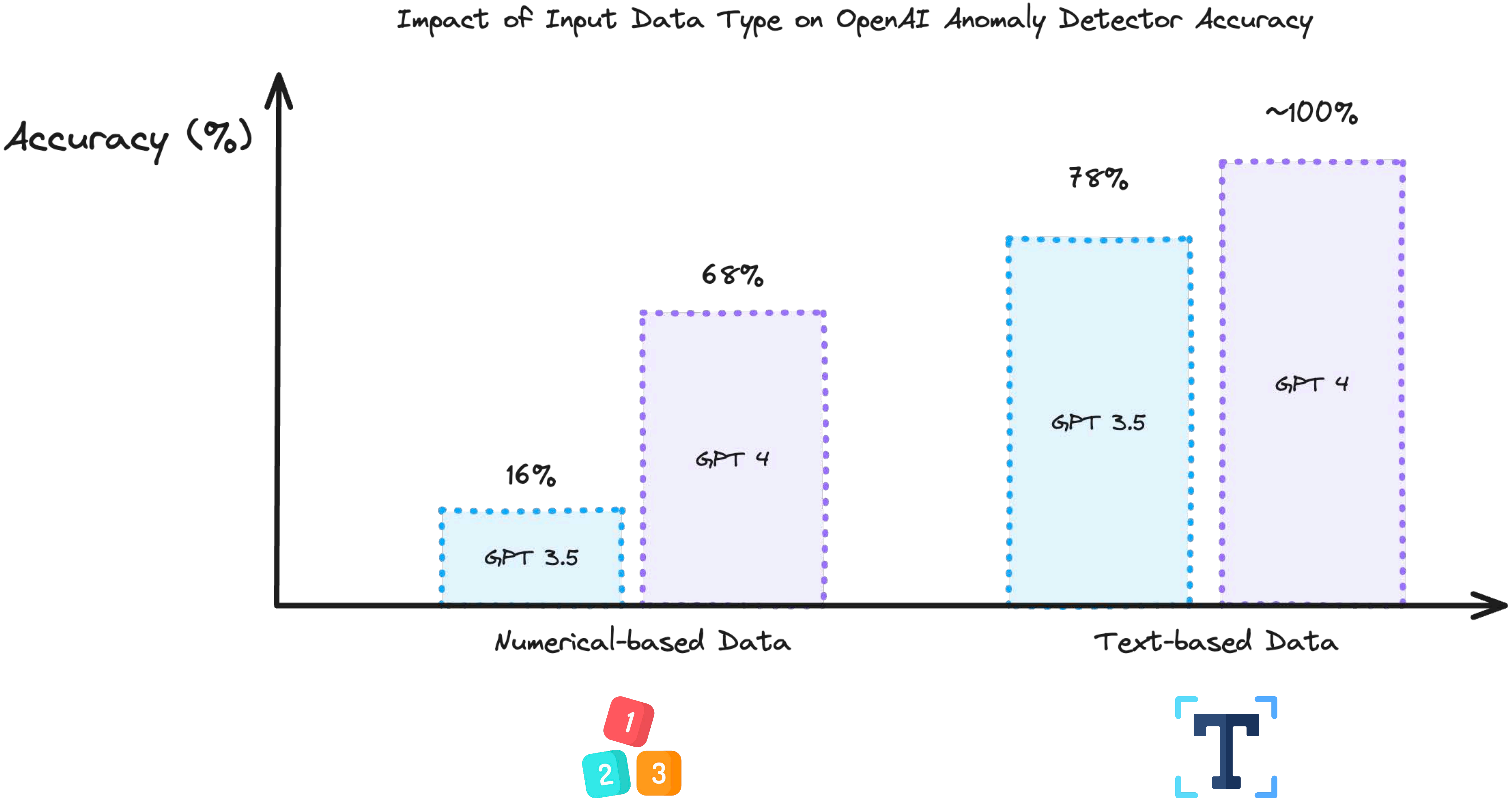
Data Type

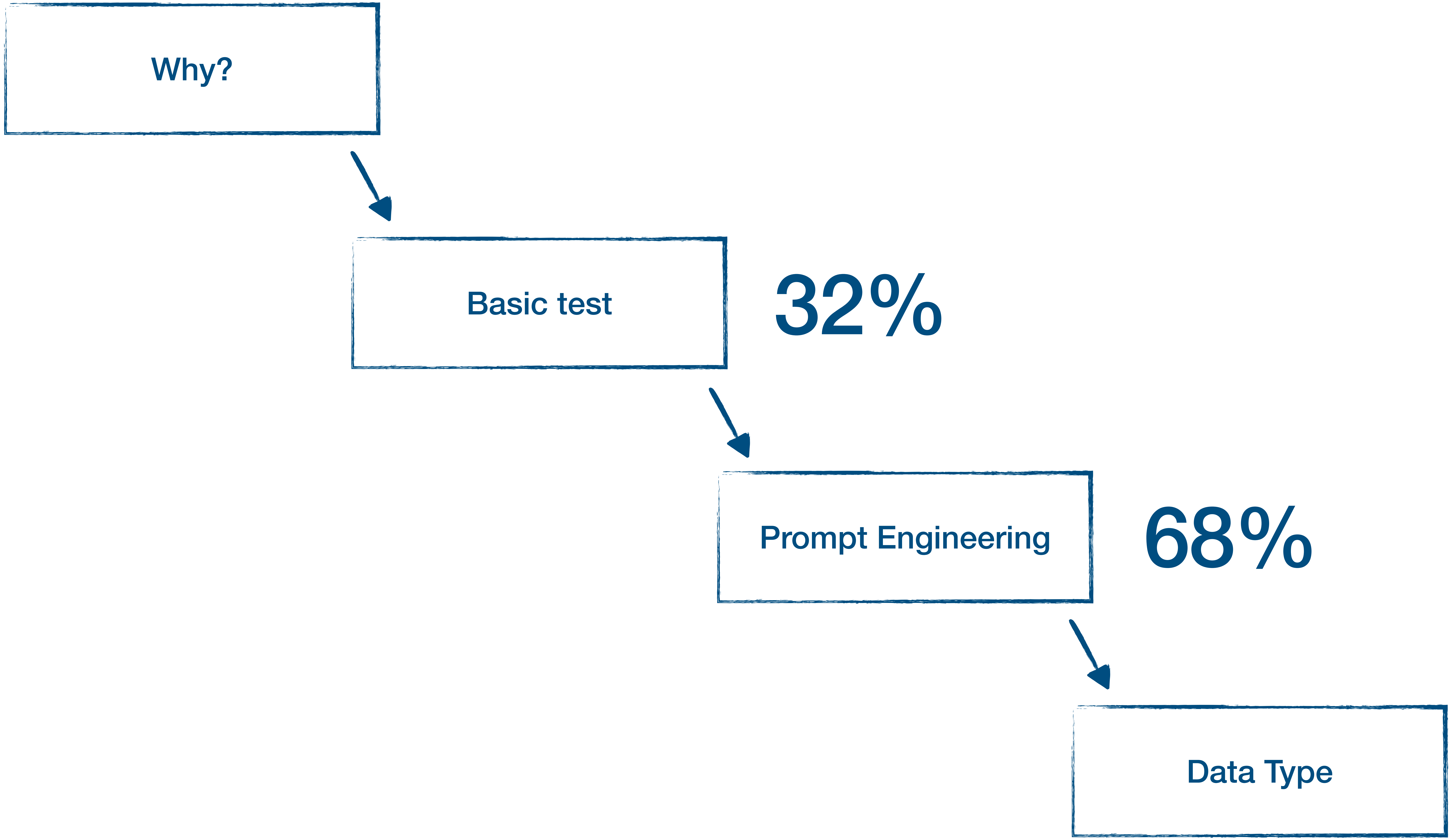
Data Type



VS

Data Type


Impact of Input Data Type on OpenAI Anomaly Detector Accuracy
Accuracy (%)
16%
GPT 3.5
68%
GPT 4
78%
GPT 3.5
~100%
GPT 4
Numerical-based Data
Text-based Data
1
2
3
T



Why?

Basic test

32%

Prompt Engineering

68%

Data Type

~100%

1
2
3

BigQuery has an in-built anomaly detector

1. Choose a **model** to fit your data, e.g. ARIMA PLUS
2. Create a **model** for each data column
3. Run your anomaly detector for each data column

Lots of code but this is the key part

```
FROM ML.DETECT_ANOMALIES(MODEL `model_name`, STRUCT(0.95 AS anomaly_prob_threshold))
```

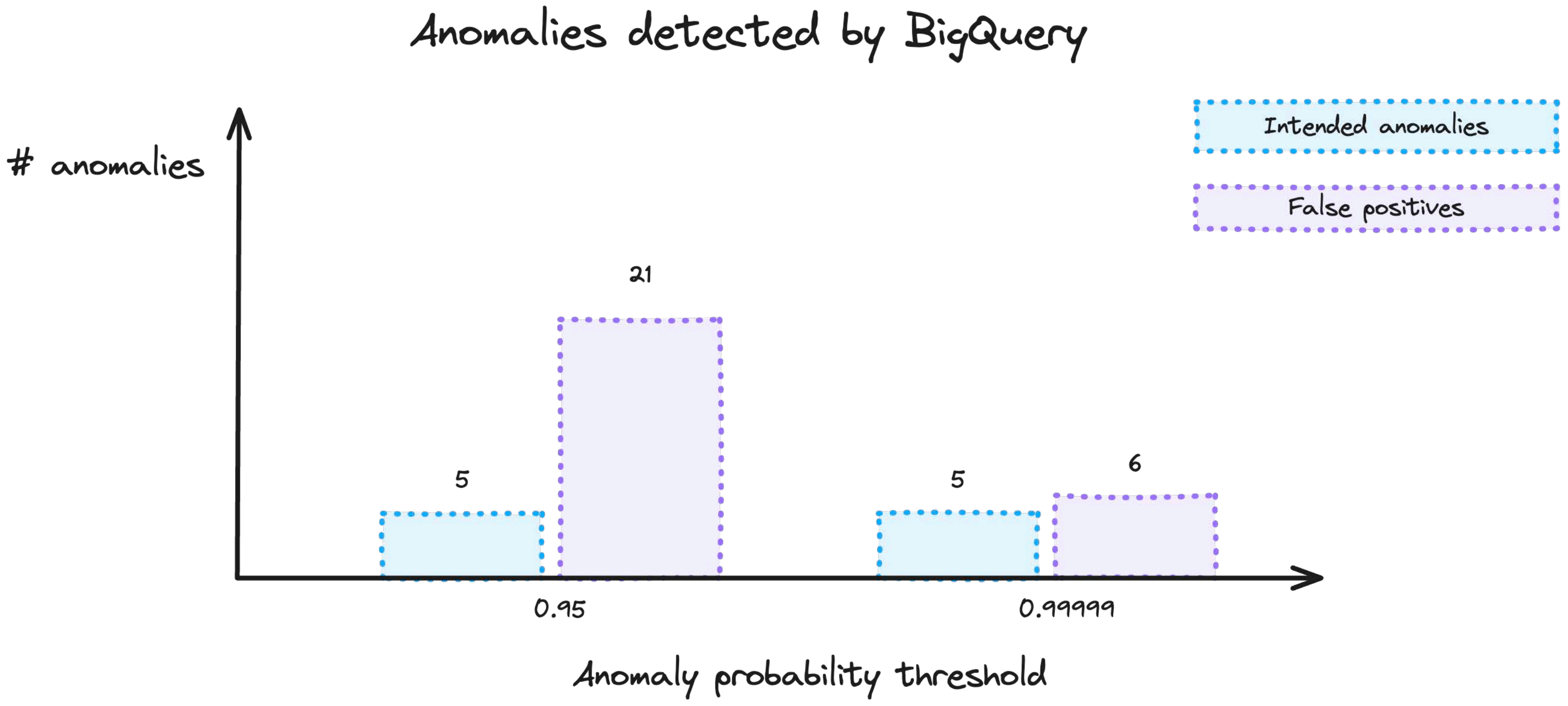
🎉 100% 🎉

Of intended anomalies picked up

21

False positives for 28 lines of data and 5 intended anomalies

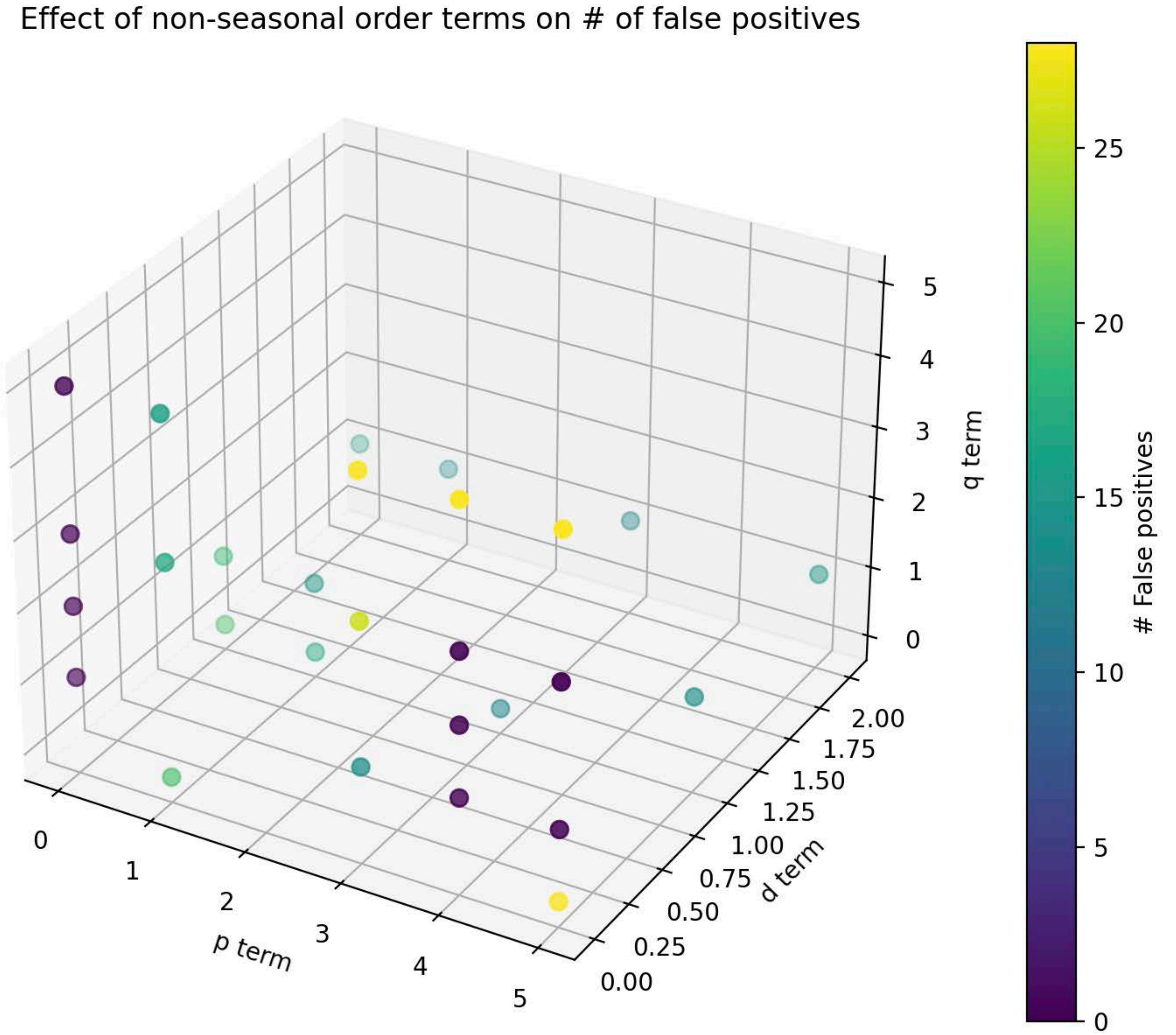
Increase the threshold



Adding separate training data

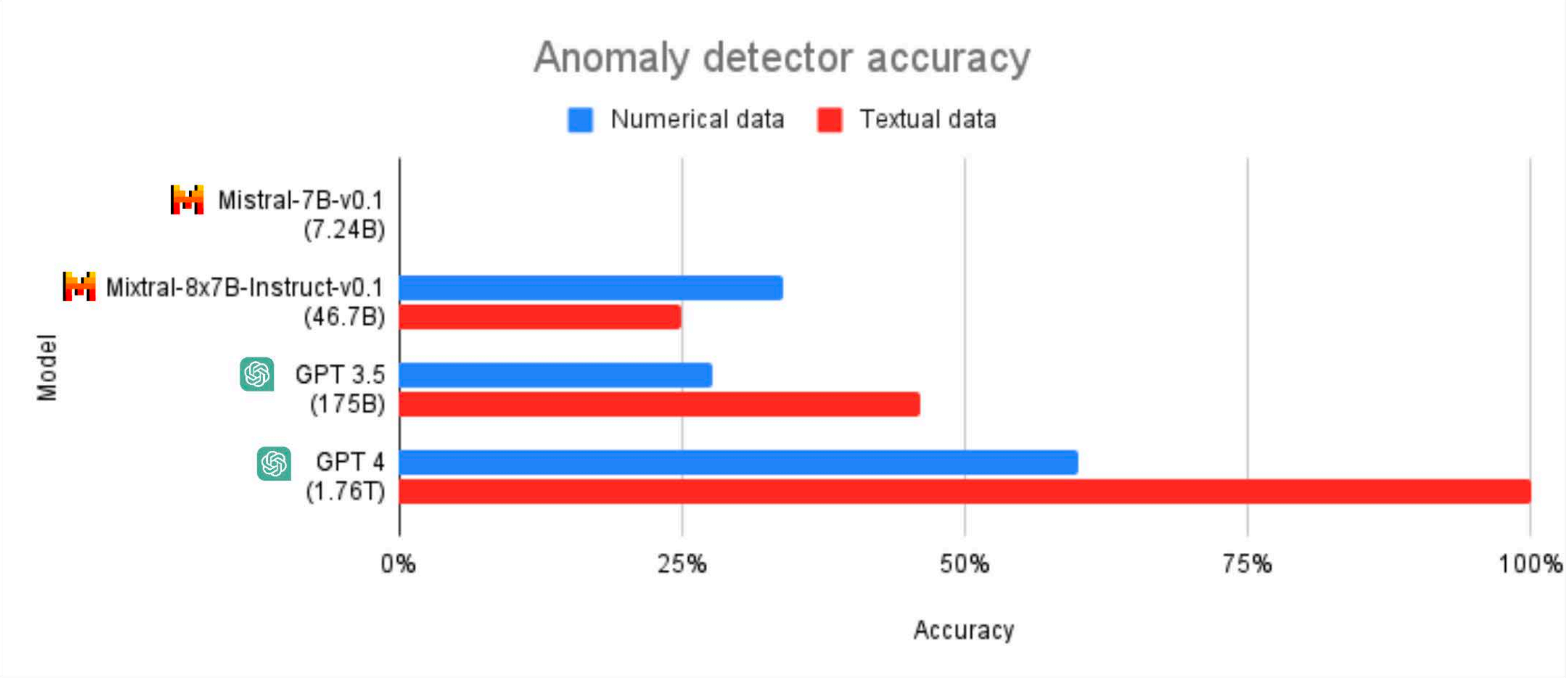
✗

Tuning non-seasonal order terms



1

False positives for 28 lines of data and 5 intended anomalies

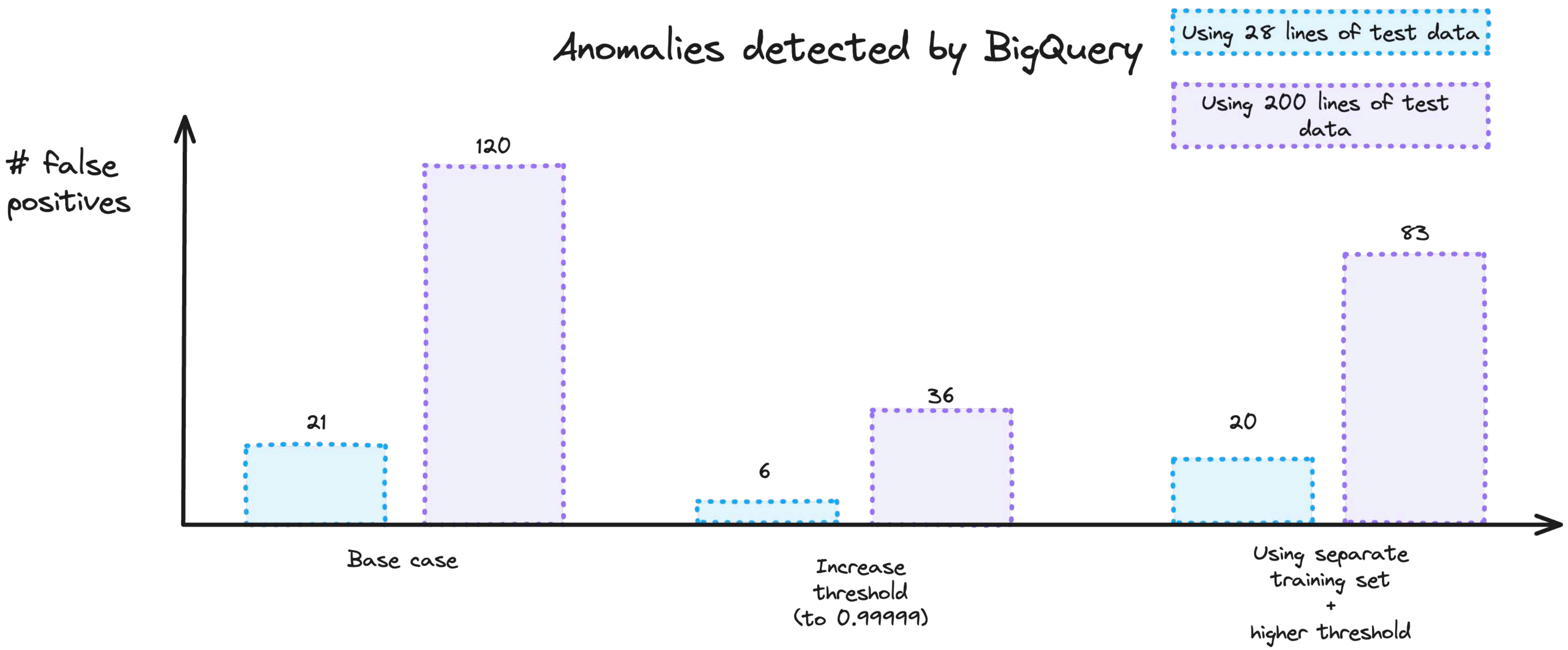

Anomaly detector accuracy
Numerical data
Textual data
Model
Mistral-7B-v0.1
(7.24B)
Mixtral-8x7B-Instruct-v0.1
(46.7B)
GPT 3.5
(175B)
GPT 4
(1.76T)
0%
25%
50%
75%
100%
Accuracy

X

@ChloeCaronEng

LinkedIn

Anomalies detected by BigQuery


Using 28 lines of test data
Using 200 lines of test data
# false positives
Base case
Increase threshold (to 0.99999)
Using separate training set + higher threshold