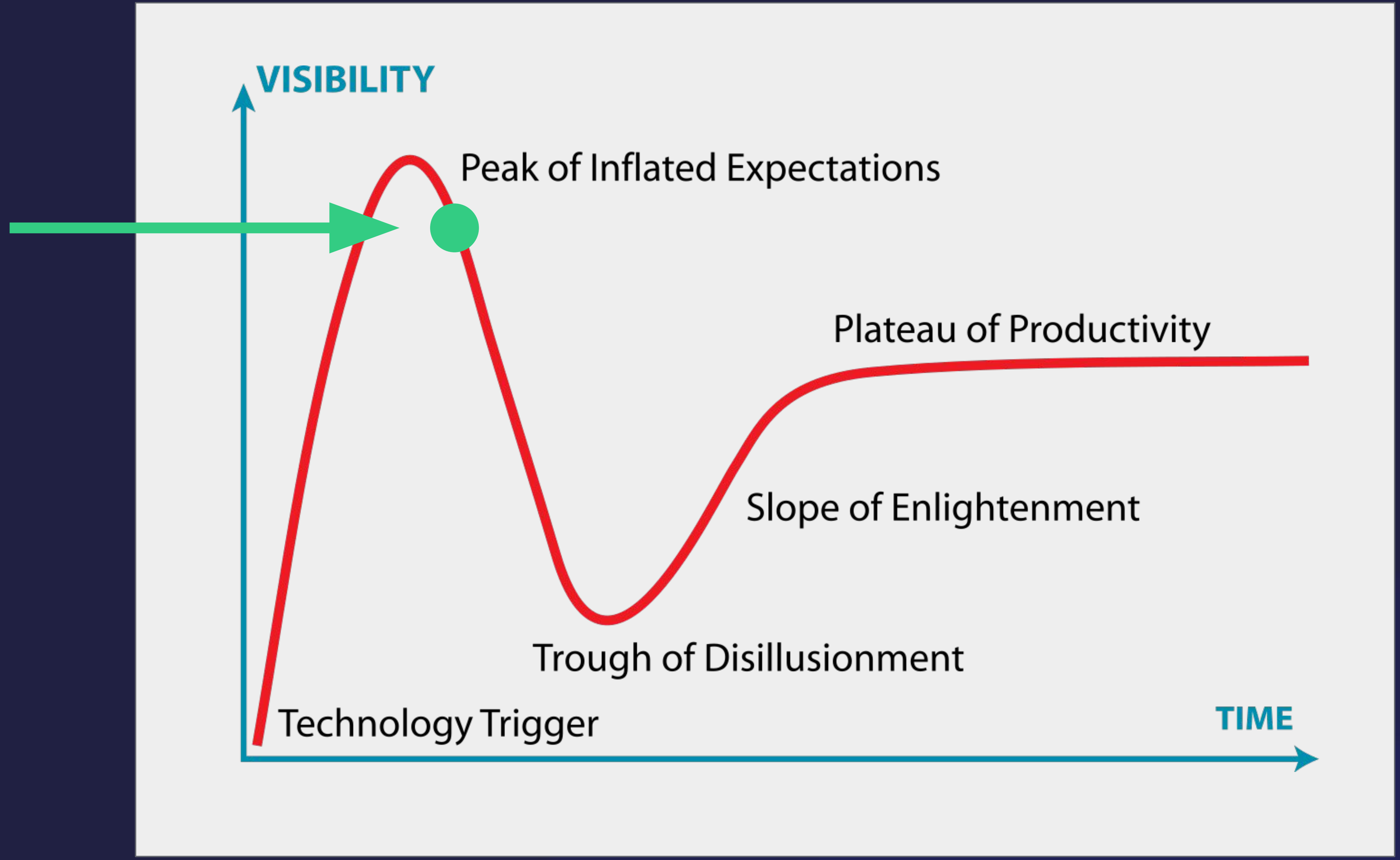
Snorkel

Snorkel

# Enterprise AI:
# All about turning *your* data into *your* AI

# The State of AI Today: From demo to production



We are here

**Enterprise AI teams are beginning to understand that FMs are a fundamental breakthrough- as a “first mile” technology**

# Why “foundation model”?

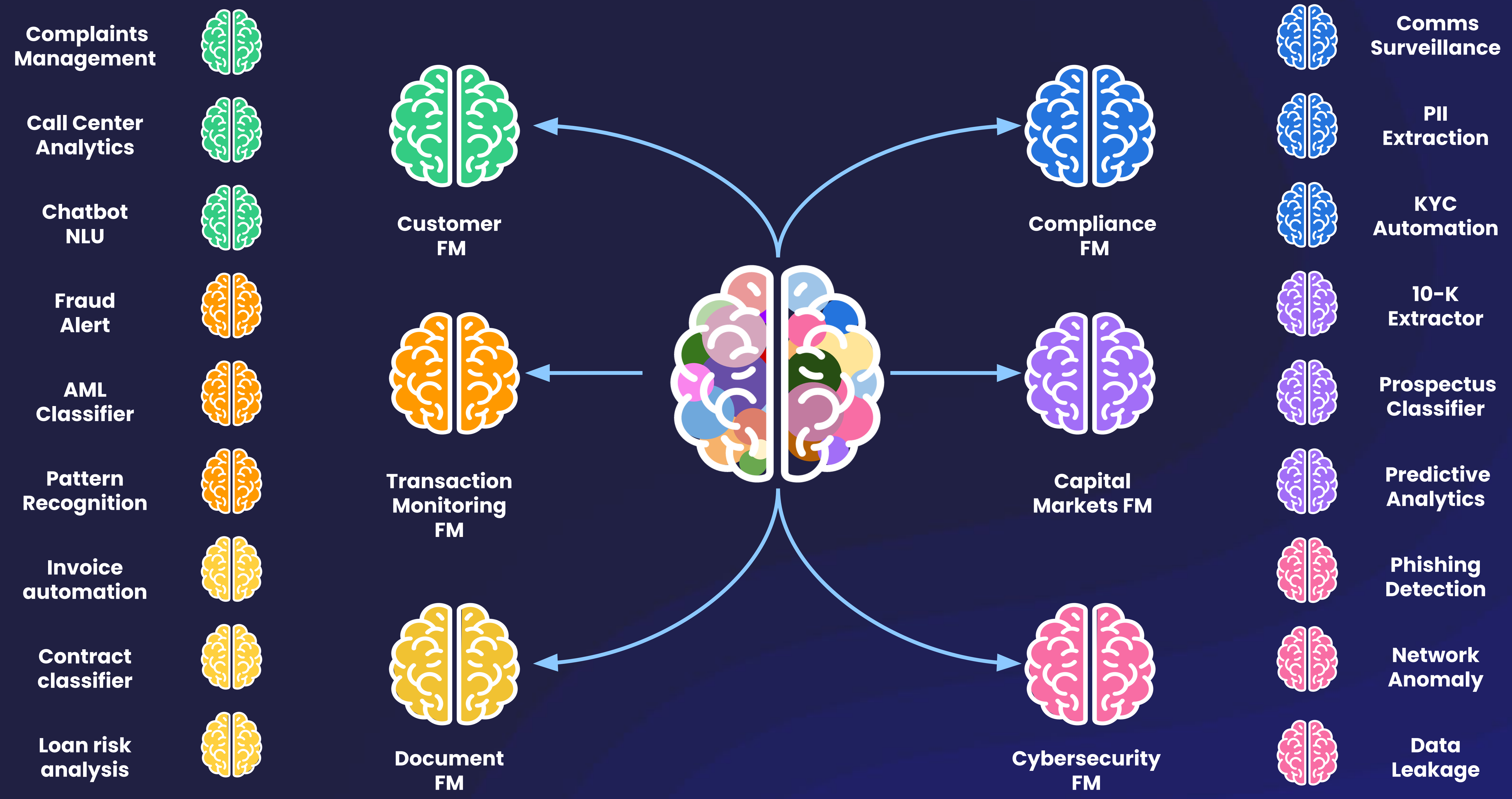
- Not just language- **all data types**
- Not just generative use cases- **all AI application types**
- Foundation models are foundations- **still need to build the house!**

# Key thesis: GPT-You, not GPT-X

Foundation models are *all about the data*

Successful enterprises will develop *their own* FMs leveraging their own data and knowledge

# The future of enterprise AI is specialized or hybrid AI



# AI today: Promising prototype, not production ready

**In a survey of 19 academic papers (151 tasks), prompted ChatGPT underperformed existing fine-tuned baselines on**

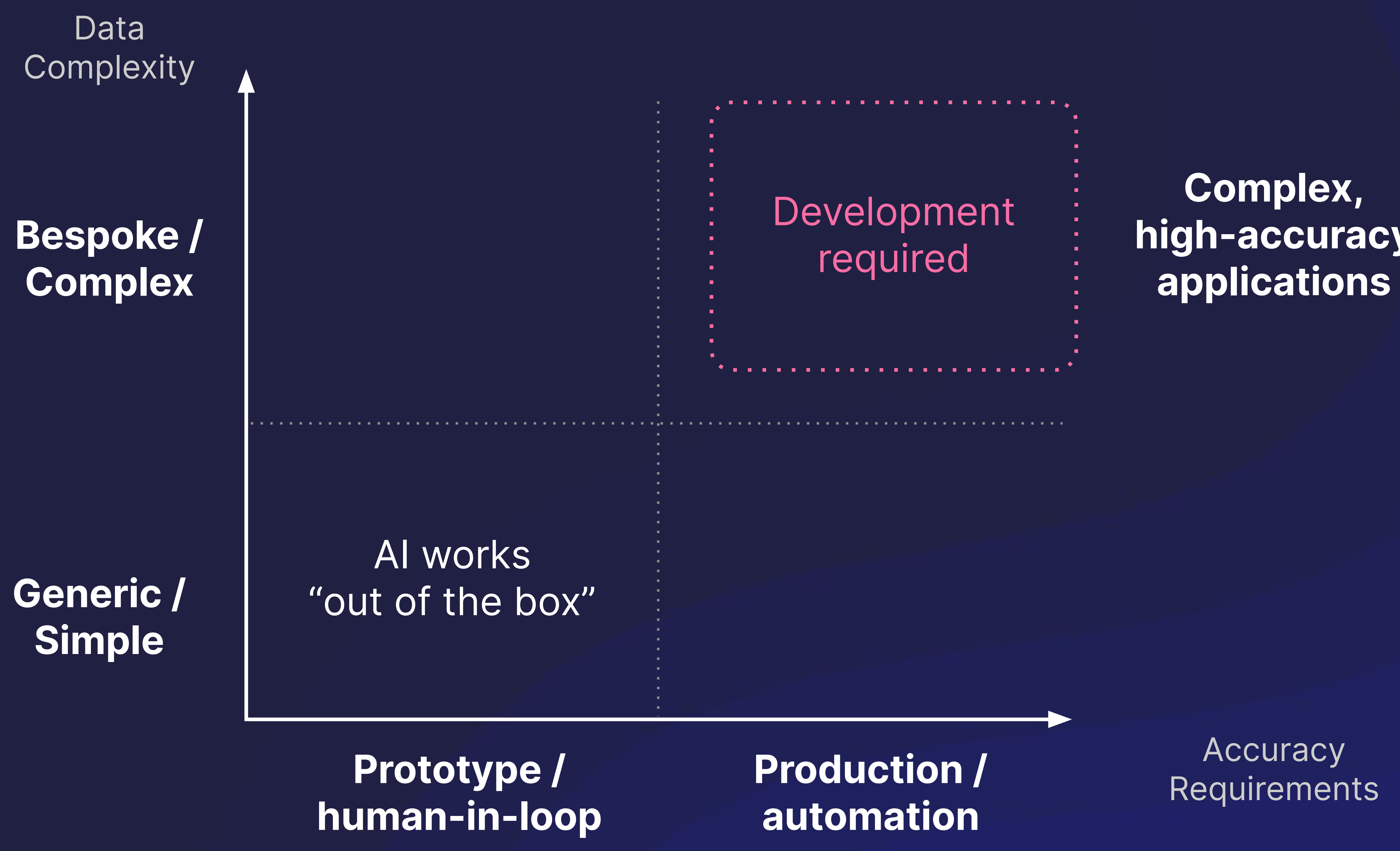
# 77.5%

**of tasks**

Author concluded that: *"....ChatGPT will be used as a quick prototype for some applications, but it will be replaced by a fine-tuned model (often smaller, for economical reasons) for most production-ready solutions..."*

**"ChatGPT Survey: Performance on NLP datasets" (Pikuliak, 2023)**

# Some use cases *will* work “out of the box”- but the hardest & most value-aligned will not



# The new (and only viable) way:
# Data-centric development

**Training Data**

**Iteratively label & develop the data**

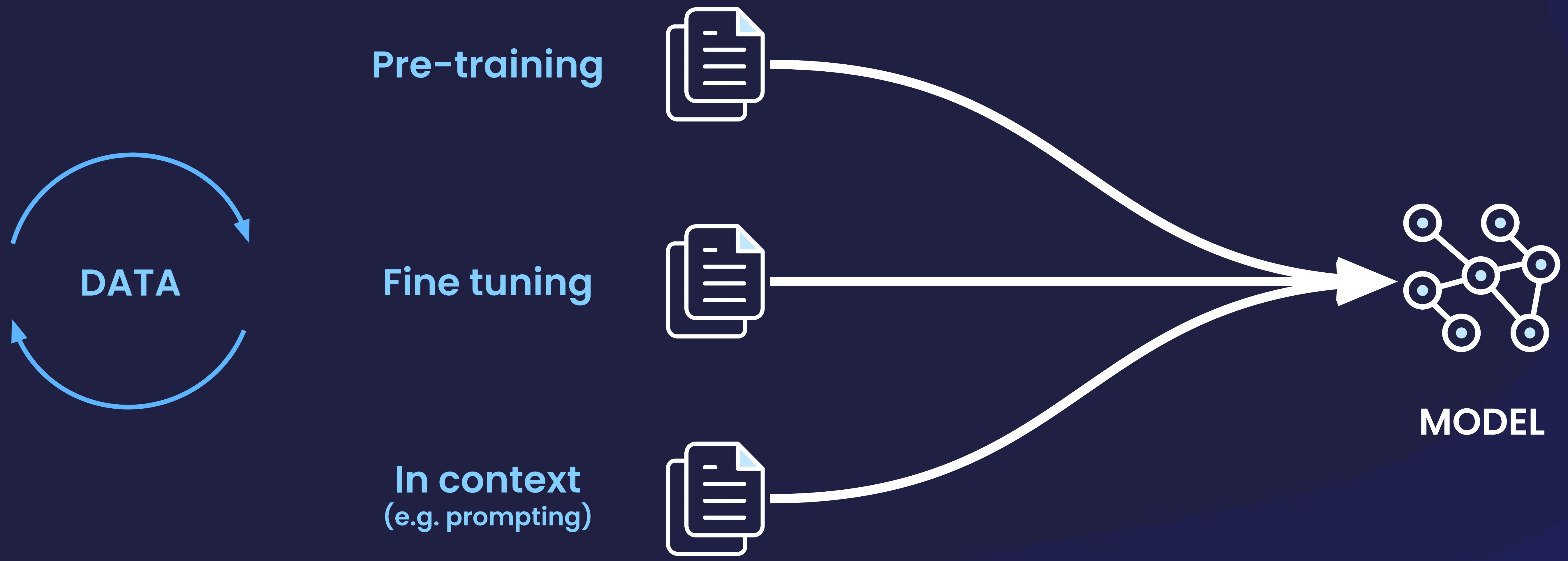
**Models**

Fine-tune standard models (e.g. transformer [2017])

**You can't "fix" errors by tweaking individual model parameters.**

**Data+supervision is the only interface**

# To do this: Data must be developed in use case-specific ways



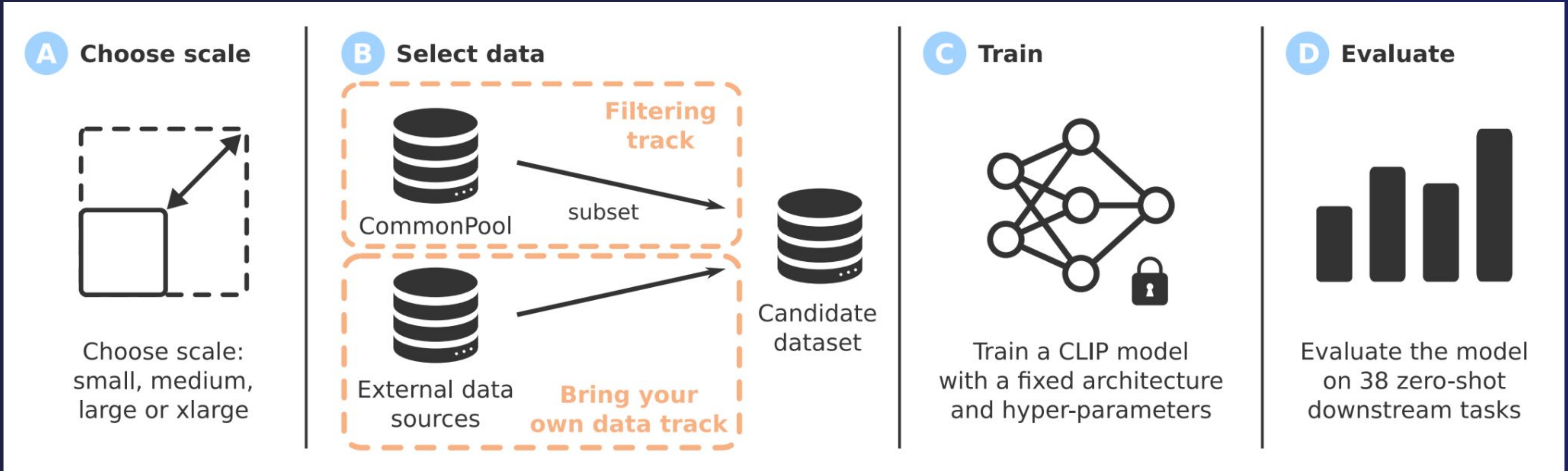
Our mission: Make this data development first-class and programmatic, like any other type of  software development

# Data development is more than labeling

| | |
|---|---|
| **Labeling** | Assigning target model outputs for input examples |
| **Cleaning** | Standardizing and formatting examples |
| **Slicing** | Tagging relevant subsets of examples |
| **Sampling** | Adjusting the distribution of examples |
| **Filtering** | Removing low quality examples |
| **Augmenting** | Creating new examples |

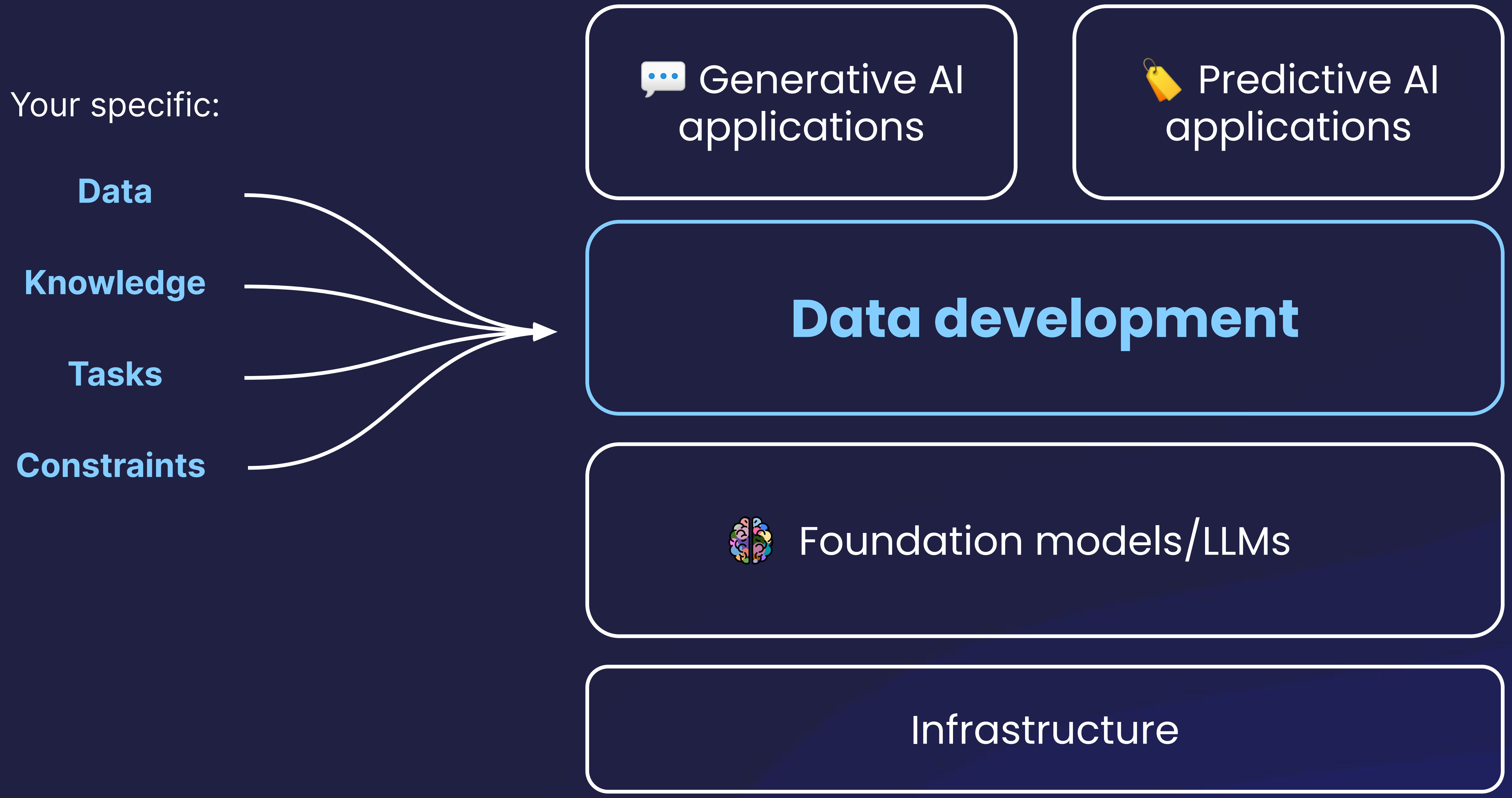
In the right framework, data development goes from **ad hoc preprocessing hacks** to a scalable and **systematic software-based process**!

# Example: DataComp


A Choose scale
Choose scale: small, medium, large or xlarge
B Select data
Filtering track
CommonPool
subset
External data sources
Bring your own data track
Candidate dataset
C Train
Train a CLIP model with a fixed architecture and hyper-parameters
D Evaluate
Evaluate the model on 38 zero-shot downstream tasks

**Key idea: If you hold everything but the pre-training dataset fixed, you get a new SOTA on CLIP!**

# Production-level AI applications need to be developed for *your* tasks using *your* data



Your specific:

Data

Knowledge

Tasks

Constraints

💬 Generative AI applications

🏷️ Predictive AI applications

**Data development**

🧠 Foundation models/LLMs

Infrastructure

# Data labeling and development is manual, slow, and expensive

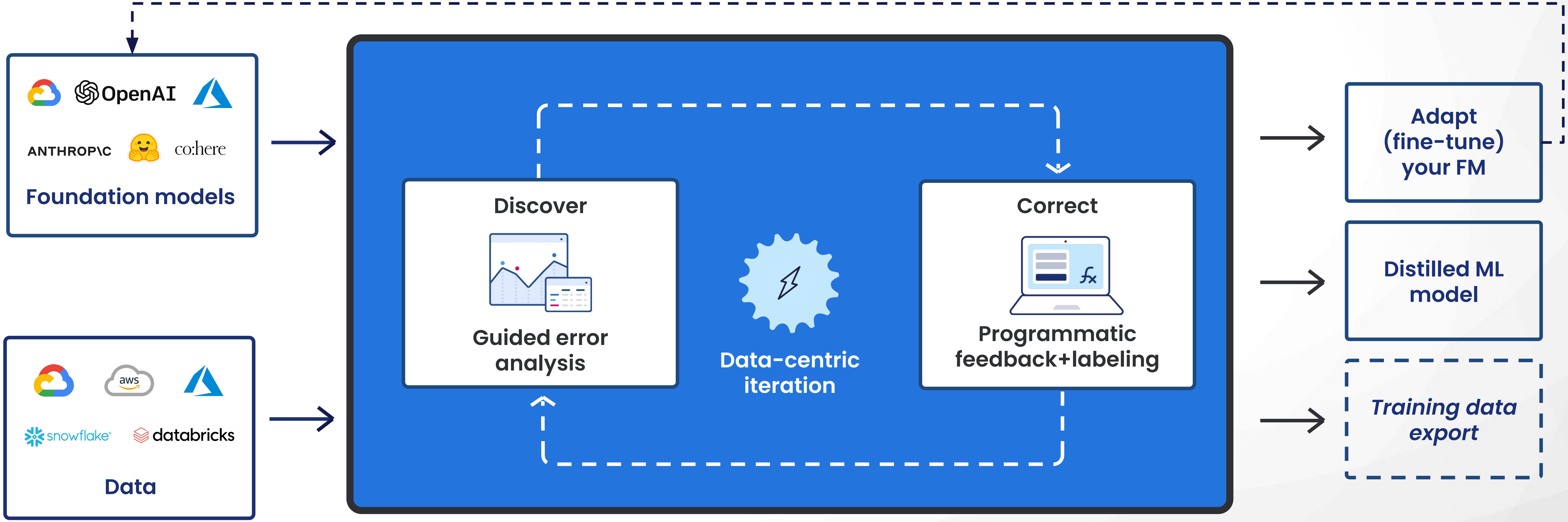
Whether outsourced...

...or labeled in-house

Manual labeling unscalable for data that is
private, fast-changing, and/or requires subject matter expertise

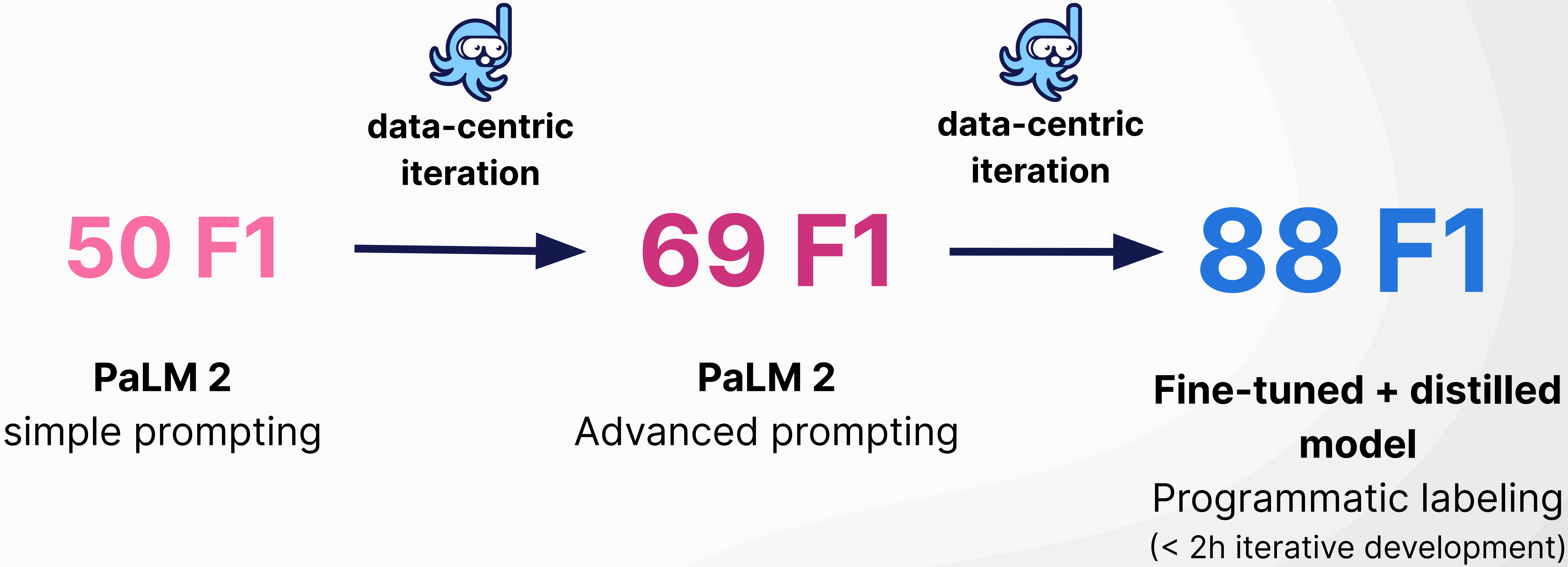
# Snorkel Flow

## The data development platform for training & fine-tuning AI



**10-100x+ faster *programmatic* data labeling & development for rapidly fine-tuning and customizing AI**

# Case study: Bank chat classification with Google PaLM 2



50 F1

**PaLM 2**
simple prompting

data-centric iteration

69 F1

**PaLM 2**
Advanced prompting

data-centric iteration

88 F1

**Fine-tuned + distilled model**
Programmatic labeling
(< 2h iterative development)

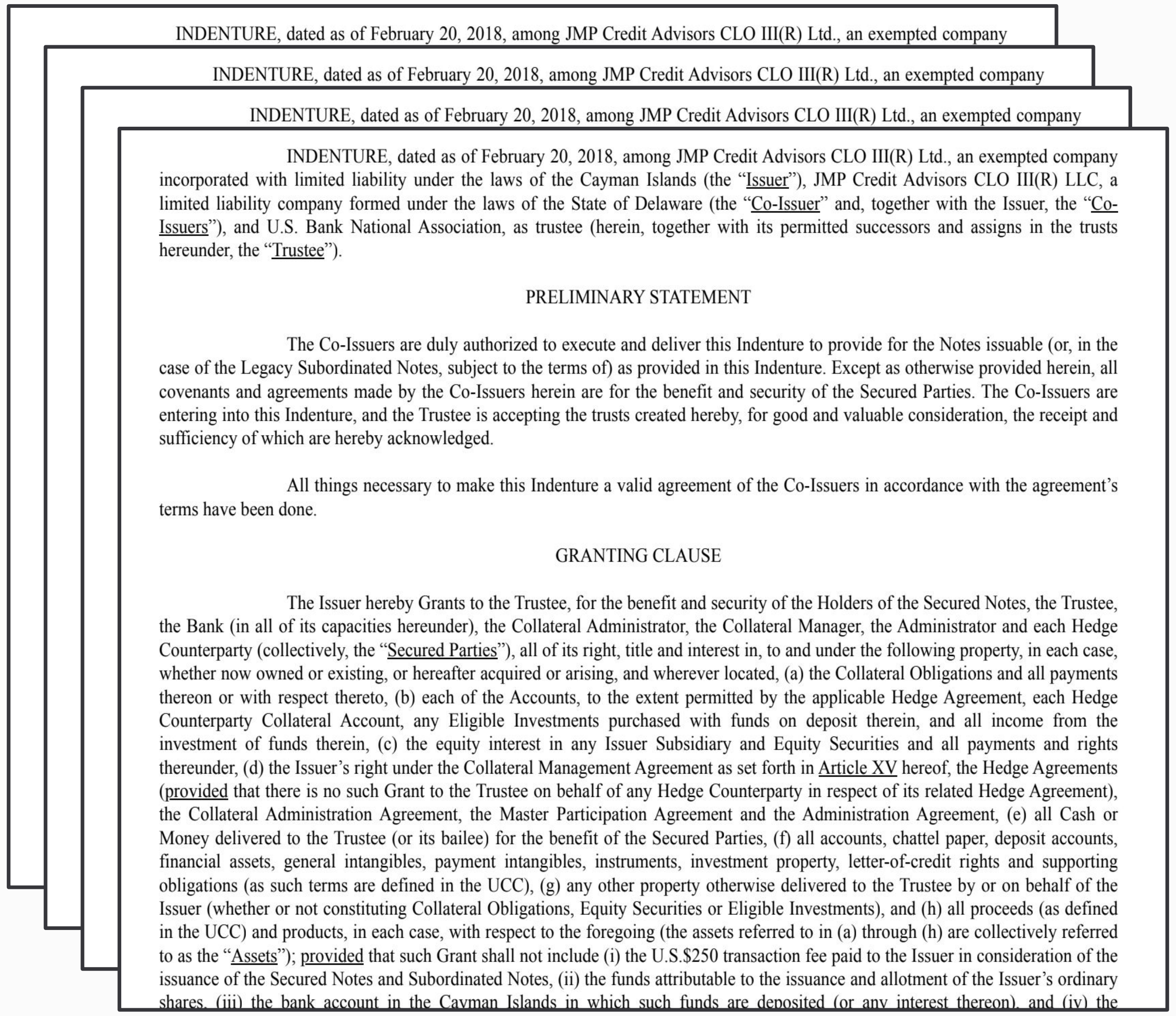
**Snorkel Flow builds on top of foundation models, enabling rapid tuning + distilling for production use**

# (Bigger) Key idea:

# Programmatic data development is critical for tuning all parts of an LLM system

# Top-10 US bank: Full system fine-tuning and data development for CLO document Q&A

## Unstructured CLO Documents



INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company incorporated with limited liability under the laws of the Cayman Islands (the "Issuer"), JMP Credit Advisors CLO III(R) LLC, a limited liability company formed under the laws of the State of Delaware (the "Co-Issuer" and, together with the Issuer, the "Co-Issuers"), and U.S. Bank National Association, as trustee (herein, together with its permitted successors and assigns in the trusts hereunder, the "Trustee").

PRELIMINARY STATEMENT

The Co-Issuers are duly authorized to execute and deliver this Indenture to provide for the Notes issuable (or, in the case of the Legacy Subordinated Notes, subject to the terms of) as provided in this Indenture. Except as otherwise provided herein, all covenants and agreements made by the Co-Issuers herein are for the benefit and security of the Secured Parties. The Co-Issuers are entering into this Indenture, and the Trustee is accepting the trusts created hereby, for good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged.

All things necessary to make this Indenture a valid agreement of the Co-Issuers in accordance with the agreement's terms have been done.

GRANTING CLAUSE

The Issuer hereby Grants to the Trustee, for the benefit and security of the Holders of the Secured Notes, the Trustee, the Bank (in all of its capacities hereunder), the Collateral Administrator, the Collateral Manager, the Administrator and each Hedge Counterparty (collectively, the "Secured Parties"), all of its right, title and interest in, to and under the following property, in each case, whether now owned or existing, or hereafter acquired or arising, and wherever located, (a) the Collateral Obligations and all payments thereon or with respect thereto, (b) each of the Accounts, to the extent permitted by the applicable Hedge Agreement, each Hedge Counterparty Collateral Account, any Eligible Investments purchased with funds on deposit therein, and all income from the investment of funds therein, (c) the equity interest in any Issuer Subsidiary and Equity Securities and all payments and rights thereunder, (d) the Issuer's right under the Collateral Management Agreement as set forth in Article XV hereof, the Hedge Agreements (provided that there is no such Grant to the Trustee on behalf of any Hedge Counterparty in respect of its related Hedge Agreement), the Collateral Administration Agreement, the Master Participation Agreement and the Administration Agreement, (e) all Cash or Money delivered to the Trustee (or its bailee) for the benefit of the Secured Parties, (f) all accounts, chattel paper, deposit accounts, financial assets, general intangibles, payment intangibles, instruments, investment property, letter-of-credit rights and supporting obligations (as such terms are defined in the UCC), (g) any other property otherwise delivered to the Trustee by or on behalf of the Issuer (whether or not constituting Collateral Obligations, Equity Securities or Eligible Investments), and (h) all proceeds (as defined in the UCC) and products, in each case, with respect to the foregoing (the assets referred to in (a) through (h) are collectively referred to as the "Assets"); provided that such Grant shall not include (i) the U.S.$250 transaction fee paid to the Issuer in consideration of the issuance of the Secured Notes and Subordinated Notes, (ii) the funds attributable to the issuance and allotment of the Issuer's ordinary shares, (iii) the bank account in the Cayman Islands in which such funds are deposited (or any interest thereon), and (iv) the

→

**25% Accurate**

With out-of-the-box LLM + RAG

*Failed to answer hardest questions*

→

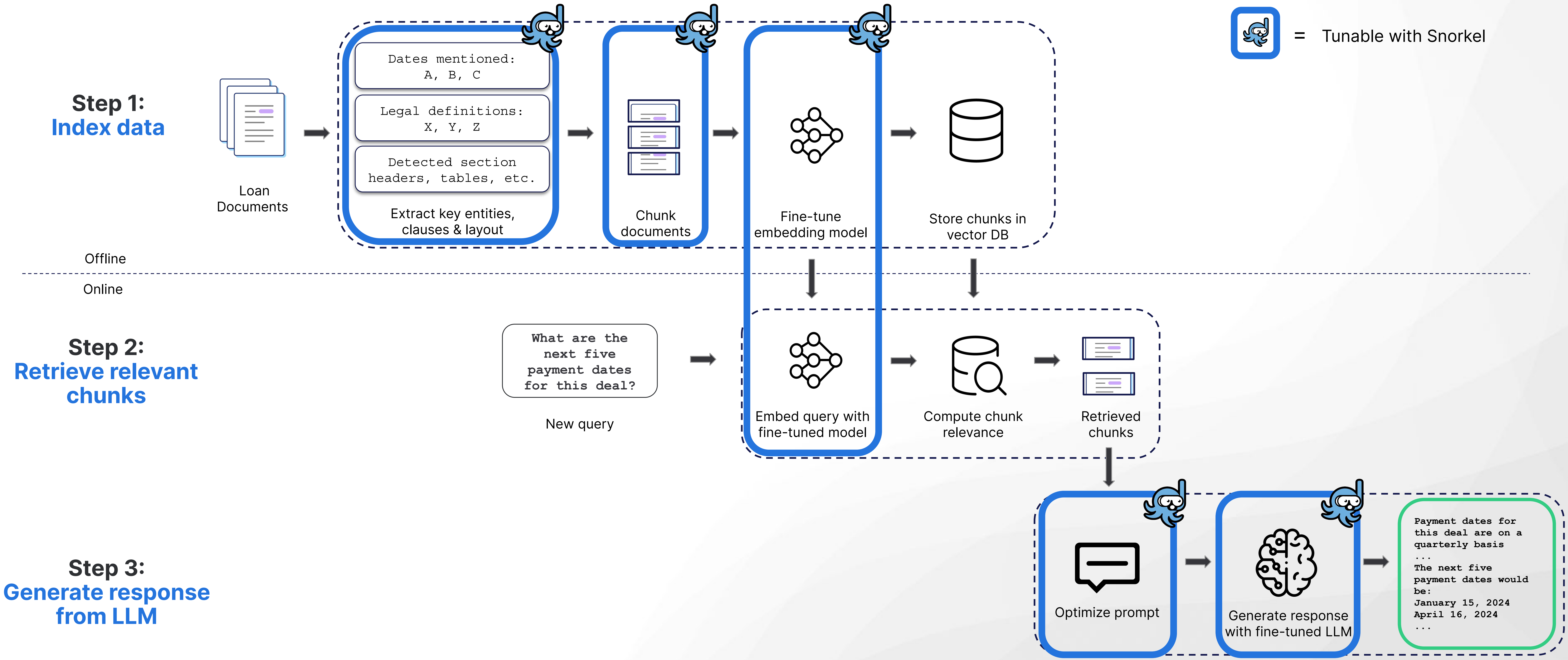
in 3 weeks

**79% Accurate**

With fine-tuning + other data development

**LLMs do not work out of the box for enterprise use cases - development is required**

# A Retrieval-augmented LLM with fine-tunable components



= Tunable with Snorkel

**Step 1:**
**Index data**

Loan Documents

Dates mentioned:
A, B, C

Legal definitions:
X, Y, Z

Detected section headers, tables, etc.

Extract key entities, clauses & layout

Chunk documents

Fine-tune embedding model

Store chunks in vector DB

Offline

Online

**Step 2:**
**Retrieve relevant chunks**

What are the next five payment dates for this deal?

New query

Embed query with fine-tuned model

Compute chunk relevance

Retrieved chunks

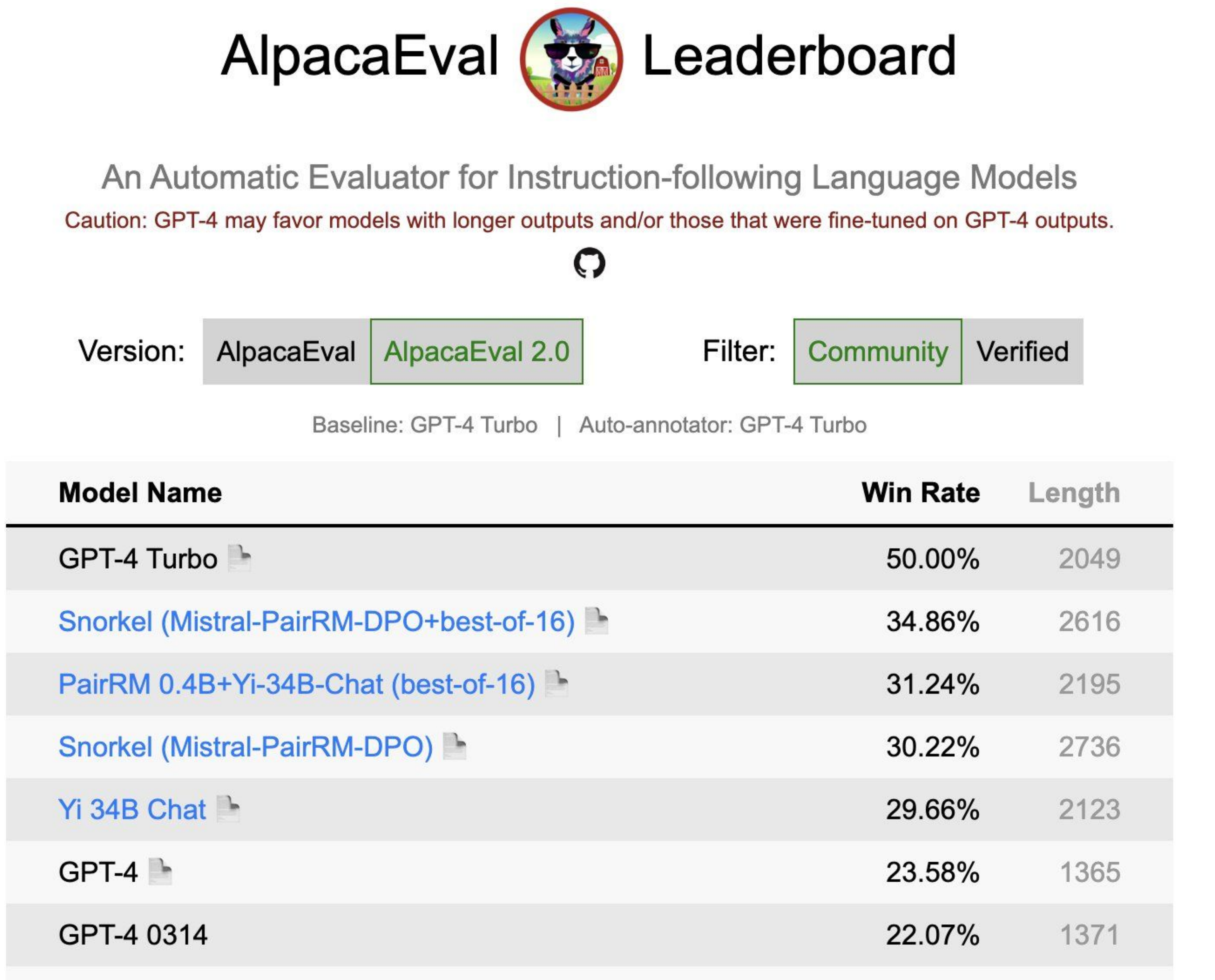
**Step 3:**
**Generate response from LLM**

Optimize prompt

Generate response with fine-tuned LLM

Payment dates for this deal are on a quarterly basis
...
The next five payment dates would be:
January 15, 2024
April 16, 2024
...

# Latest results: programmatic alignment



## AlpacaEval Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Caution: GPT-4 may favor models with longer outputs and/or those that were fine-tuned on GPT-4 outputs.

Version: AlpacaEval | AlpacaEval 2.0 Filter: Community | Verified

Baseline: GPT-4 Turbo | Auto-annotator: GPT-4 Turbo

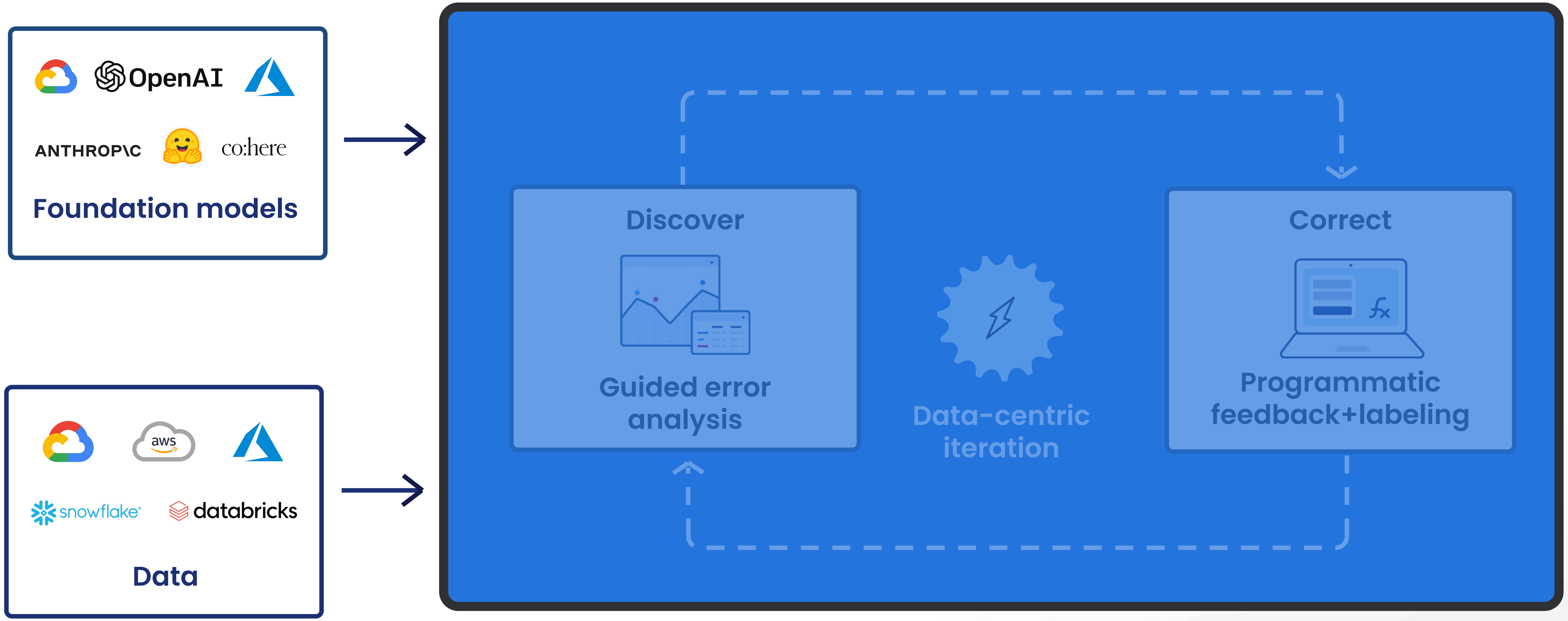
| Model Name | Win Rate | Length |
|---|---|---|
| GPT-4 Turbo | 50.00% | 2049 |
| Snorkel (Mistral-PairRM-DPO+best-of-16) | 34.86% | 2616 |
| PairRM 0.4B+Yi-34B-Chat (best-of-16) | 31.24% | 2195 |
| Snorkel (Mistral-PairRM-DPO) | 30.22% | 2736 |
| Yi 34B Chat | 29.66% | 2123 |
| GPT-4 | 23.58% | 1365 |
| GPT-4 0314 | 22.07% | 1371 |

**Key idea:** Rapidly develop custom reward models with programmatic data development- use to steer LLMs without manual annotation!

# Walking through a data-centric workflow in Snorkel Flow

# Snorkel Flow

## The data development platform for training & fine-tuning AI



Foundation models

Data

Discover

Guided error analysis

Data-centric iteration

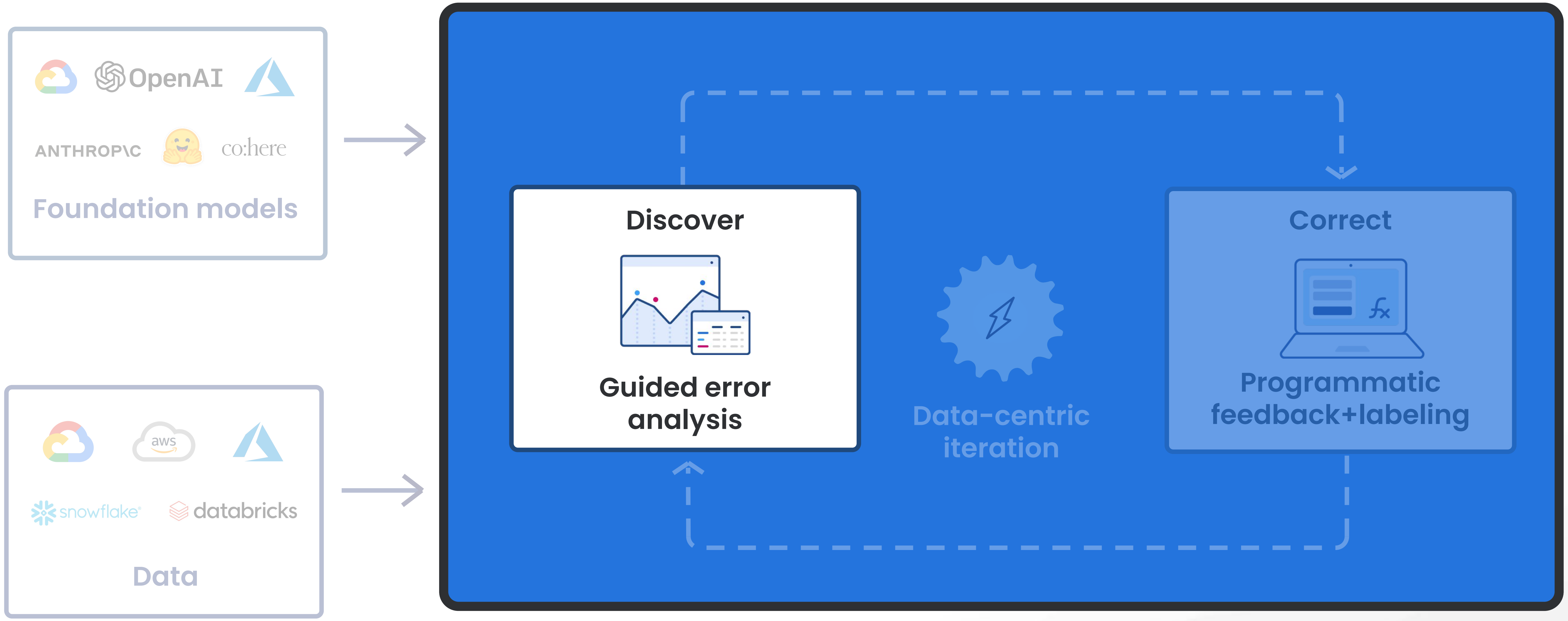
Correct

Programmatic feedback+labeling

**Start:** Define your task, connect your data, and select your model

# Snorkel Flow

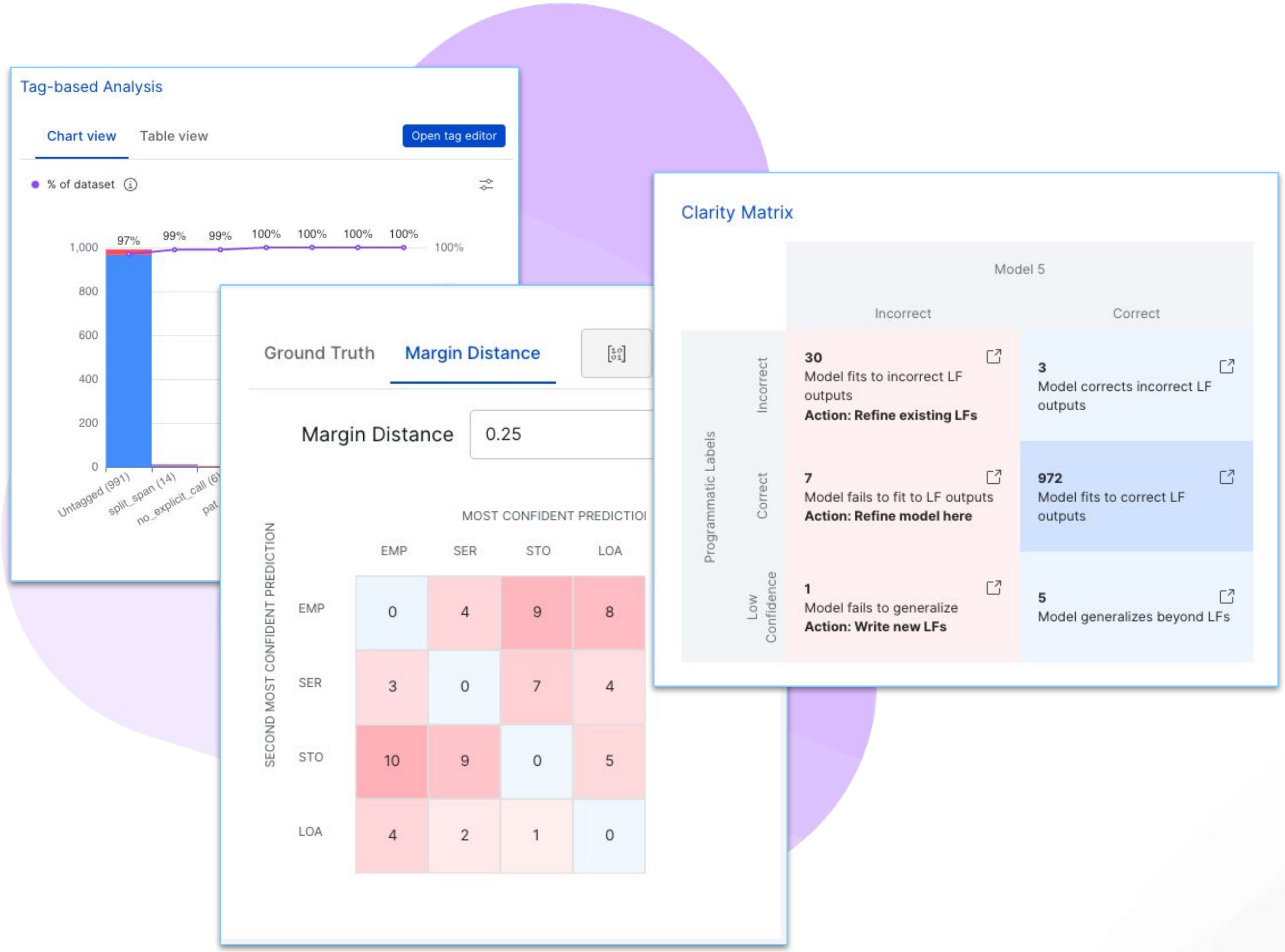
## The data development platform for training & fine-tuning AI



**Discover:** Use model-guided error analysis tools to pinpoint model errors

# Model-guided error analysis

## Quickly identify actionable next steps to improve model performance



Margin Distance 0.25

MOST CONFIDENT PREDICTIO

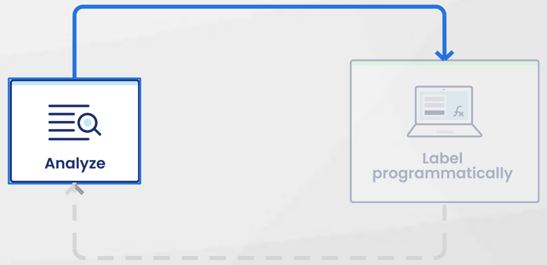
SECOND MOST CONFIDENT PREDICTION

|  | EMP | SER | STO | LOA |
|---|---|---|---|---|
| EMP | 0 | 4 | 9 | 8 |
| SER | 3 | 0 | 7 | 4 |
| STO | 10 | 9 | 0 | 5 |
| LOA | 4 | 2 | 1 | 0 |

Clarity Matrix

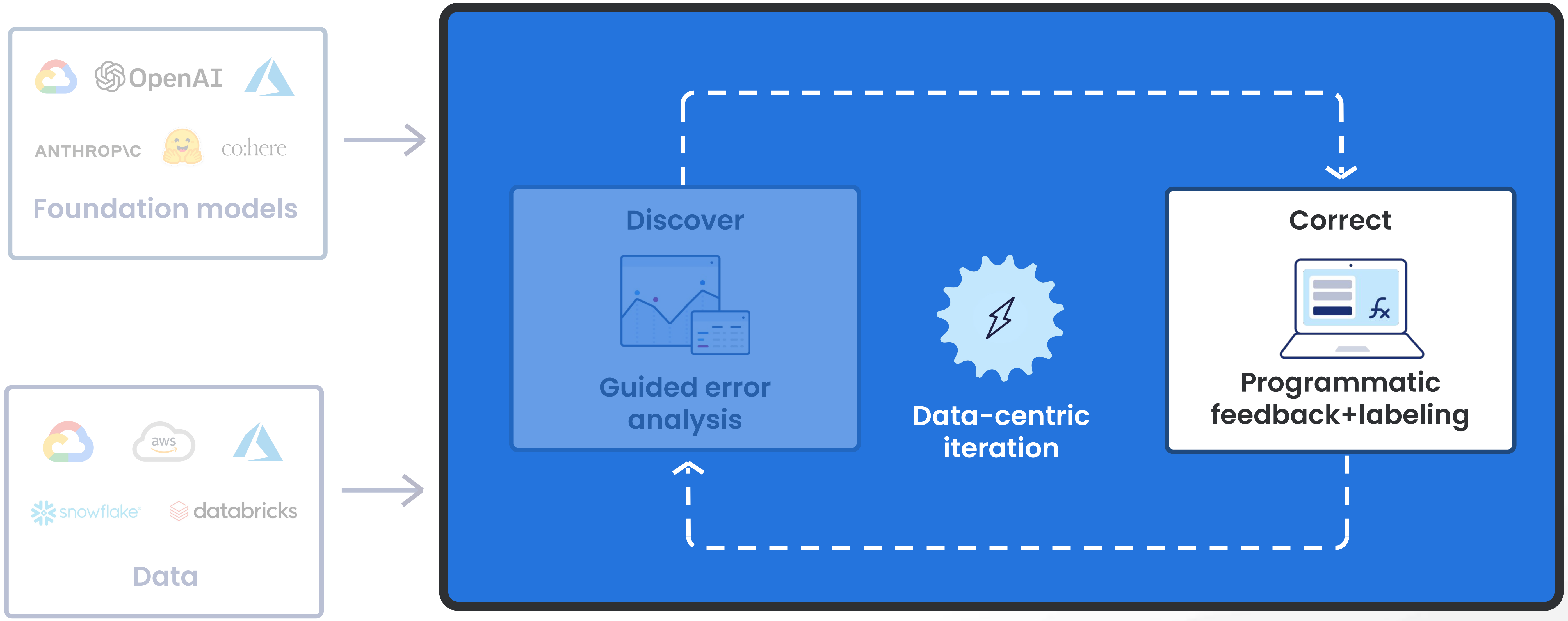
| Programmatic Labels | Model 5: Incorrect | Model 5: Correct |
|---|---|---|
| Incorrect | 30<br>Model fits to incorrect LF outputs<br>**Action: Refine existing LFs** | 3<br>Model corrects incorrect LF outputs |
| Correct | 7<br>Model fails to fit to LF outputs<br>**Action: Refine model here** | 972<br>Model fits to correct LF outputs |
| Low Confidence | 1<br>Model fails to generalize<br>**Action: Write new LFs** | 5<br>Model generalizes beyond LFs |

- **Evaluation of FMs on complex datasets and increasingly open-ended tasks is emerging as one of the most significant challenges**
- **Snorkel Flow's guided error analysis** helps to identify data slices with likely error modes, powered by both model- and SME-driven inputs



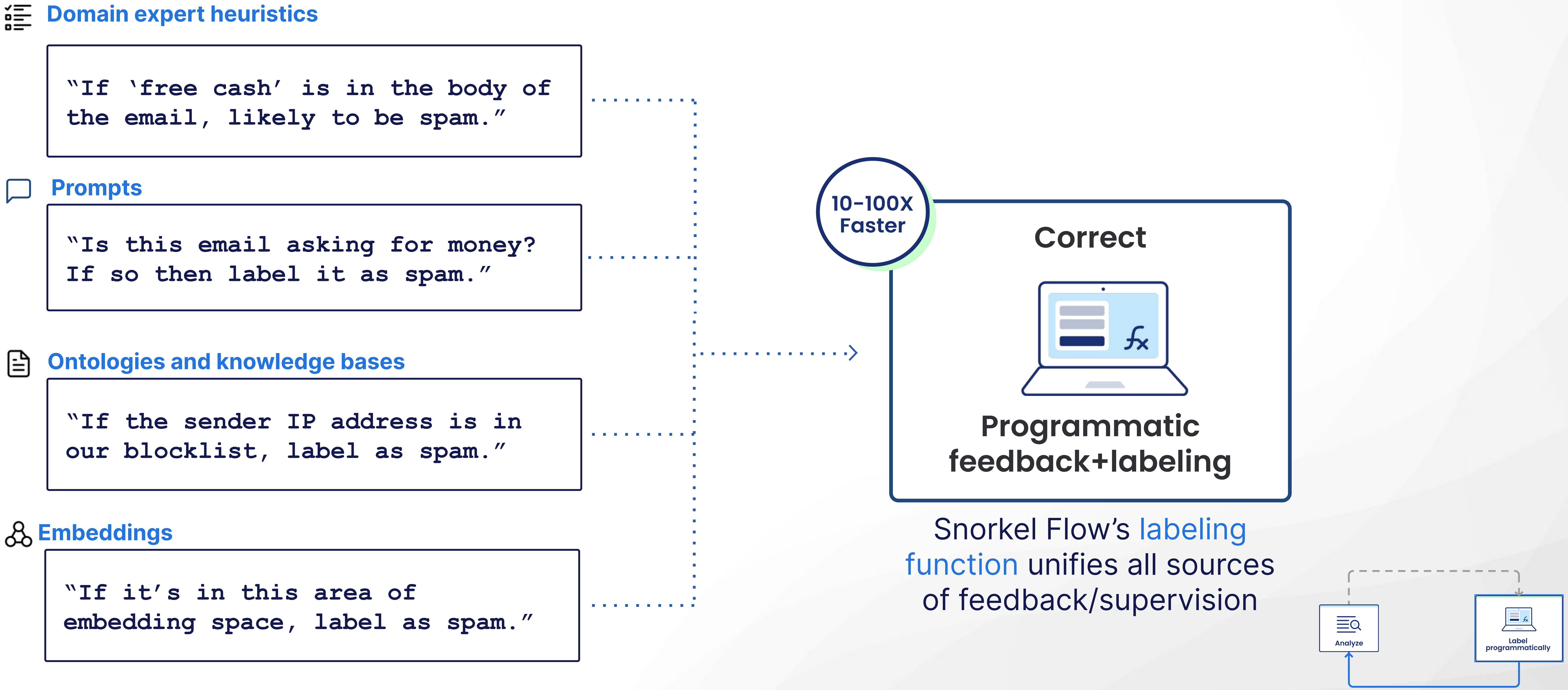
# Snorkel Flow

## The data development platform for training & fine-tuning AI



**Correct:** Use programmatic labeling, prompts and other knowledge resources to create training data and iteratively improve model quality

# Apply all knowledge sources to correct errors



**Domain expert heuristics**

"If 'free cash' is in the body of the email, likely to be spam."

**Prompts**

"Is this email asking for money? If so then label it as spam."

**Ontologies and knowledge bases**

"If the sender IP address is in our blocklist, label as spam."

**Embeddings**

"If it's in this area of embedding space, label as spam."

10-100X Faster

**Correct**

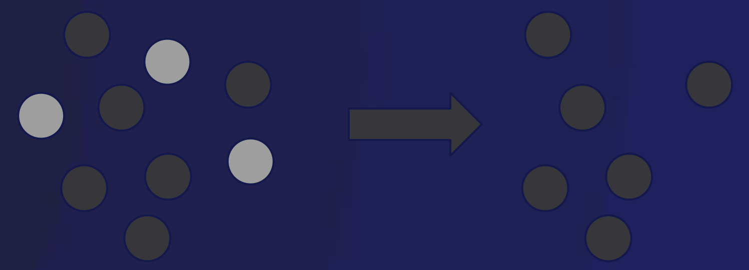
**Programmatic feedback+labeling**

Snorkel Flow's labeling function unifies all sources of feedback/supervision

# Programmatic data development beyond labeling

- ***Sampling***: Choosing the right distribution of prompts & responses

- ***Filtering:*** Filtering for high-quality responses



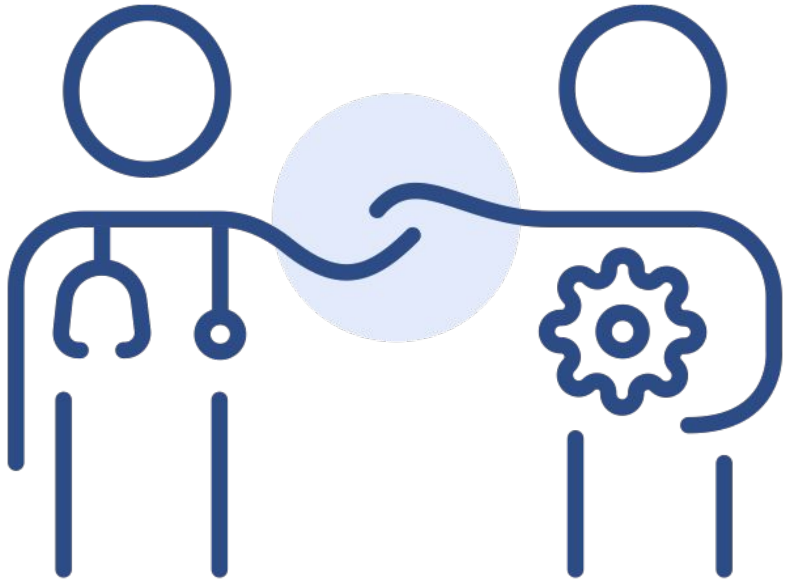
- ***Ranking:*** Providing absolute and relative feedback on prompt/response pairs

(x,y,z) → (y,x,z)

- ***Annotation & routing:*** Estimating response/author quality, routing to correct expertise areas, etc.

**The mix, quality, and annotation of data determines the quality of the statistical model trained on top**

# Data is the common ground for efficient collaboration across the team



**Domain expert**

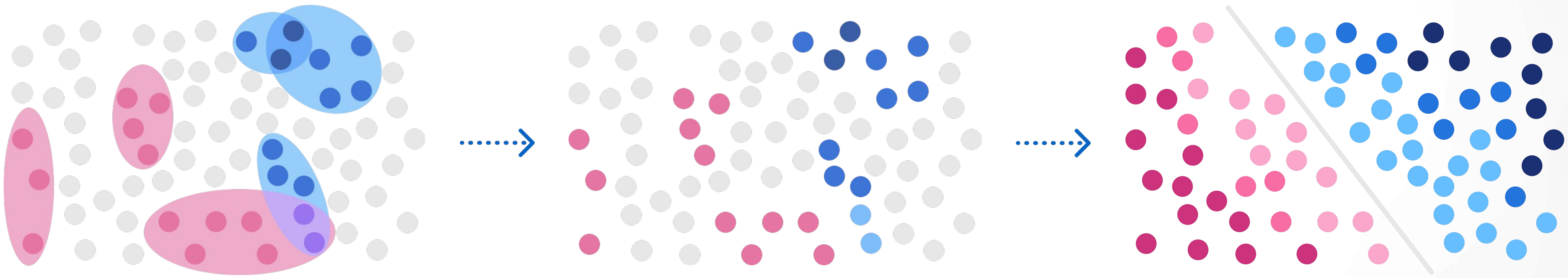
- Tags
- Comments
- Insights

**Data scientist**

- Labeling functions
- Spot checks
- Data slices

DS + SME collaboration is the key– and a data-centric approach facilitates this!

# Snorkel Flow uses theoretically grounded label modeling techniques to denoise and combine



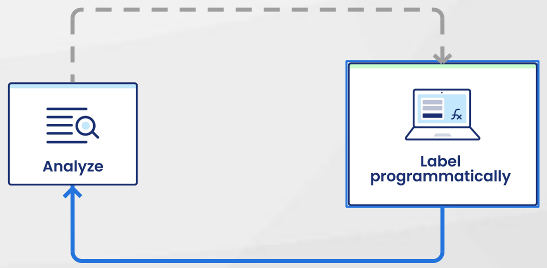
**Labeling functions**
Output: noisy weak labels

**Label model**
Output: denoised training labels

**End ML model**
Output: predictions on all data

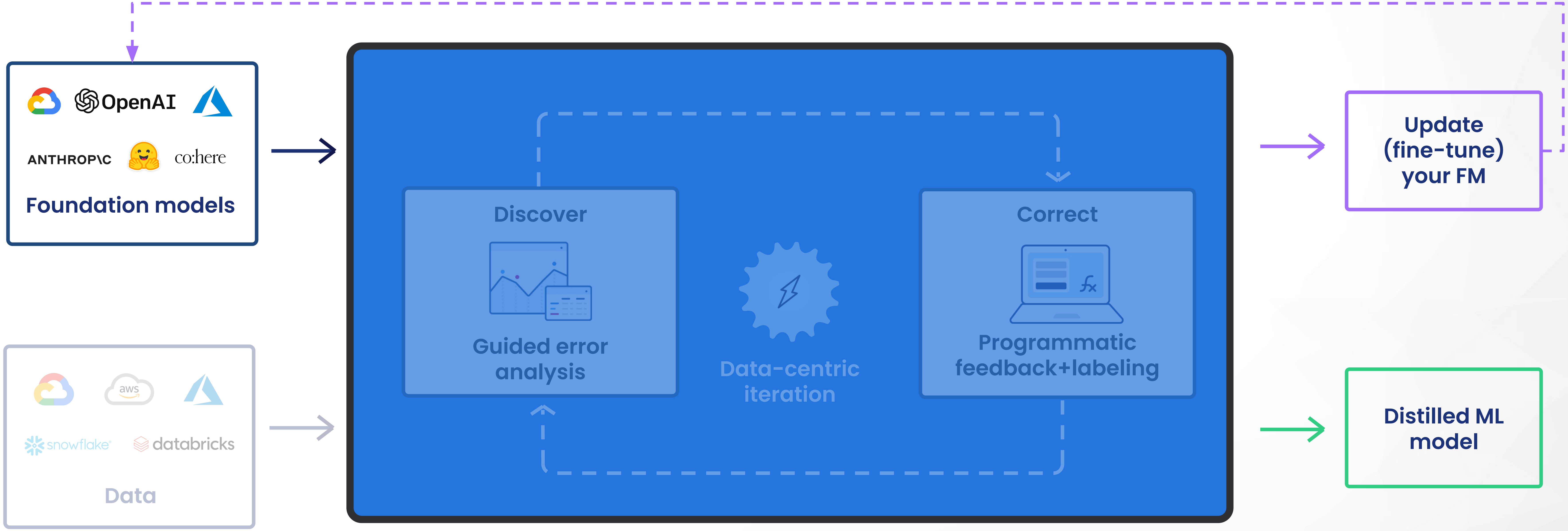
**Based on peer-reviewed research**

- Ratner et. al., NeurIPS’16
- Bach et. al., ICML’17
- Ratner et. al., VLDB’18
- Ratner et. al., AAAI’19
- Varma et. al., ICML’19
- Chen et. al., UAI'22
- and others



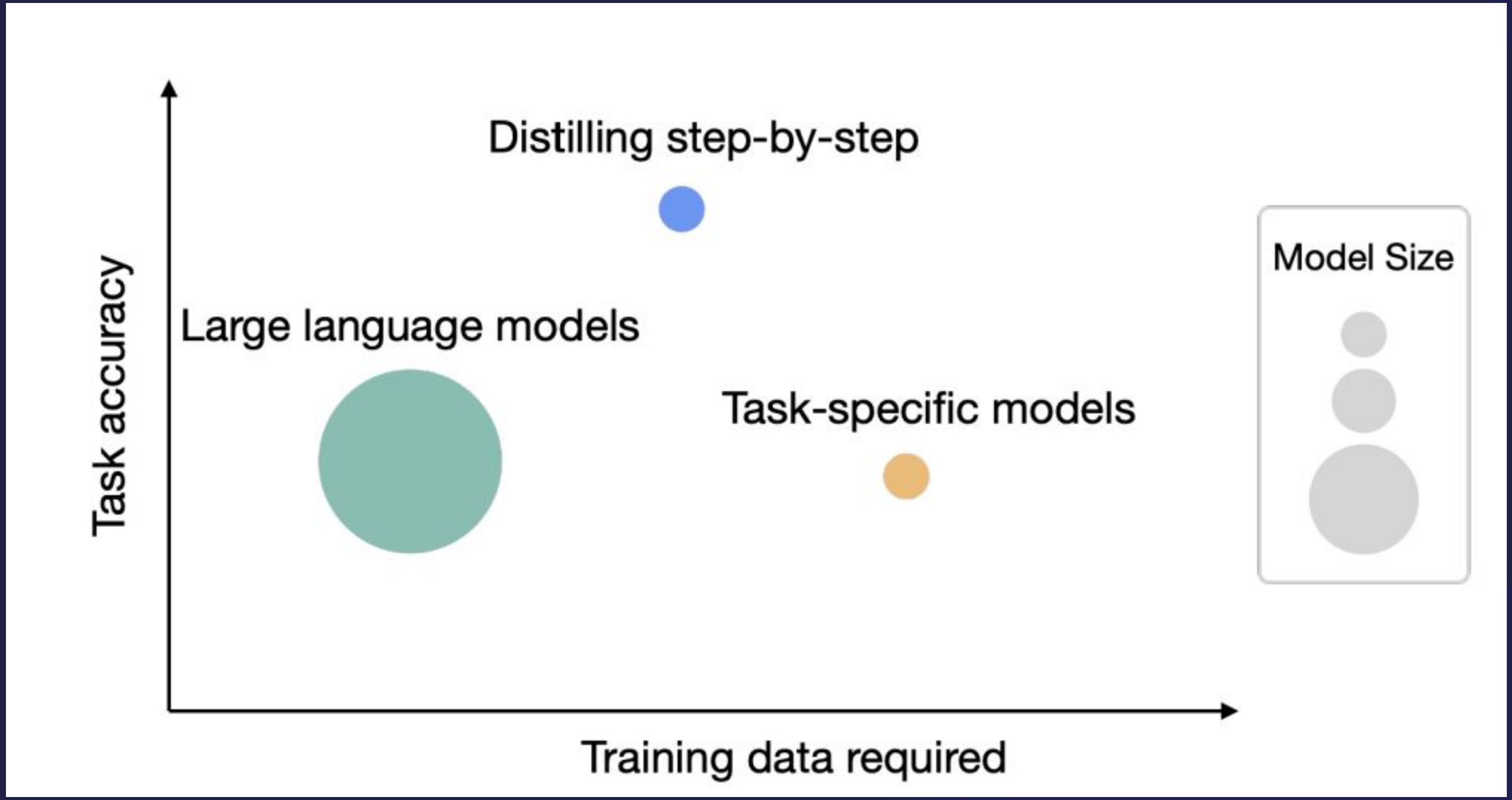
# Snorkel Flow

## The data development platform for training & fine-tuning AI



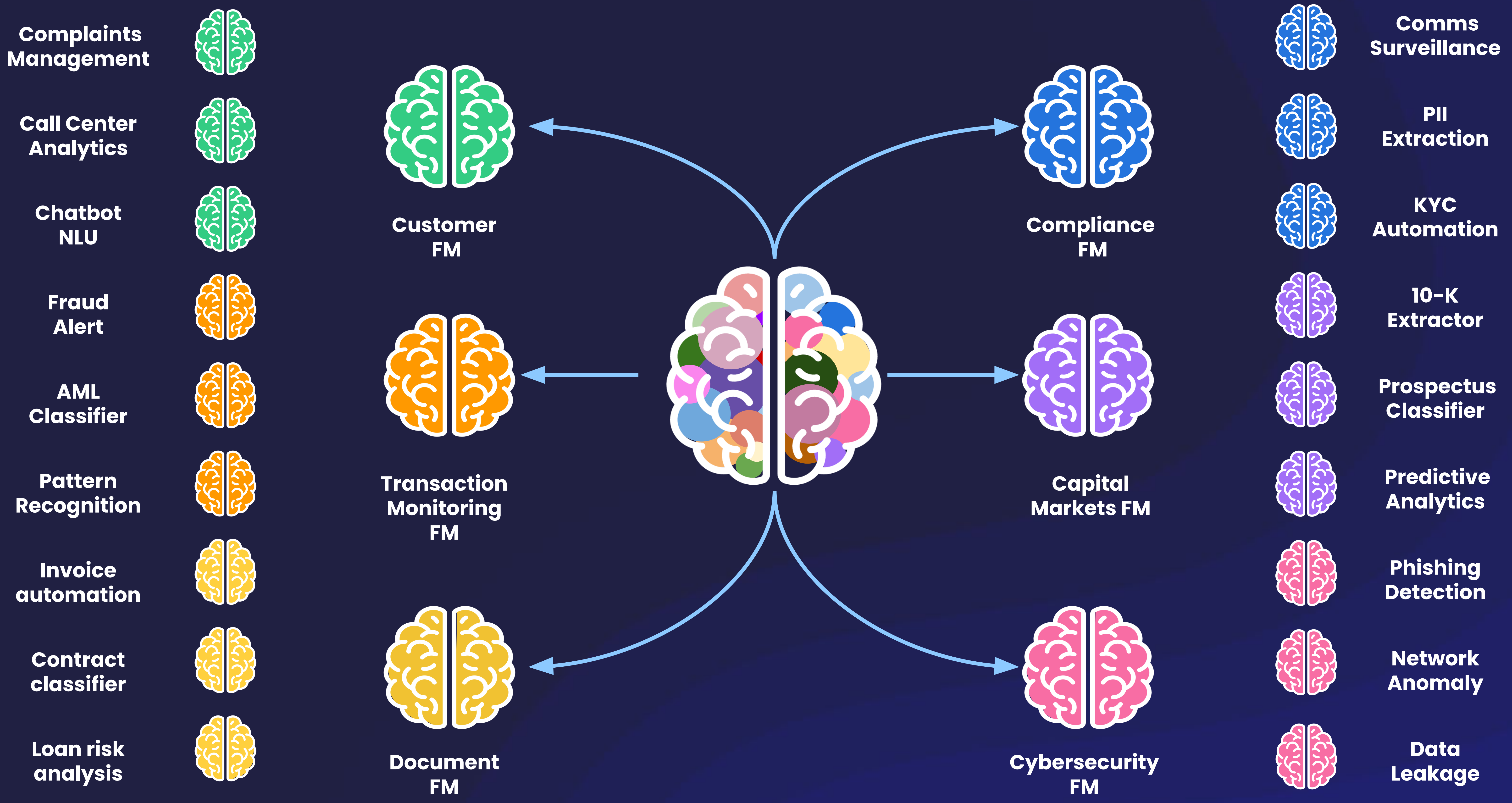
Update or Deploy: Update your FM to build “GPT-You”, or distill into a smaller “specialist” model for deployment

# Example: Distilling Step-By-Step



Key idea: For a specific task, smaller “specialist” models work best and are cheaper too!

# The future of enterprise AI is specialized or hybrid AI



Complaints Management

Call Center Analytics

Chatbot NLU

Fraud Alert

AML Classifier

Pattern Recognition

Invoice automation

Contract classifier

Loan risk analysis

Customer FM

Transaction Monitoring FM

Document FM

Compliance FM

Capital Markets FM

Cybersecurity FM

Comms Surveillance

PII Extraction

KYC Automation

10-K Extractor

Prospectus Classifier

Predictive Analytics

Phishing Detection

Network Anomaly

Data Leakage

# Key thesis: GPT-You, not GPT-X

Foundation models are *all about the data*

Successful enterprises will develop *their own* FMs leveraging their own data and knowledge