

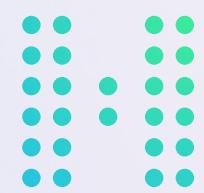
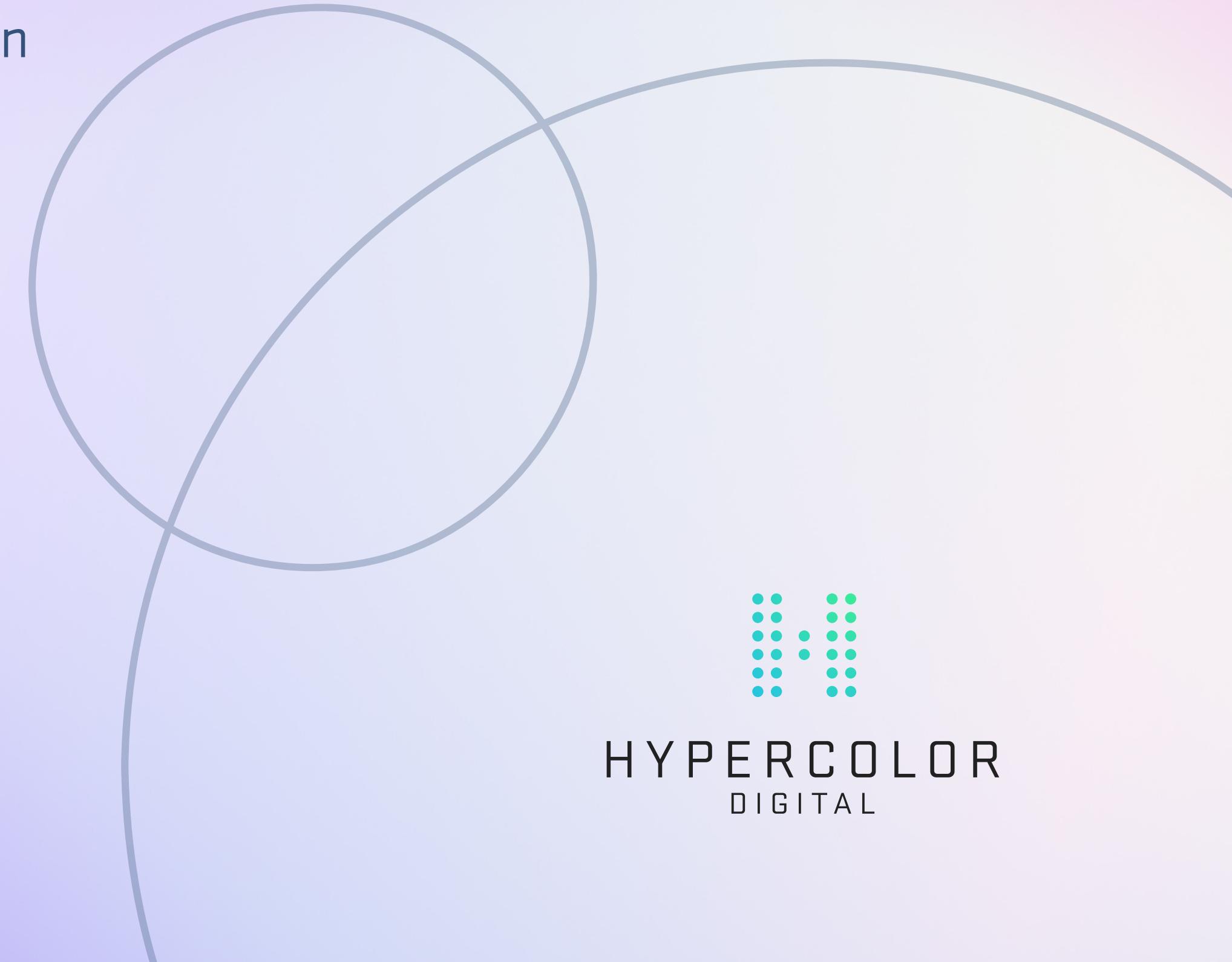
# From Generalist to Specialist

Fine-Tuning Models for Spam Detection

November 10, 2023



Emerging  
Technologies  
North



HYPERCOLOR  
DIGITAL

<https://github.com/hypercolor/applied-ai-fine-tuning>

<https://www.linkedin.com/in/andrew-aarestad>

<https://www.linkedin.com/in/seth-uschuk>

# Agenda

- Notebook 1 - Fine Tuning
- Slides
- Notebook 2 - Prompt Engineering
- Notebook 3 - Performance Testing
- Exercise



# Content Warning

<https://archive.ics.uci.edu/dataset/228/sms+spam+collection>

Dataset contains real text messages, some of which contain potentially offensive language.

“Rated R”

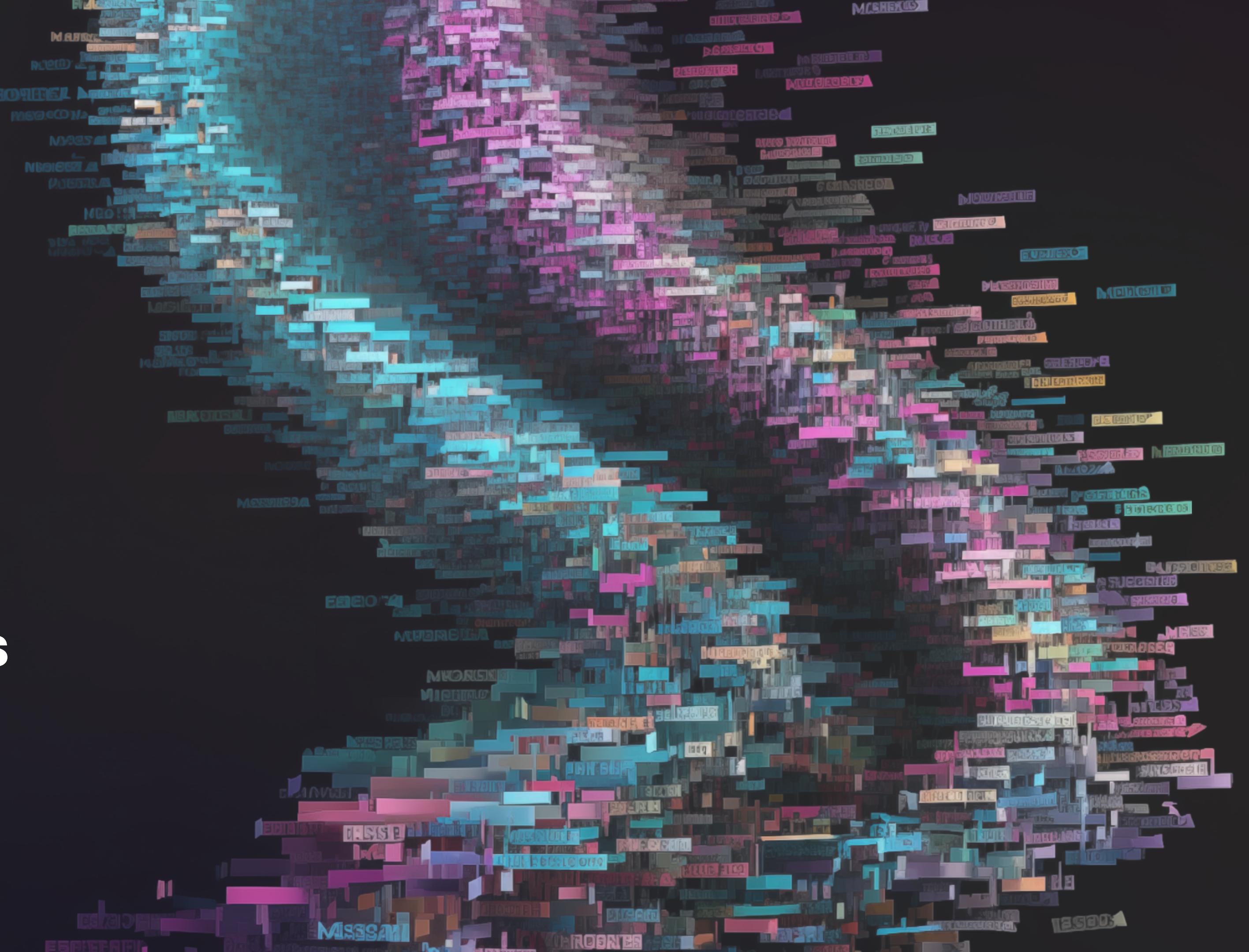
# Notebook 1 - Fine Tuning

<https://github.com/hypercolor/applied-ai-fine-tuning/blob/main/notebooks/part-1-fine-tuning.ipynb>

In this notebook we train fine-tuned models using  
the SMS spam dataset



# What Is Spam?



# Fine Tuning: “Showing”

## Spam

- You can stop further club tones by replying \"STOP MIX\\\" See my-tone.com/enjoy.html for terms. Club tones cost GBP4.50/week. MFL
- Claim a 200 shopping spree, just call 08717895698 now! Have you won! MobStoreQuiz10ppm

## Ham

- Free any day but i finish at 6 on mon n thurs...
- Lol! Nah wasn't too bad thanks. Its good to b home but its been quite a reality check. Hows ur day been? Did u do anything with website?

# Prompt Engineering: “Telling”

You are a spam classifier. Your job is to examine the content of a message and determine if it is spam or ham. Spam is a message that is spam, harmful, abusive, an attempt to access private information, harrassment, or otherwise unwanted. Ham is any message that is not classified as spam. Response should be a single word: Spam or Ham.

# When to Fine Tune?

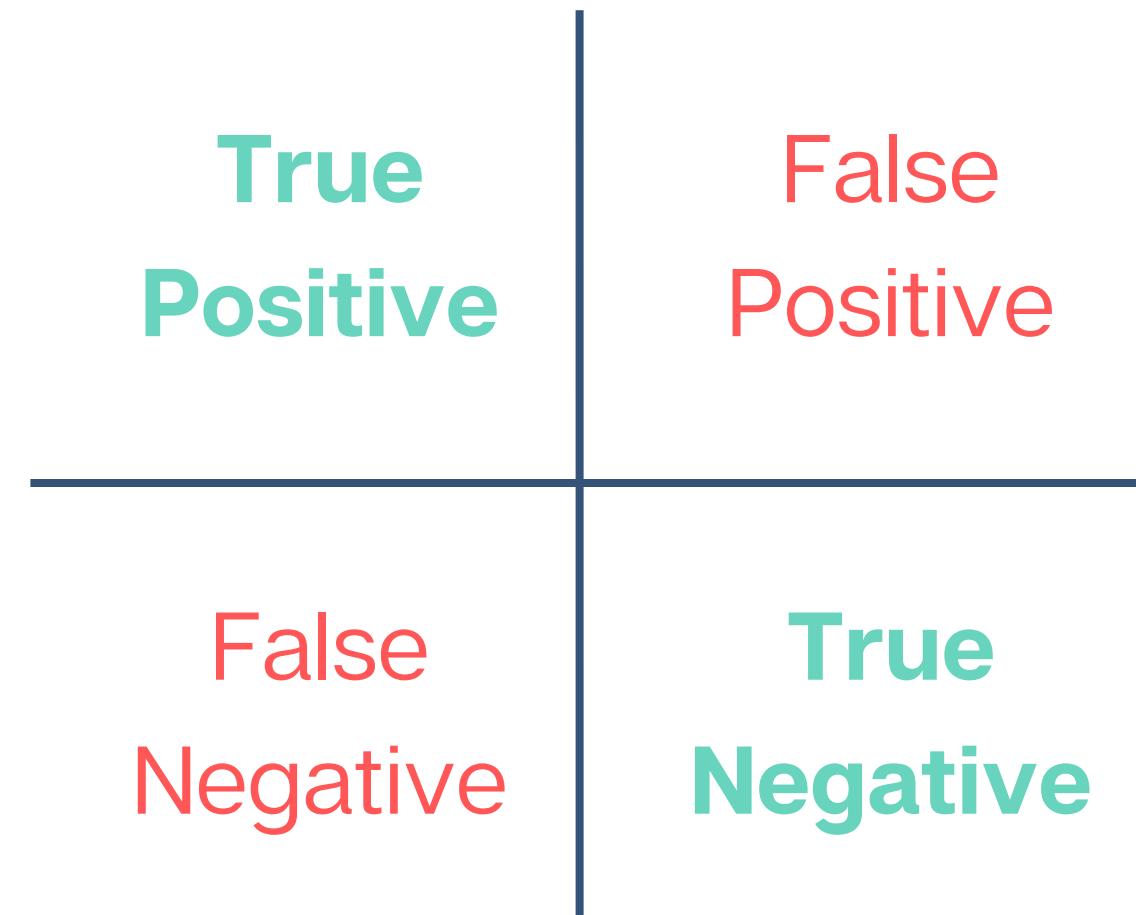
## Benefits

- Smaller Prompts
- Cost at Scale
- Handle Edge Cases  
w/Large Training Data

## Drawbacks

- Cost of Obtaining Datasets
- Development Cycle
- Extensibility to  
New Datasets

# Confusion Matrix



$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Total}$$

# Notebook 2 - Prompt Engineering

<https://github.com/hypercolor/applied-ai-fine-tuning/blob/main/notebooks/part-2-prompt-engineering.ipynb>

In this notebook, we build a simple prompt-driven spam classifier



# Notebook 3 - Performance Testing

<https://github.com/hypercolor/applied-ai-fine-tuning/blob/main/notebooks/part-3-performance-testing.ipynb>

In this notebook, we compare the performance of prompt-driven and fine-tuned spam classifiers



# Notebook 4 - Exercise

<https://github.com/hypercolor/applied-ai-fine-tuning/blob/main/notebooks/part-4-exercise.ipynb>

In this notebook, we explore extending the spam classifier model to more datasets

