Observability on LLMs

Grupo 8

Javier Carrasco Arango

Sergio Mulet Alonso

Saúl Martín Fernández

Adrián Martínez Fuentes

Index

1. *Observability*

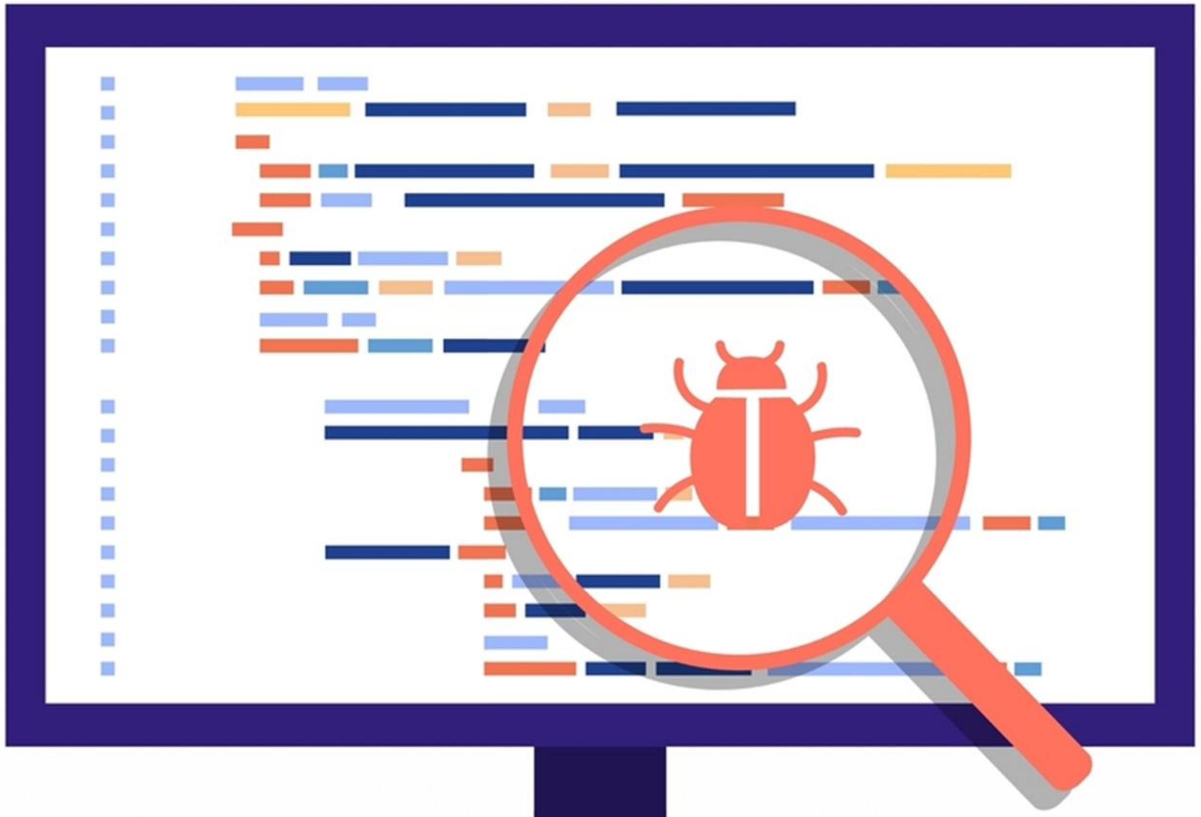
2. *Large Language Models*

3. *Integration between LLMs and Observability*

4. *Observability Driven Development*

What is observability

- *A new concept, quite recent*
- *Each company has its own definition for it*
- *Rooting the definition on the problem*



Rooting the definition on the problem it solves

Generally, the process goes like this:

1. identify errors
2. Debug
3. find the error, analyze it
4. decide on the approach, on how to change the system to account for this or prevent it.

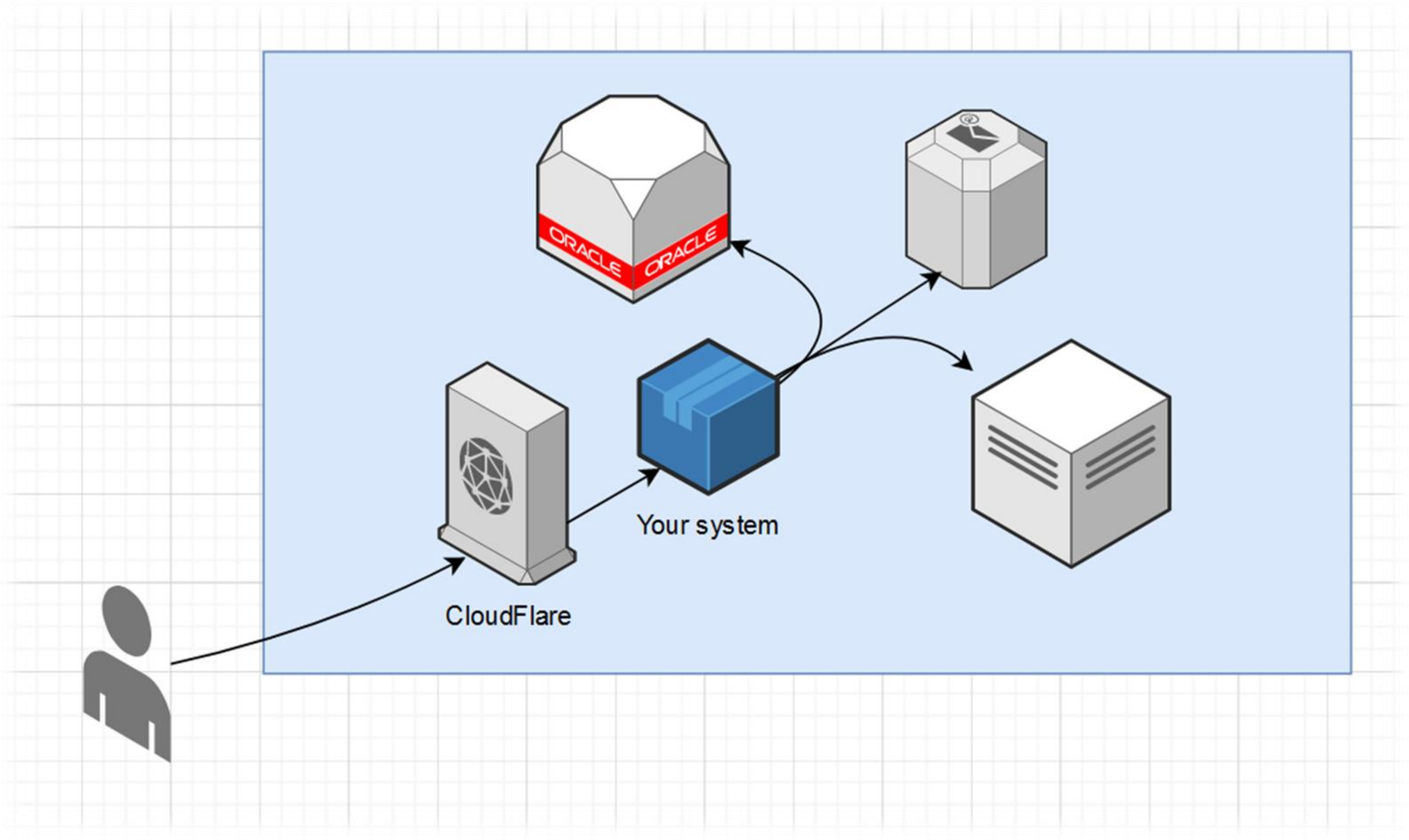
But what if you can't do that?

Observability to solve interaction

When using closed off systems

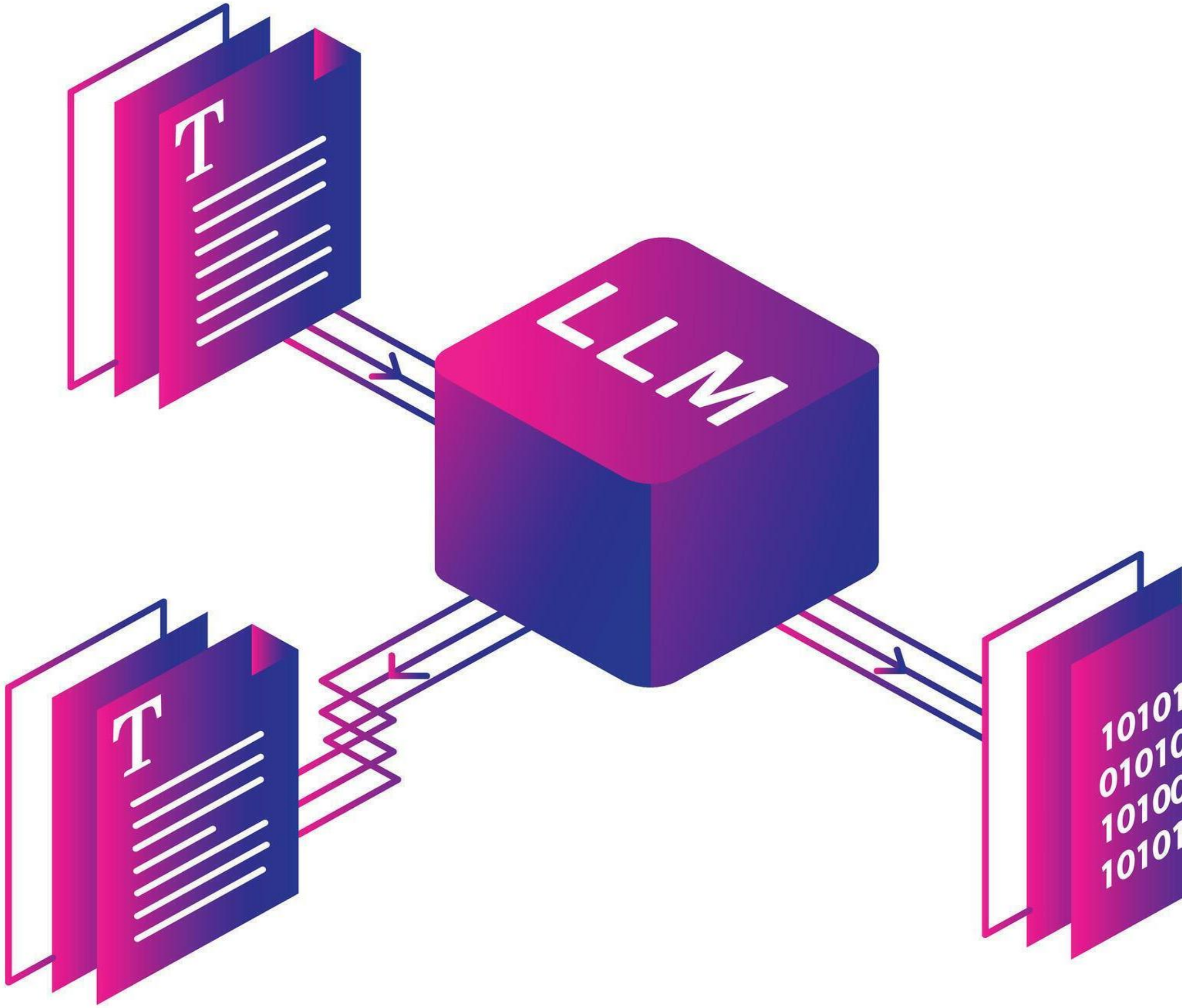
Collecting and analyzing telemetry to spot the issues

Ie. Latency, non-cohesiveness



What are LLMs

- *Computer programs understanding human language*
- *Transformers Architecture*
- *Training and hosting an LLM*



Training and hosting a Large Language Model

Two training phases:

- Pre - Training
- Fine tuning

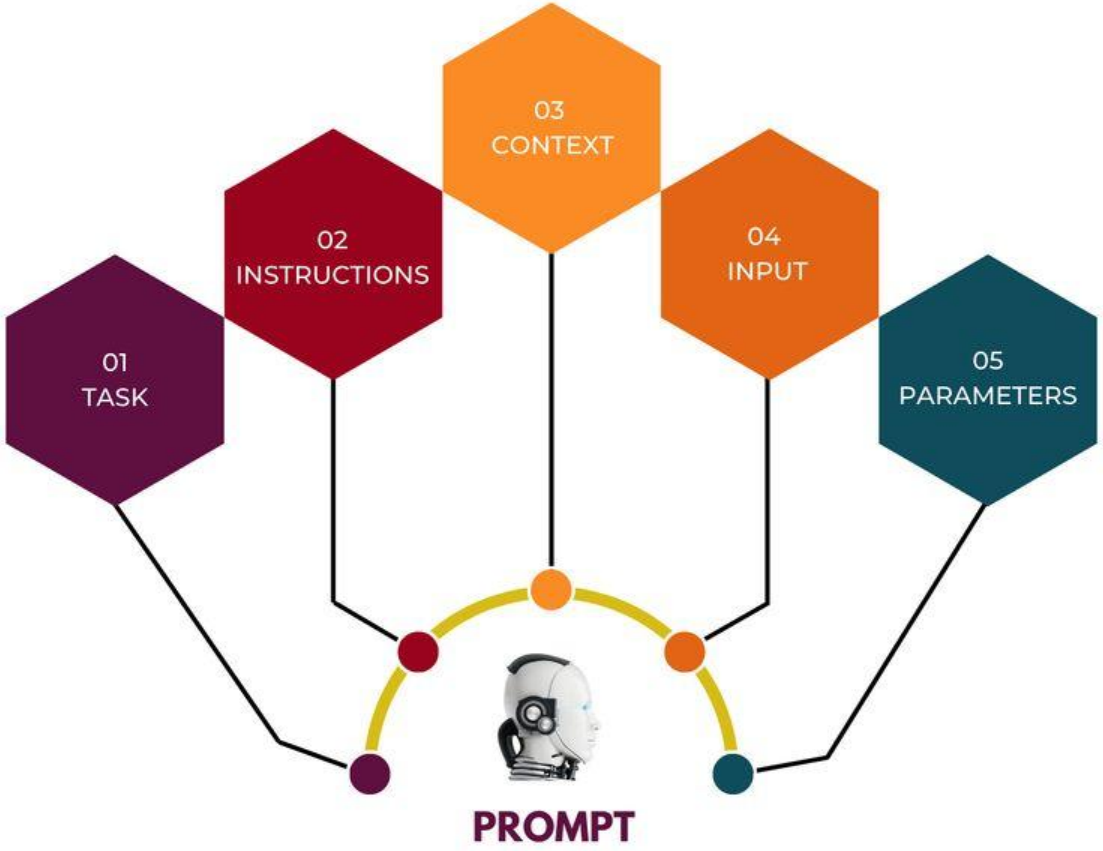
Hosting:

- With cloud services
- And specific hardware

Prompt engineering

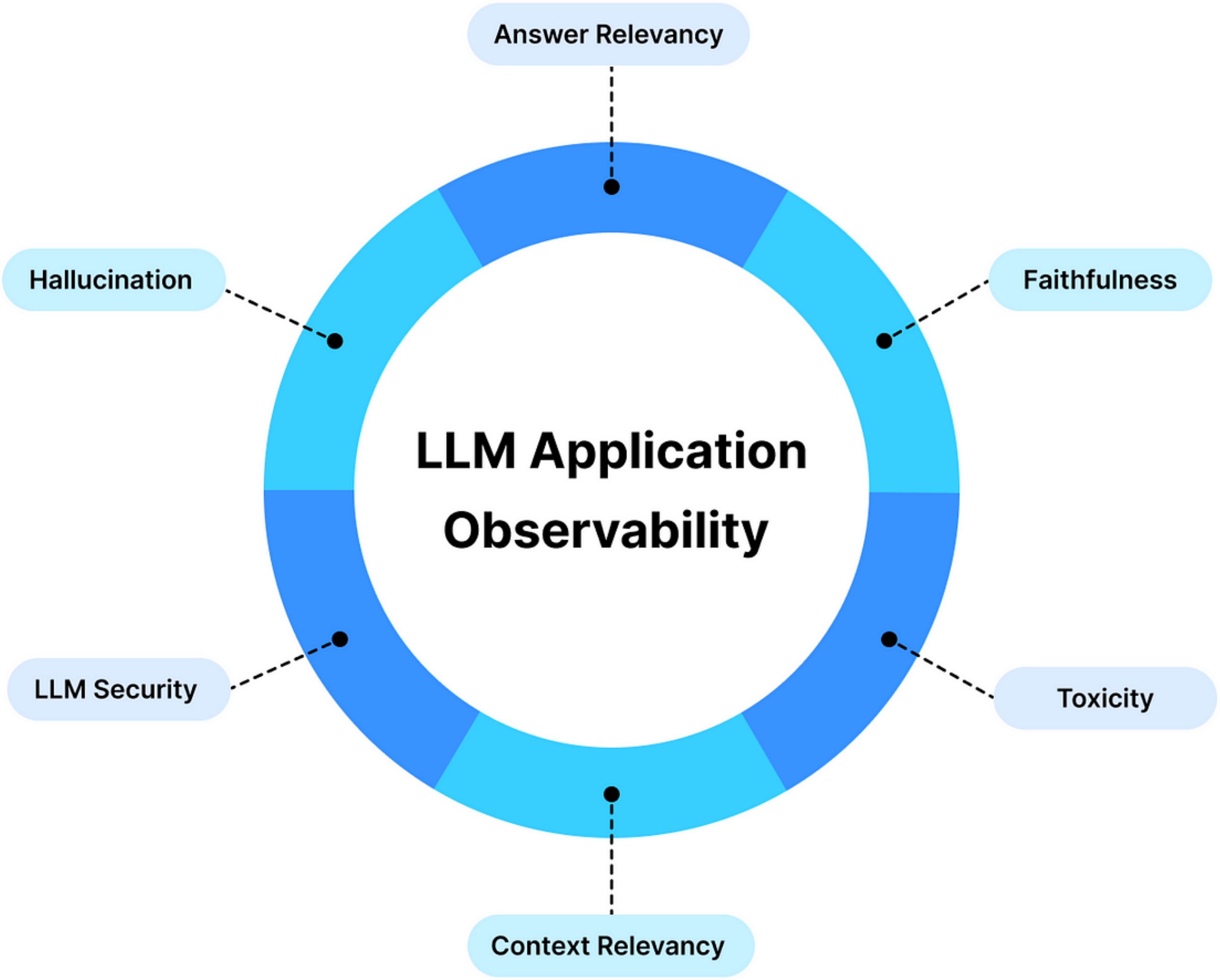
What is a prompt?

Optimizing the queries we make to the LLM



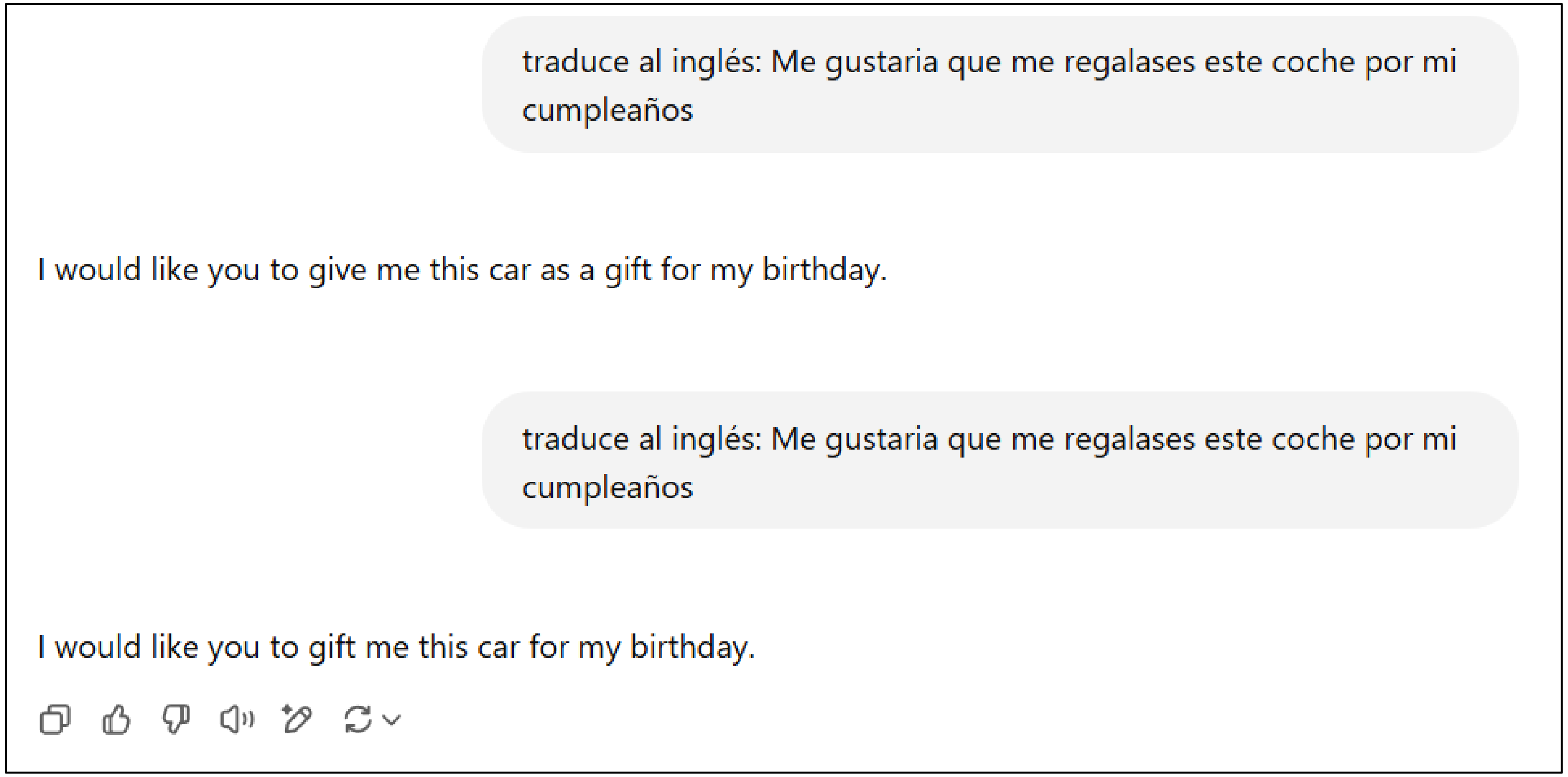
Observability with LLMs

- Observability matters
- Challenges
- Solutions



Observability matters

- Answers are non-deterministic
- Unpredictable user inputs
- Hard to debug



traduce al inglés: Me gustaria que me regalases este coche por mi cumpleaños

I would like you to give me this car as a gift for my birthday.

traduce al inglés: Me gustaria que me regalases este coche por mi cumpleaños

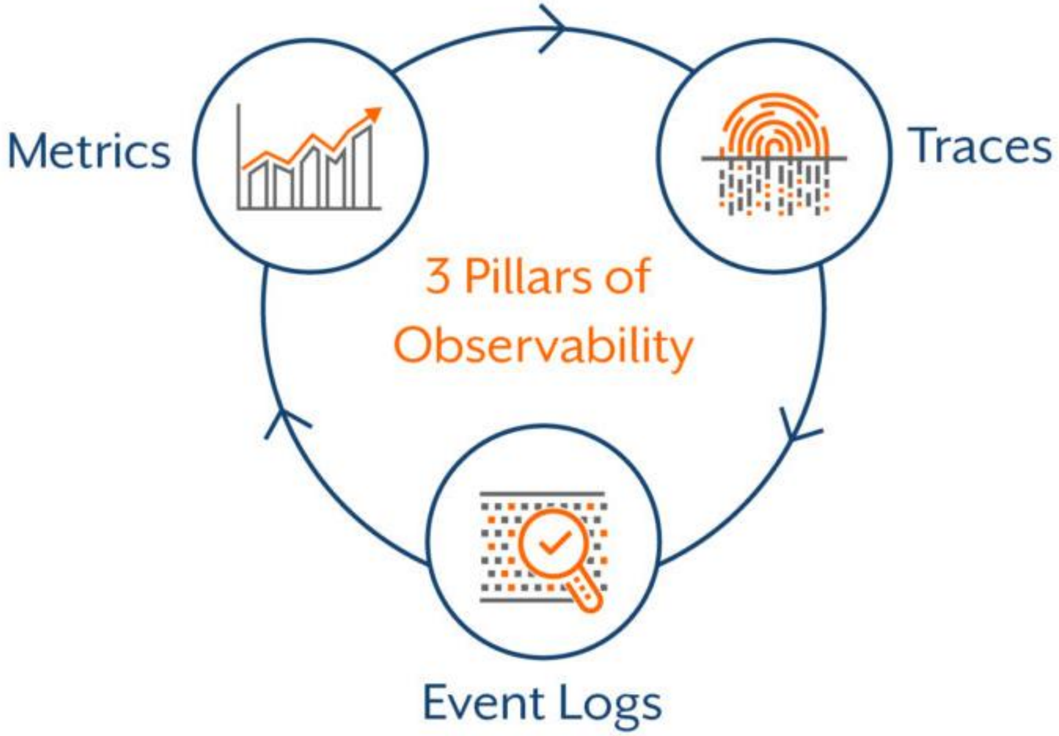
I would like you to gift me this car for my birthday.

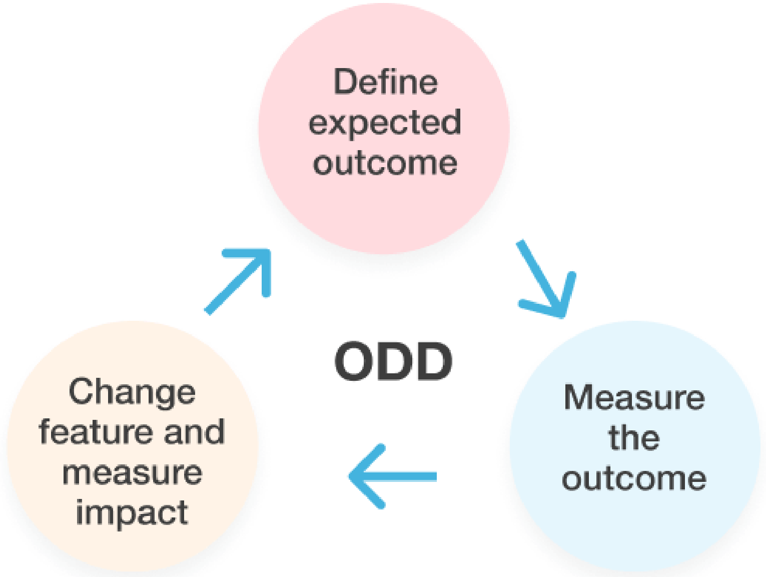
Challenges and solutions

Challenges	Solutions
Complex decision chains	Log inputs and outputs
Mistakes	Track upstream and downstream
Latency	Optimize performance

Tools for Observability

- *Structured **Logging***
- *Tracing with **OpenTelemetry***
- *Data Analysis with **Honeycomb***
- *Other Tools (**Prometheus** & **Grafana**)*





Observability Driven Development (ODD)

- *Iterative Improvement Based on Real Data*
- *Feedback Loop*
- *Balancing Reliability And Innovation*

The end

Thanks for you attention.