

# How to Ground LLM's to minimize hallucinations



# Cameron Vetter

Principal Architect  
Microsoft MVP for AI  
Deep Learning  
Generative AI  
Cloud Architecture



- 1 Artificial Intelligence**
- 2 Productionalizing Machine Learning**
- 3 Cloud Architecture**
- 4 Technology Roadmaps**





WHO ARE YOU?

AUDIENCE

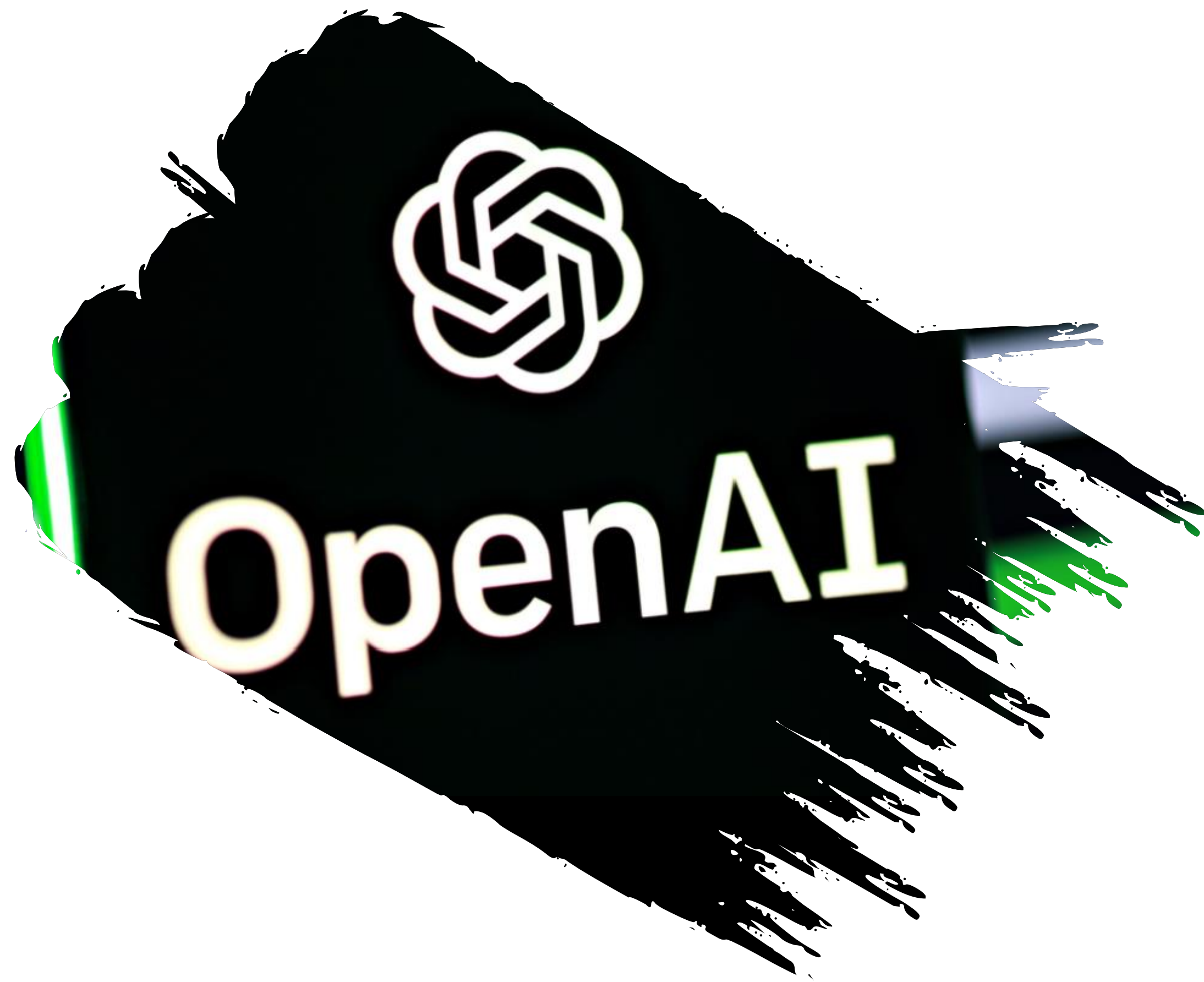




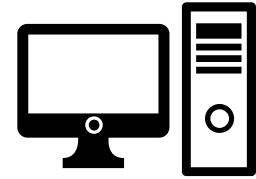
HOW ARE YOU USING AI  
CURRENTLY?

AUDIENCE





WHAT ARE LARGE  
LANGUAGE  
MODELS (LLM)?



**LLM'S ARE NOT THE  
SOLUTION TO  
EVERY PROBLEM**

**HAMMER**



**AI ISN'T THE  
SOLUTION TO  
EVERY PROBLEM**

**SLEDGEHAMMER**

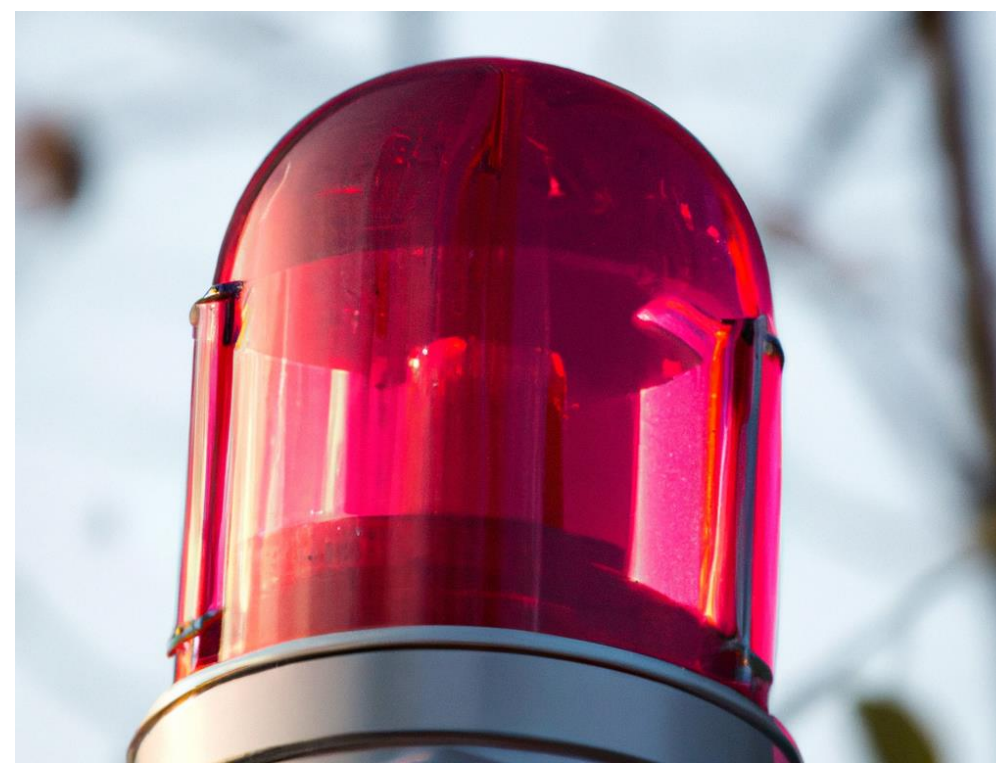


WHAT ARE  
HALLUCINATIONS?





**OPENAI STATES:** “*ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers.*”







WHO HAS EXPERIENCED  
HALLUCINATIONS?

AUDIENCE



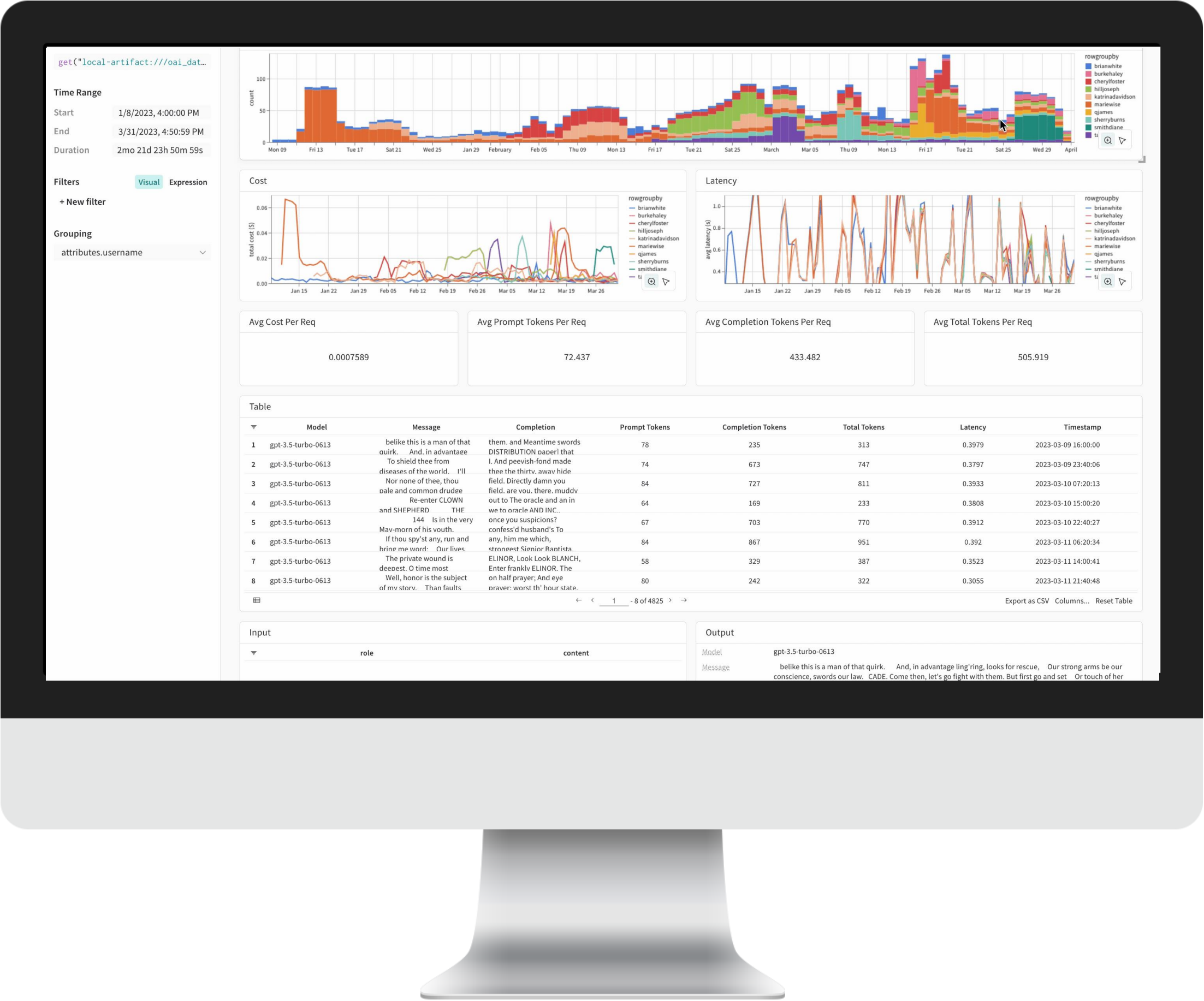
WHAT DO WE DO?

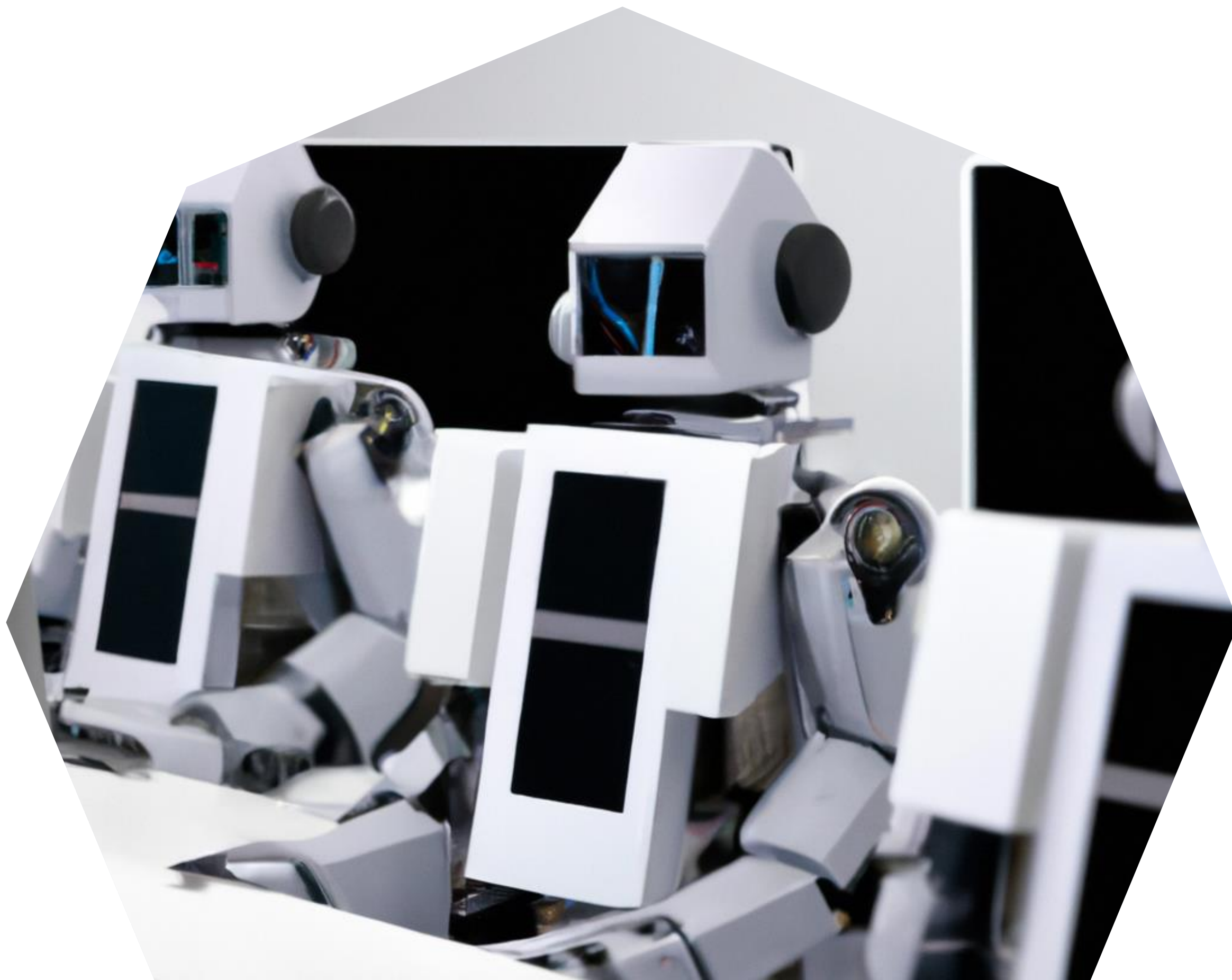
UNDERSTANDING THE  
PROBLEM!





# Weights & Biases Instrumentation

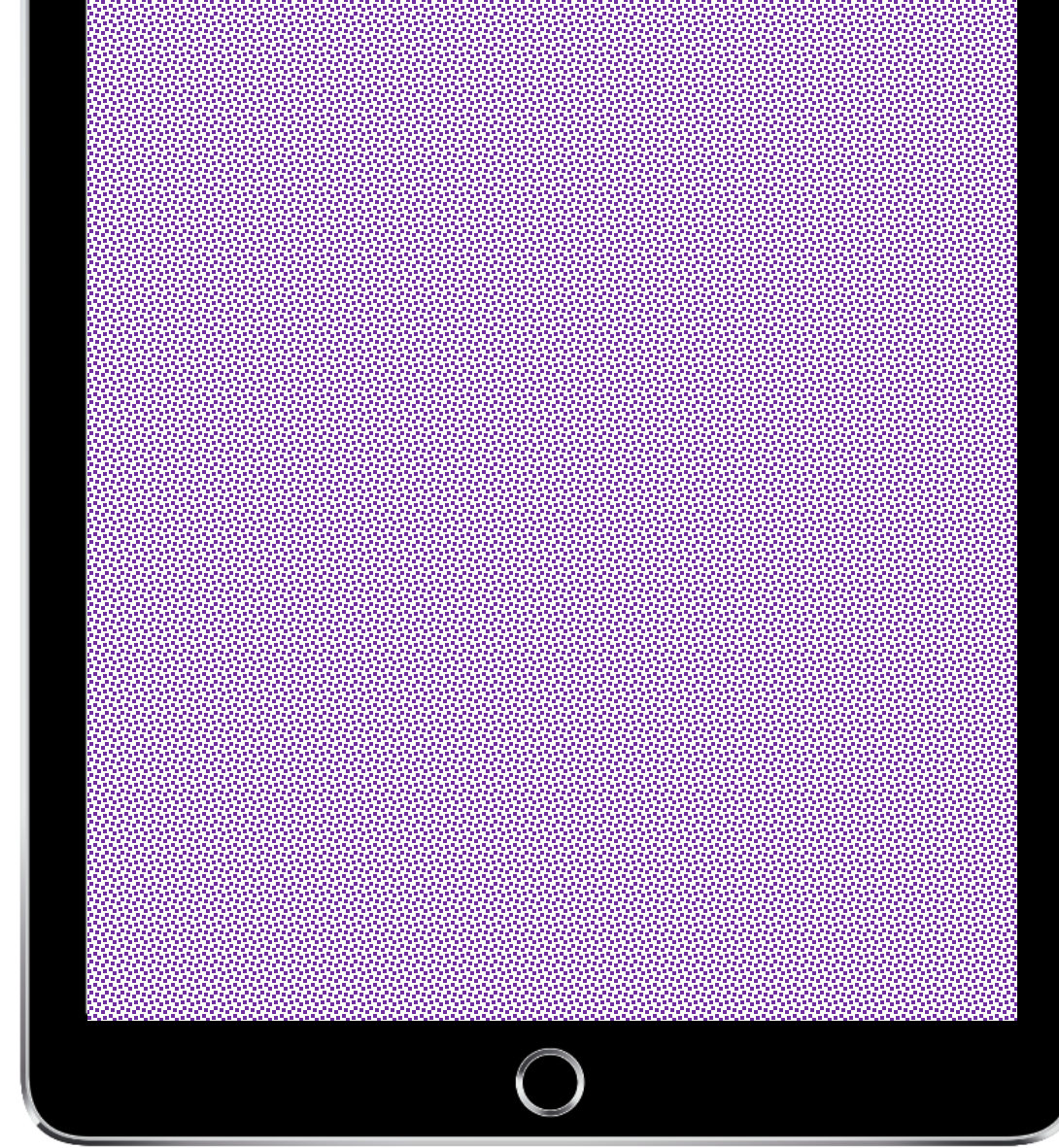
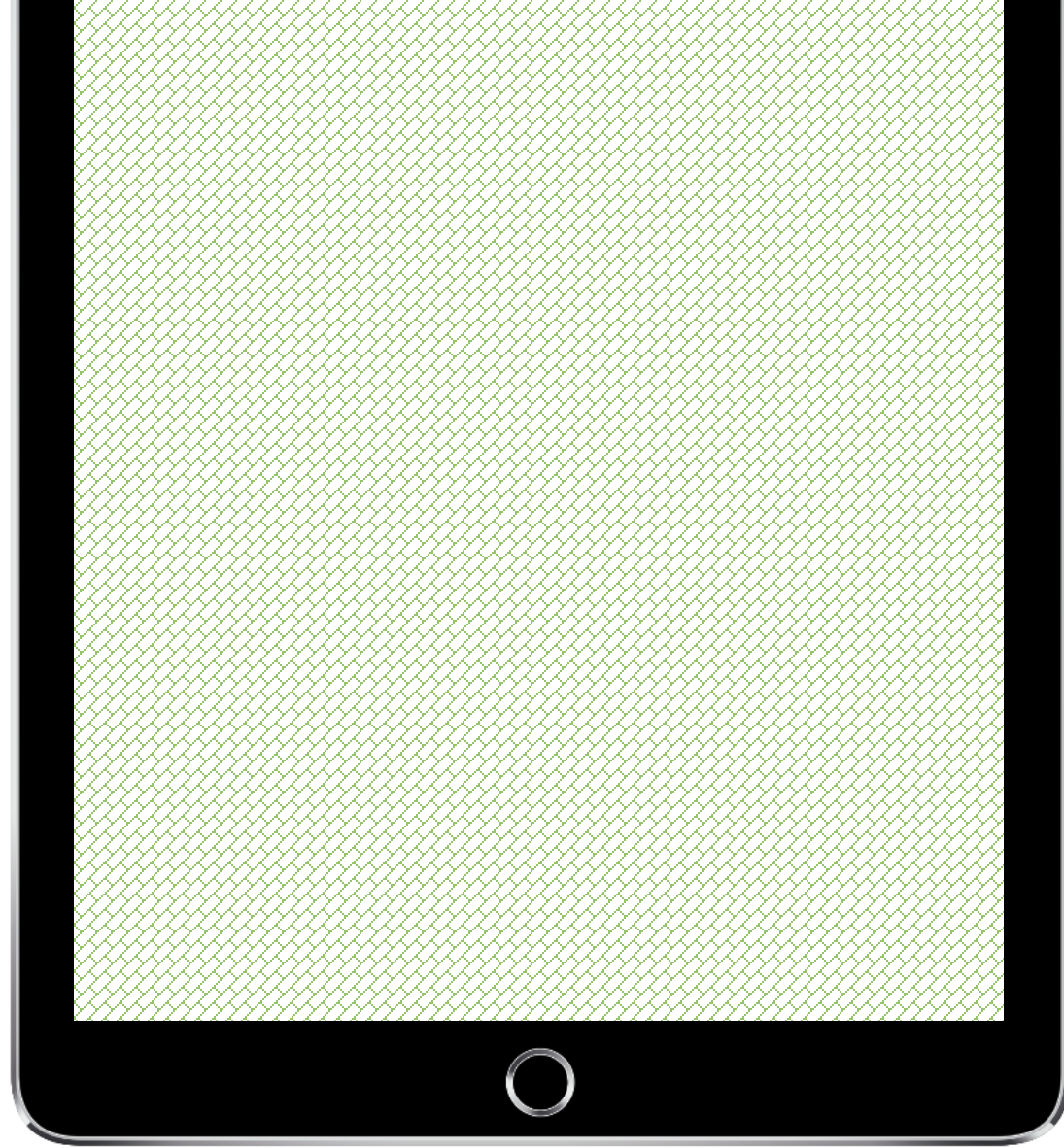




## AUTOMATED TEST SUITE

Just like all software  
development, a good  
automated test suite is crucial.

But yet very different...



# AUTOMATED A / B TESTING

As you make changes, use your automated tests to perform automated testing to check for regressions, quality, and hallucinations!



SOLUTIONS  
FOR **OFF THE**  
**SHELF** LLM's



RAG DESIGN PATTERN



FINE TUNING




PROMPT ENGINEERING



MULTI MODEL SOLUTIONING



MODEL SELECTION



# RETRIEVAL AUGMENTED GENERATION

A methodology for building  
Generative AI applications over  
private or custom datasets.

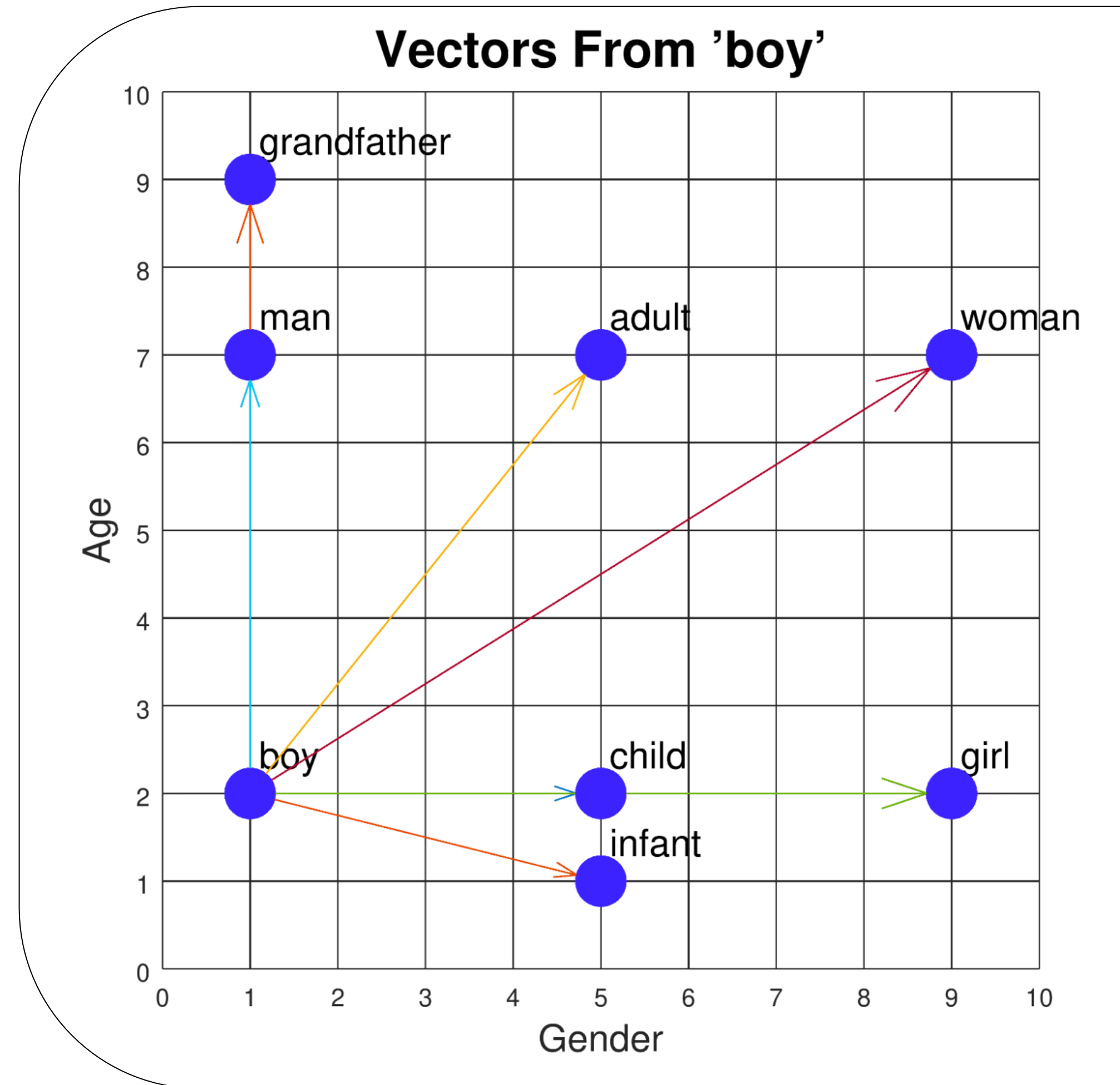
```
array([-0.5968882 , -0.33086956, -0.32643065, -0.3670732 ,  0.628059
       -0.3692328 , -0.37902787, -0.12308089, -0.38124698, -0.03940517
        0.2260839 ,  0.10852845, -0.2873811 , -0.42781743,  0.06604357
       -0.07114276, -0.29775023, -0.99628943, -0.54497653, -0.11718027
       -0.15935768,  0.09587188, -0.2503798 ,  0.06768776,  0.3311586
        0.43098116,  0.06936899,  0.24311952,  0.14515282,  0.19245838
        0.10462623, -0.45676082,  0.5662387 ,  0.69908774,  0.48064467
        0.27378514, -0.45430255,  0.17282294, -0.40275463, -0.38083532
        0.47487524,  0.31950948, -0.1109335 ,  0.2165357 ,  0.034114
        0.05689918,  0.20939653,  0.15209009, -0.24204595,  0.03478364
        0.1616051 , -0.5827333 , -0.47017908,  0.26226178, -0.11884775
        0.40180743, -0.5173988 , -0.19270805,  0.660391 , -0.24518126
       -0.42860952, -0.22274768,  0.4887834 ,  0.49302152,  0.38799986
       -0.041193 , -0.38600504, -0.37632987,  0.04570564,  0.50462466
       -0.14396502,  0.33490512, -0.15964787, -0.21363072, -0.25445372
        0.52389127,  0.5747422 , -0.25075617, -0.5339069 ,  0.2582965
       -0.16139959,  0.09748188,  0.04540966, -0.27768216, -0.51260656
       -0.06189002, -0.54032195, -0.21863565,  0.06233869,  0.13287479
        0.49741864,  0.1772418 ,  0.02064824, -0.04775626, -0.16804916
        0.4643644 ,  0.5546319 ,  0.68051434,  0.7790246 ,  0.5617202
dtype=float32)
```

# VECTOR DATABASES

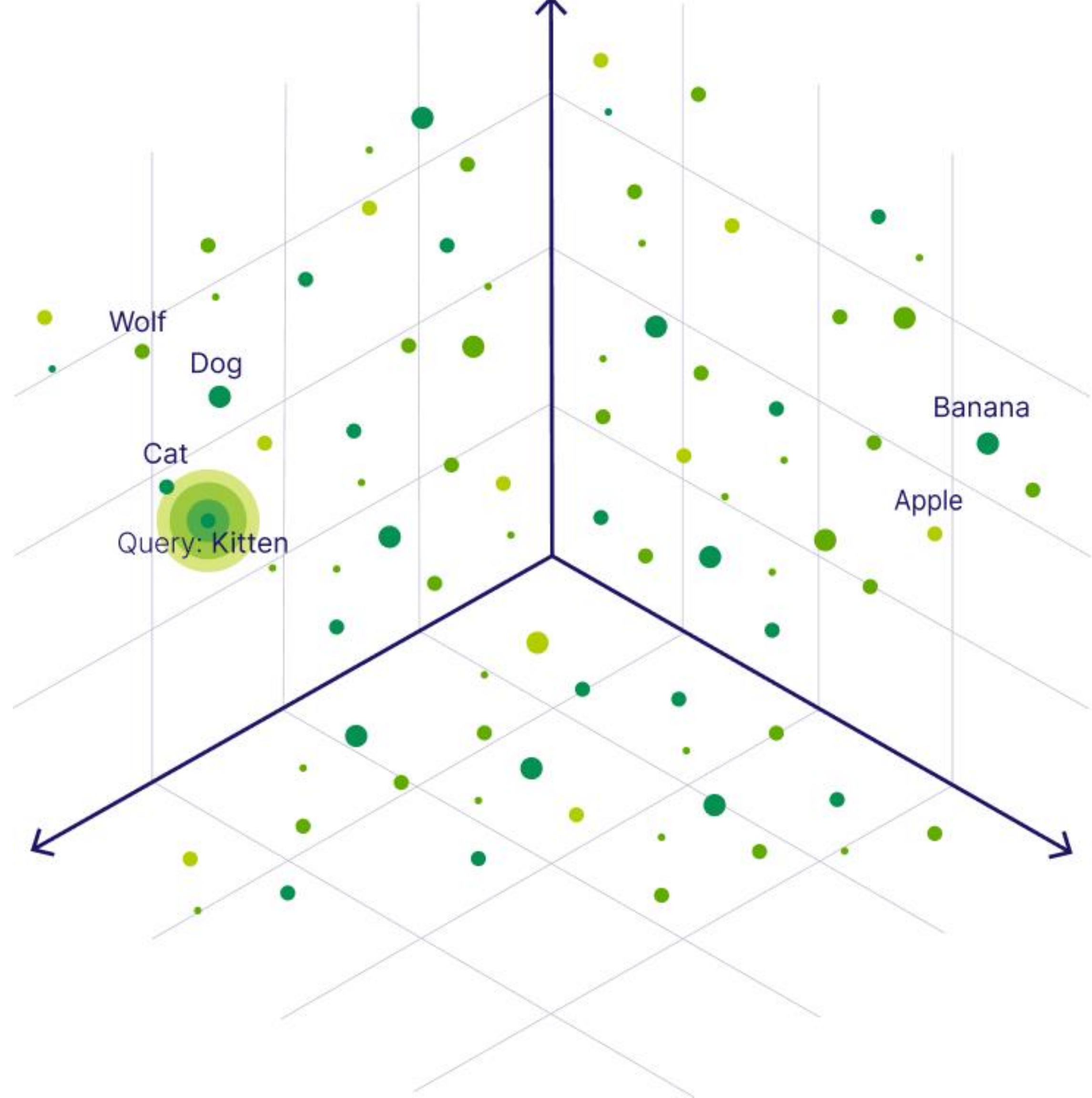
A vector describes the relationship  
between two points with respect to a  
given number of dimensions.



# VECTOR DATABASES



# VECTOR DATABASES





# FINE TUNING

Transfer Learning...

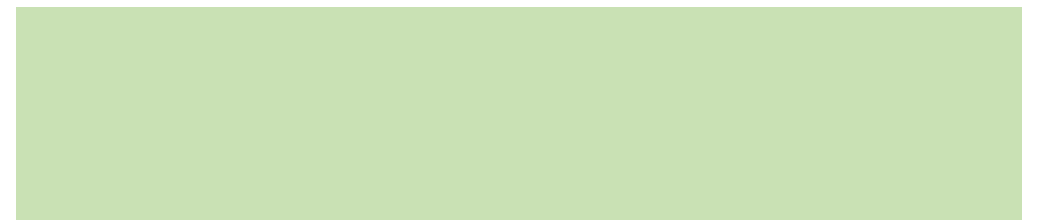
Sort of...



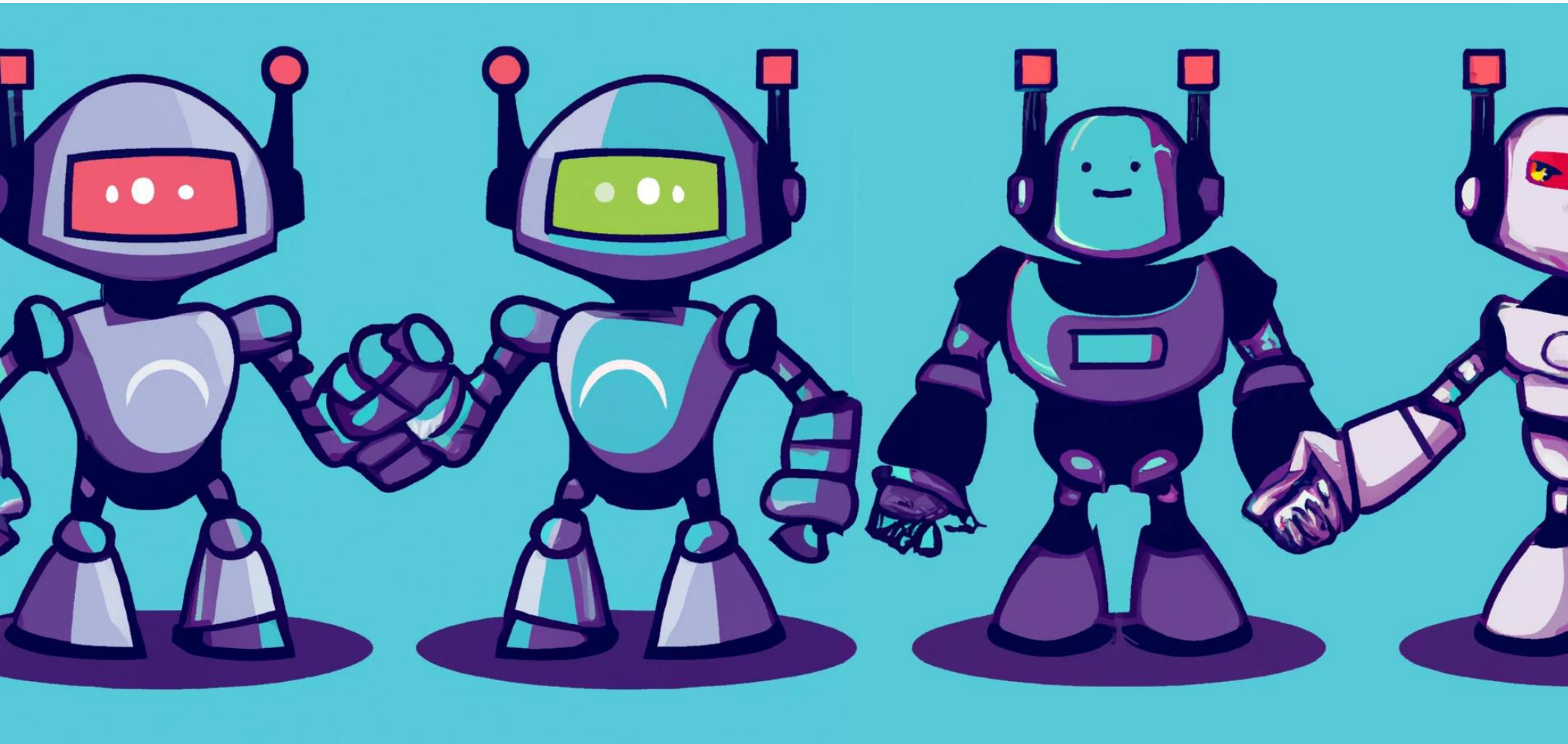


# PROMPT ENGINEERING

It's not just a bogus new  
career path.







## MULTI MODEL SOLUTIONING

Many frameworks such as LangChain help combine the results of models together.



## MODEL SELECTION

GPT3.5, GPT 4, Bard, Llama, Flacon, PaLM, Claude, etc

SOLUTIONS  
FOR **CUSTOM**  
**TRAINED** AND  
**FINE TUNED**  
LLM's



**BAD DATA**



**OVERFITTING**



**IDIOMS OR SLANG IN  
PROMPTS**

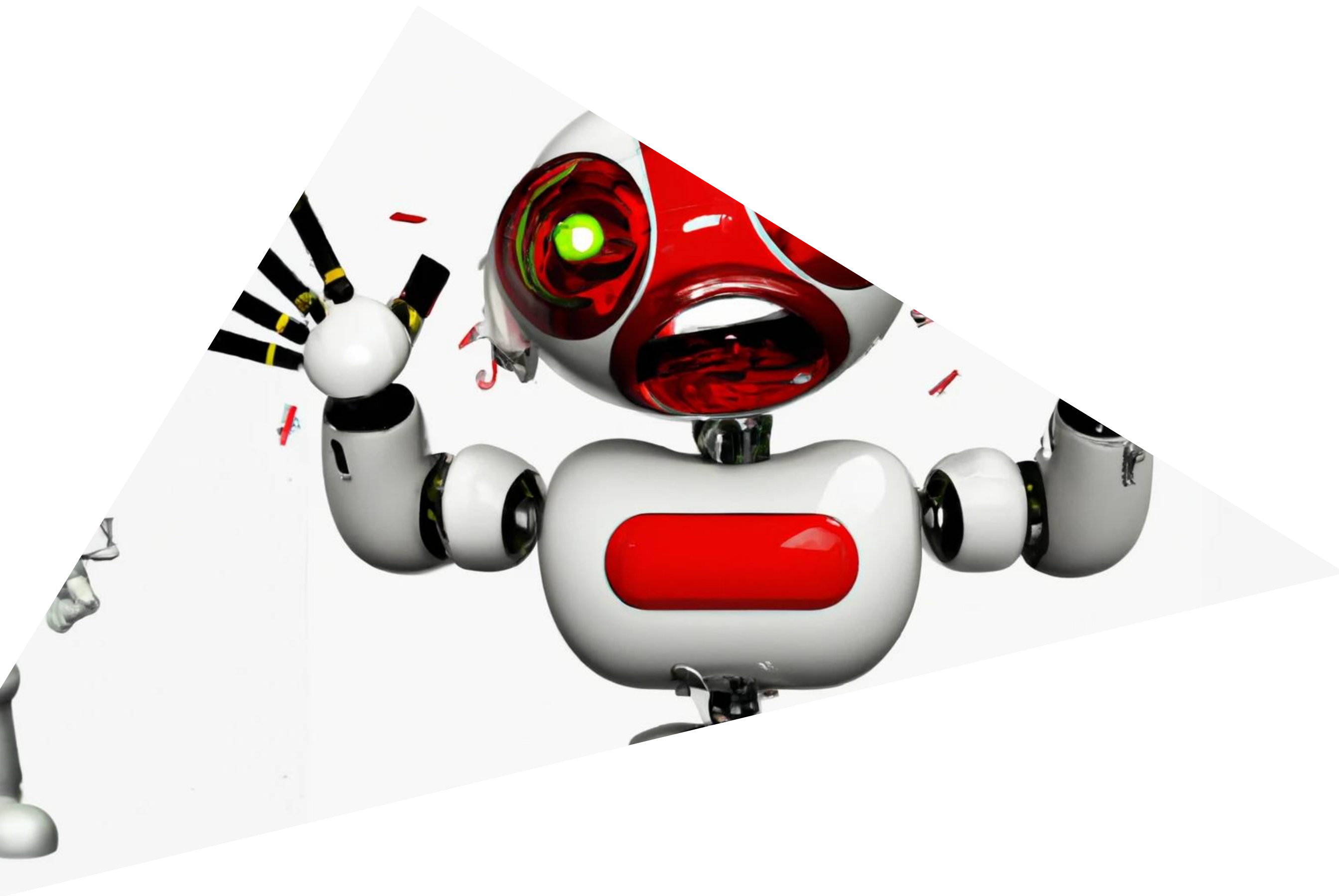


**ADVERSARIAL ATTACKS**

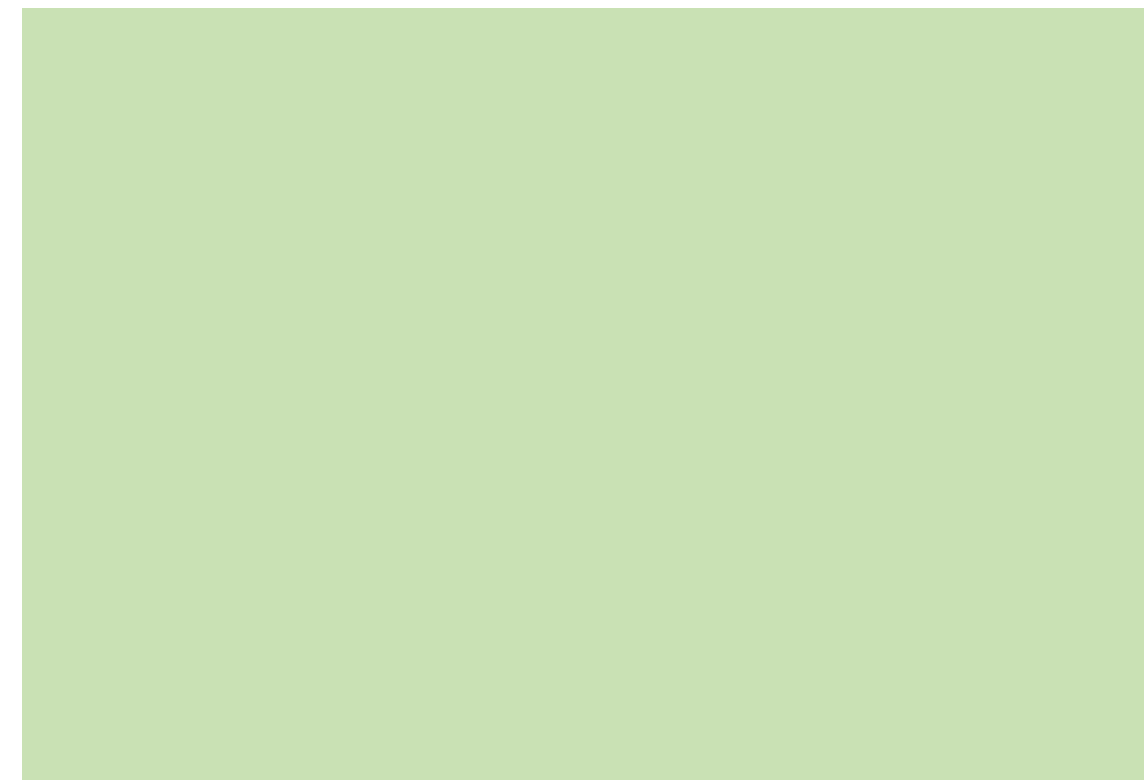


**MODEL SELECTION**





BAD AI  
SOLUTIONS  
ARE BEING  
DEVELOPED  
EVERYWHERE



Expectations



Plateau will be reached:

# Gartner Hype Cycle for Emerging Tech 2023

**HALLUCINATIONS  
ARE**

**THE PATH  
TO THE**

**TROUGH OF  
DISILLUSIONMENT**





Tread Lightly, BUT MOVE QUICKLY!

Effective use of AI will elevate your organization above  
the competition.

# THANK YOU!



**Any  
Questions?**

[cameron@zecilconsulting.com](mailto:cameron@zecilconsulting.com)

[www.zecilconsulting.com](http://www.zecilconsulting.com)