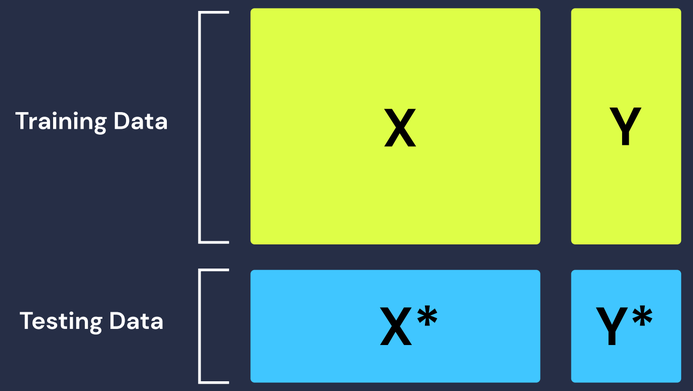
Introducing Human-Centered AI

The AI Market

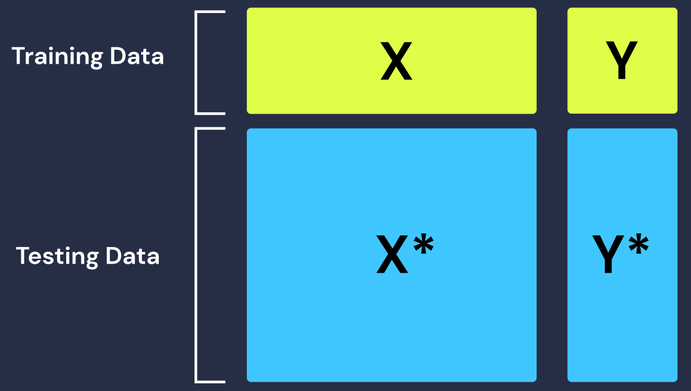
How does modern AI/ML works

there are **millions of rows of training data (X)** and **corresponding labels (Y).** Given new testing data (X*), the goal is to predict a corresponding label (Y*)

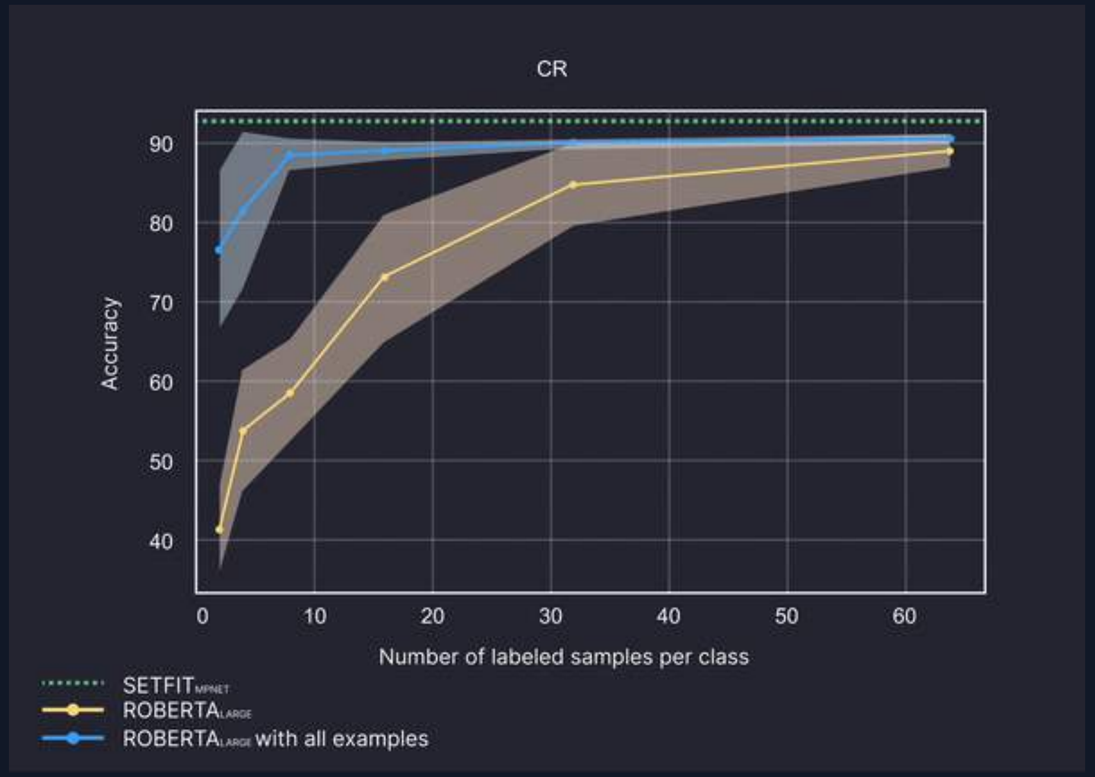


In few shot learning...

the goal is to **predict label (Y*)** with just a **few rows of training data (X)** and corresponding labels (Y). This requires novel foundation models.

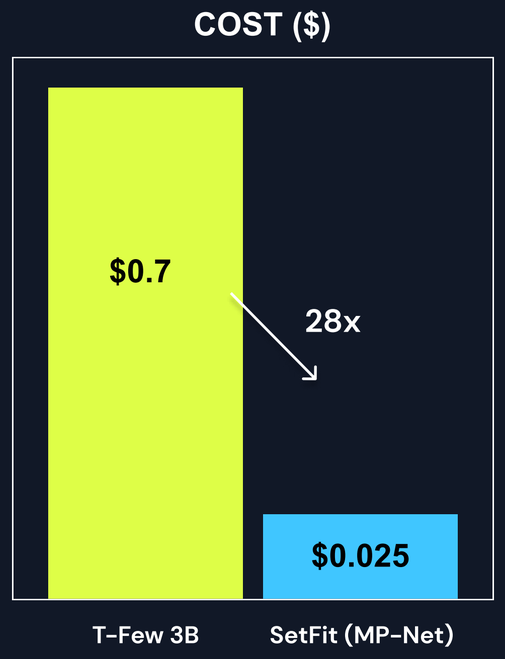


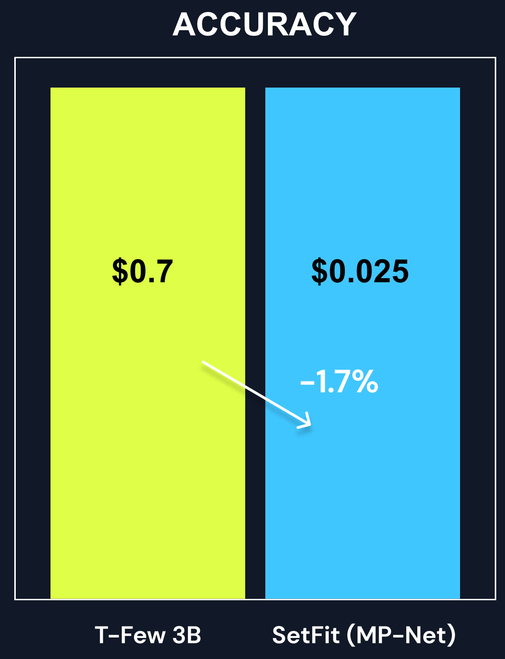
Setfit: Few Shot Learning in Practice



Result

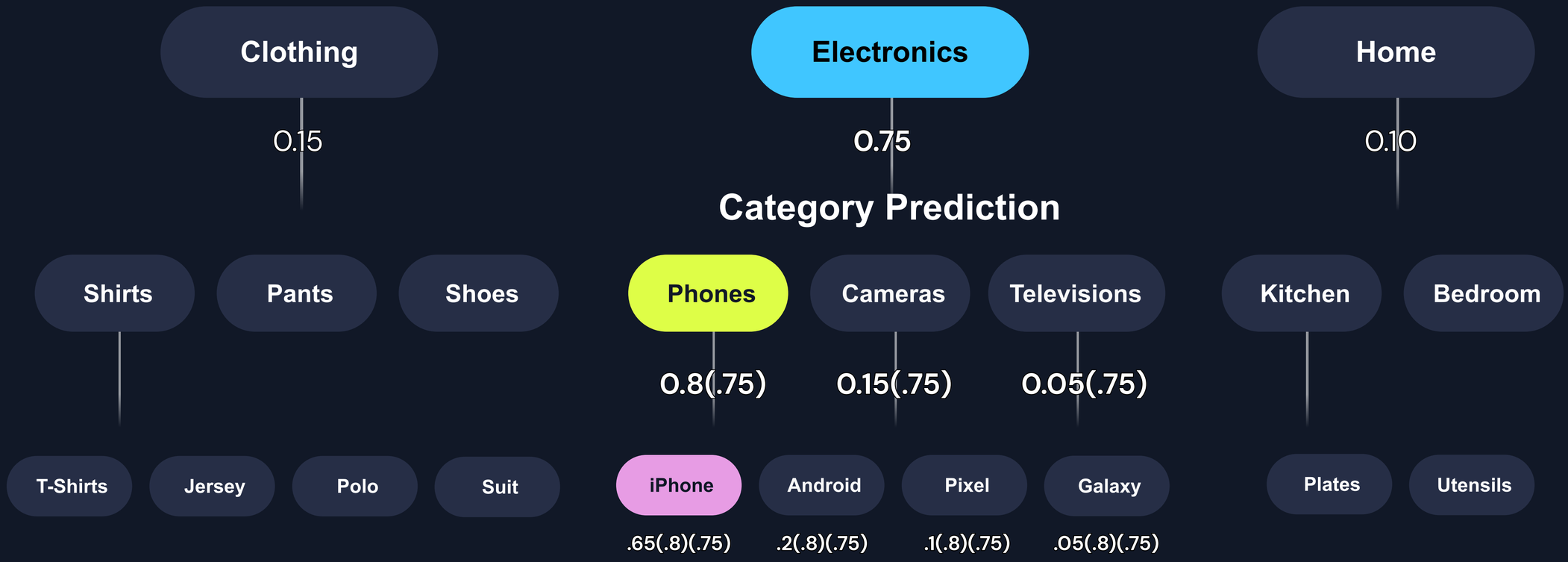
Rank	Method	Accuracy	Model Size
2	T-Few	75.8	11B
4	Human Baseline	73.5	N/A
6	SetFit (Roberta Large)	71.3	355M
9	PET	69.6	235M
11	SetFit (MP-Net)	66.9	110M
12	GPT-3	62.7	175B



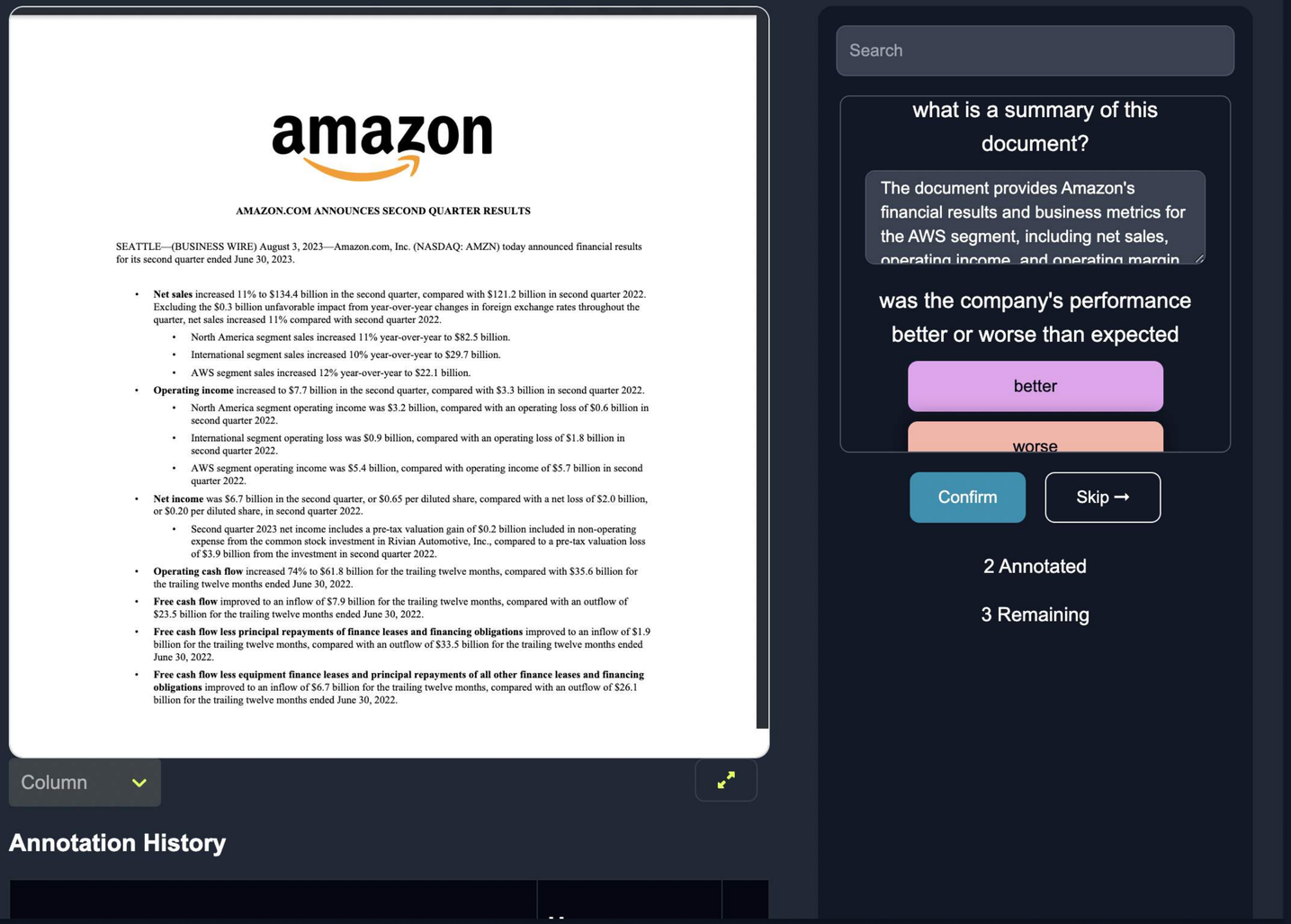


Few Shot Hierarchical Classification

I am at the Apple Store getting a **cellphone.** I admire Steve Job's vision of the **iPhone.** How can I get this **electronic** **device** at a good price?



Semi-Structured Prompts



Model Centric AI

In model-centric AI, the training data is treated as a fixed input. Your training data is something you download as a static file. New iterations of your project result from changes to the model.

Try Different Models on Static Datasets

Text Classification

Text	Label
I like my bananas	fruit
Broccoli is not good	vegetable
We like tomatoes and potatoes	vegetable

Name Entity Recognition

Text	Label
I like my bananas	I like my bananas [FRUIT]
Broccoli is not good	Broccoli [VEGETABLE] is not good
We like tomatoes and potatoes	We like tomatoes [FRUIT] **and potatoes** [VEGETABLE] is not good

Data Centric AI

In data-centric AI, data quality and quantity is increasingly the key to successful results. Teams spend more time on labeling, managing, slicing, augmenting, and curating the data, with the model itself relatively more fixed. In data centric AI, you programmatically iterate on your training data.

Try different datasets on static model

Text Classification

Text	Label
I like my bananas	fruit
Broccoli is not good	vegetable
We like tomatoes and cake	dessert
Cookies and milk are great	dessert

Name Entity Recognition

Text	Label
I like my bananas	I like my bananas [FRUIT]
Broccoli is not good	Broccoli [VEGETABLE] is not good
We like tomatoes and cake	We like tomatoes [FRUIT] **and cake** [DESSERT]
Cookies and milk are great	Cookies [DESSERT] and **cake** [DESSERT] are great

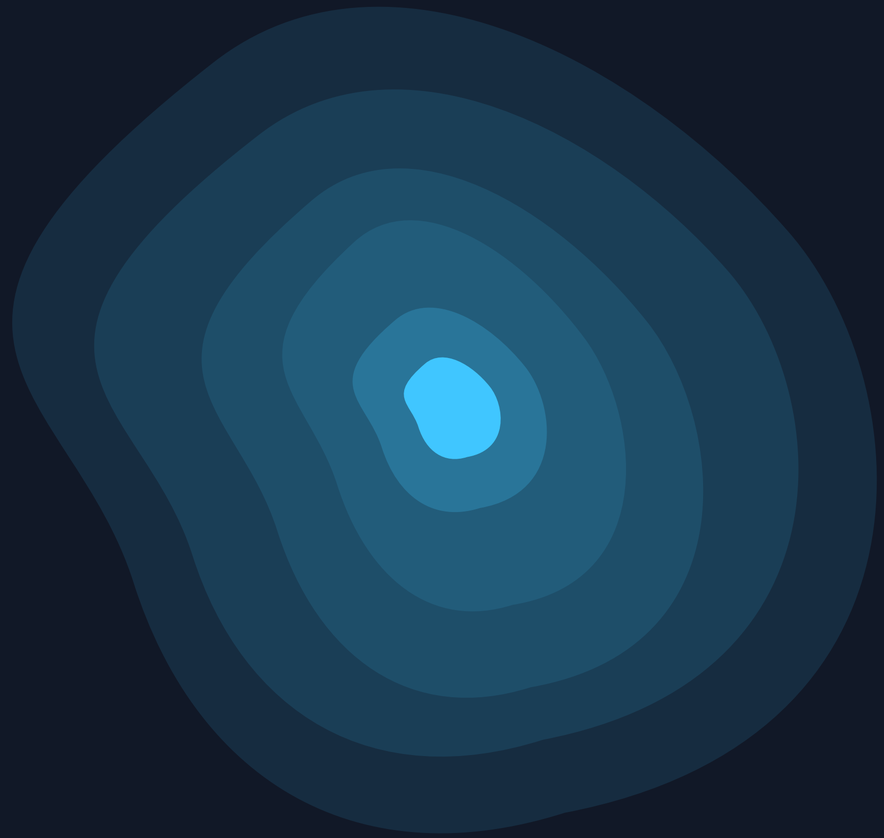
The “Scale”-ability Issue

Manual Labeling

As we label data, new classes of data are added due to changing business requirements. Oftentimes, this requires manually relabeling all of the data again, from scratch.

Progammatic Labeling

A solution that works in a specific vertical (i.e. finance) may not translate well to a different vertical (i.e. healthcare). Many labeling applications can't be solved programmatically



Less than 1/1000th of data that businesses require to be labeled is currently labeled today. As **Carl Sagan** said, we are just a pale blue dot.

A New Way of Doing AI

As we iterate on our data, we also iterate on our ensemble model
The ensemble model output is a function of the data input As we label more, the ensemble model learns which local model is best

Try different datasets on static model

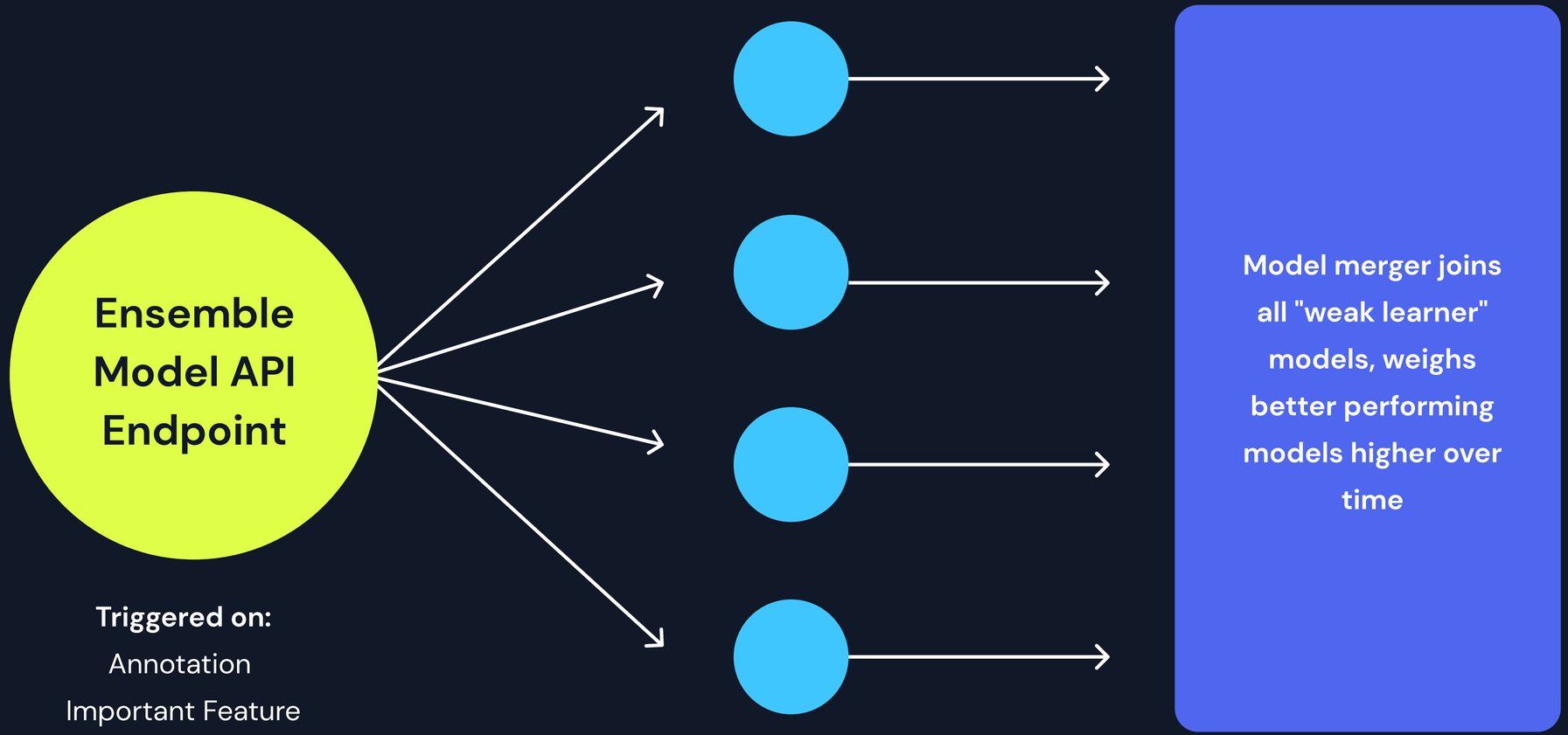
Text Classification

Text	Label
I like my bananas	fruit
Broccoli is not good	vegetable
We like tomatoes and cake	dessert
Cookies and milk are great	dessert

Name Entity Recognition

Entity_Text	Entity_Label
I like my bananas	I like my bananas [FRUIT]
Broccoli is not good	Broccoli [VEGETABLE] is not good
We like tomatoes and cake	We like tomatoes [FRUIT] and cake [DESSERT]
Cookies and milk are great	Cookies [DESSERT] and cake [BEVERAGE] are great

Enabling a Scalable Data Labeling Solution



Triggered on:

Annotation

Important Feature

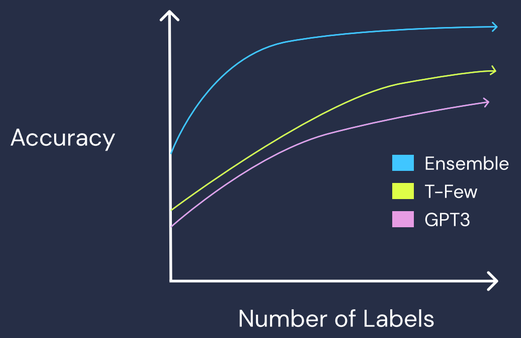
Category Added

Each "weak model" runs asynchronously, though to the user they appear to run instantaneously

Features

Few Shot Learning

After you label just a few row of data (the edge cases), we label the rest, using state of the art transformer models



Programmatic Labeling

Heuristics such as key words, entities and regex expressions are fed into the model

Decomposition

Convert unstructured text data to a spreadsheet

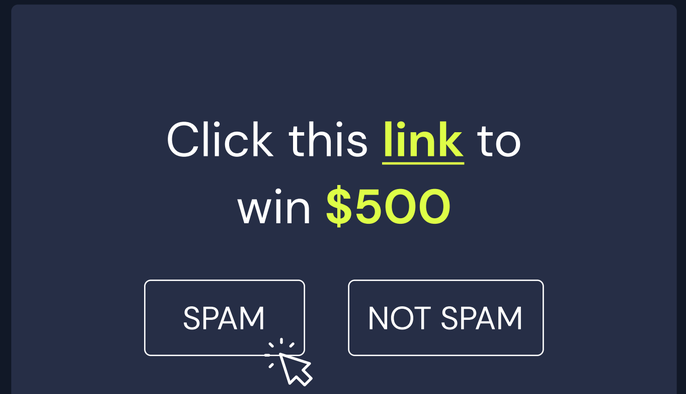
Synchronous

Get your data labeled in real time, and receive immediate feedback on the model's performance as you label data

Index	Text	Label	Probability
0	Click this link to win $500	SPAM	.98
1	This is a normal sentence	NOT SPAM	.89
2	You just won a free iPhone	SPAM	.93
3	Hope you are doing well	NOT SPAM	.87
4	Enter your credit card info	SPAM	.97

Human in the Loop

We actively learn from subject matter experts, who provide input into our model



Contextual

Our large language models extract the context of words in sentences, not just the individual words themselves