# The Data Flywheel for FMware

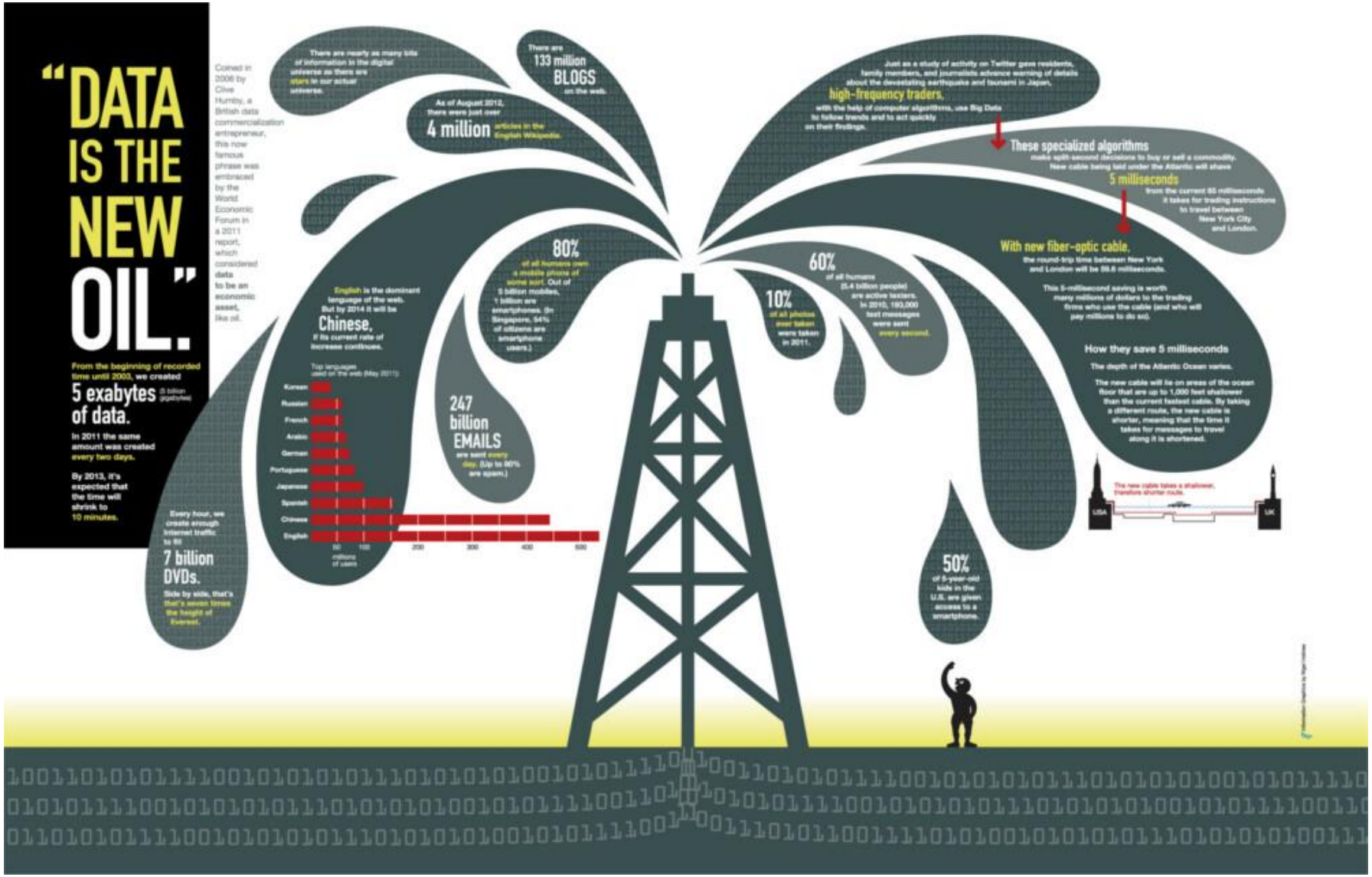
Gopi Krishnan Rajbahadur

# How to cite this session?

```
@misc{Rajbahadur2024dataflywheel,
author = {Gopi Krishnan Rajbahadur},
title = {The Data Flywheel for FMware},
howpublished = {Tutorial presented at the AIware Leadership Bootcamp 2024},
month = {November},
year = {2024},
address = {Toronto, Canada},
note = {Part of the AIware Leadership Bootcamp series.},
url = {https://aiwarebootcamp.io/slides/2024_aiwarebootcamp_rajbahadur_dataflywheelforfmware.pdf }
```

# NETFLIX

# Data is the new oil




https://www.linkedin.com/pulse/data-new-oil-clive-humby-sajithkumar-swaminathan

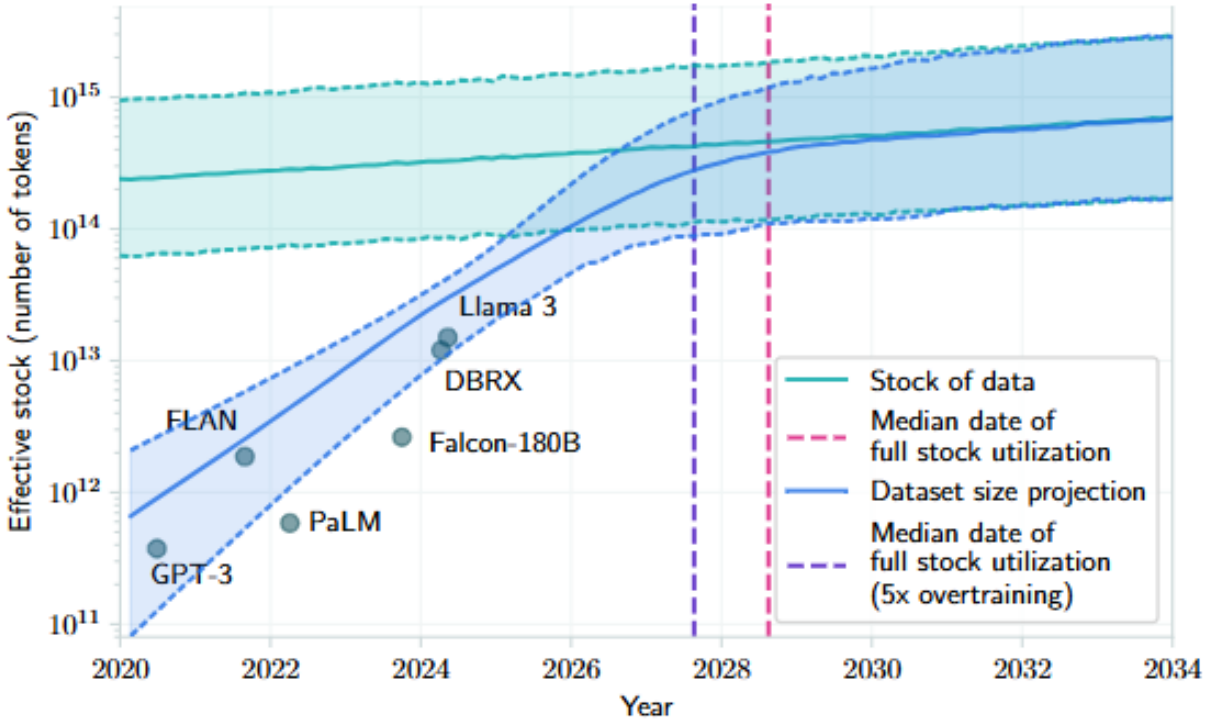
# Data is the new oil


https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

# What happens when we run out of data?

**Will we run out of data? Limits of LLM scaling based on human-generated data**

Pablo Villalobos [1] Anson Ho [1] Jaime Sevilla [1 2] Tamay Besiroglu [1 3] Lennart Heim [1 4] Marius Hobbhahn [1 5]




"Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data **between 2026 and 2032, or slightly earlier if models are overtrained.**"

Vilalobos et al., Will we run out of data? Limits of LLM scaling based on human-generated data

# What happens when we run out of data?

# We turn to renewable resources

But what does it mean in terms of data for FMware?

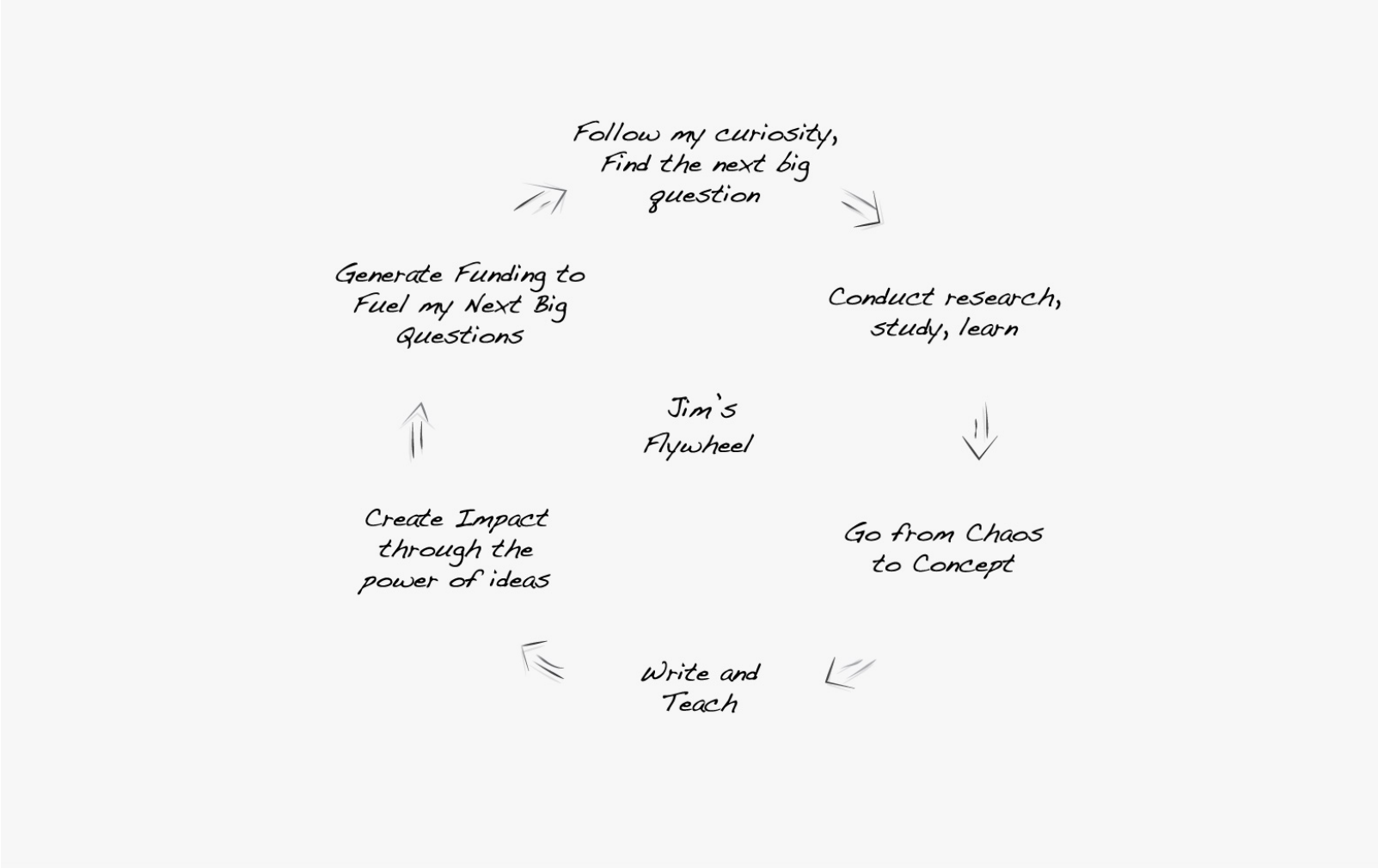
# Say Hello to our savior!

# Can anyone tell me what this is?

# Can anyone tell me what this is?

**Flywheel!**

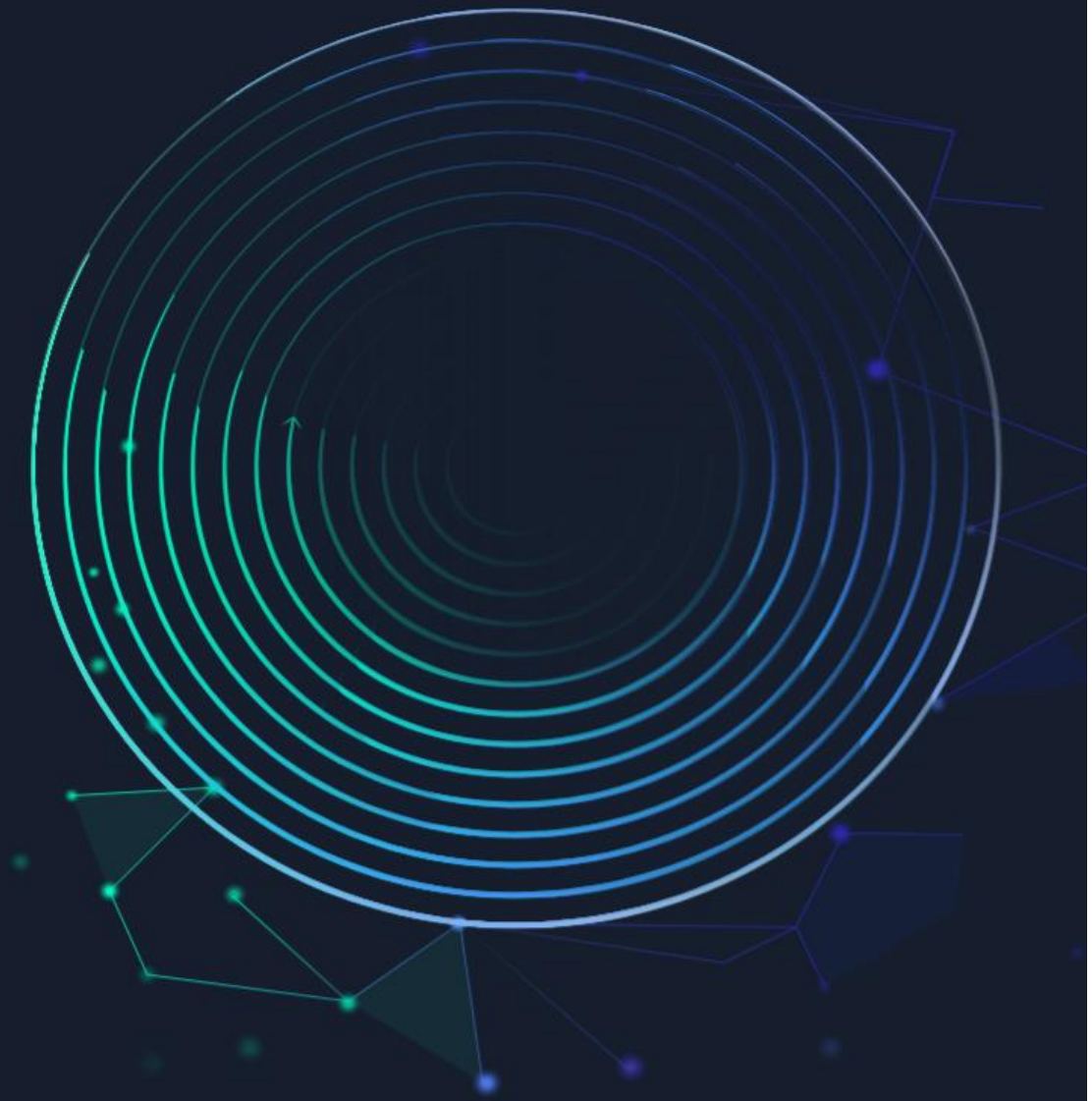
# Jim Collin's Flywheel effect from the book "Good to Great"



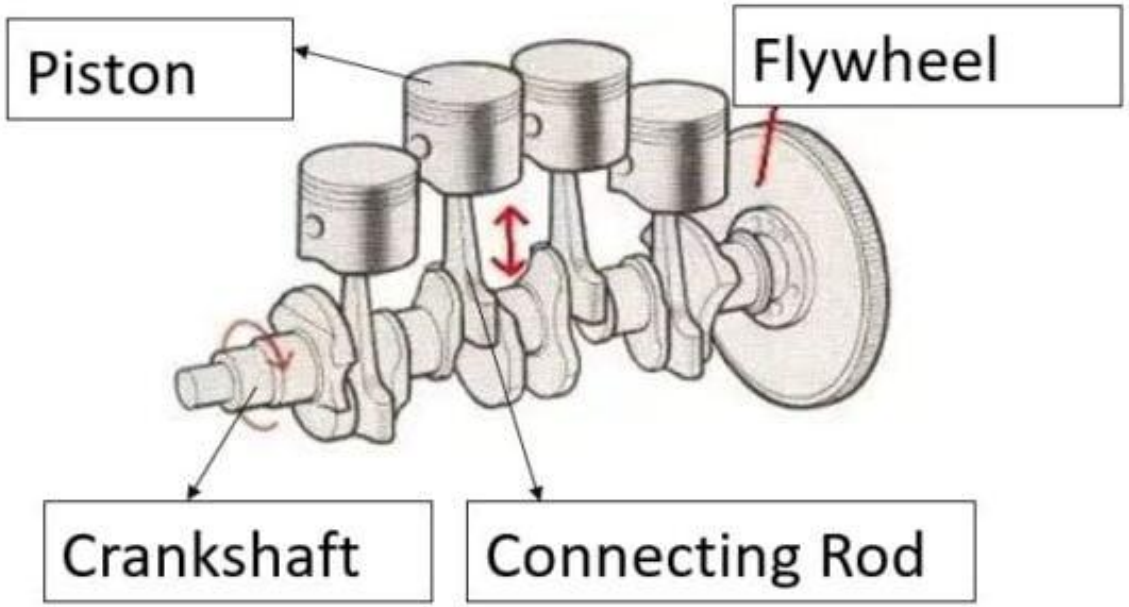
aws databases and analytics

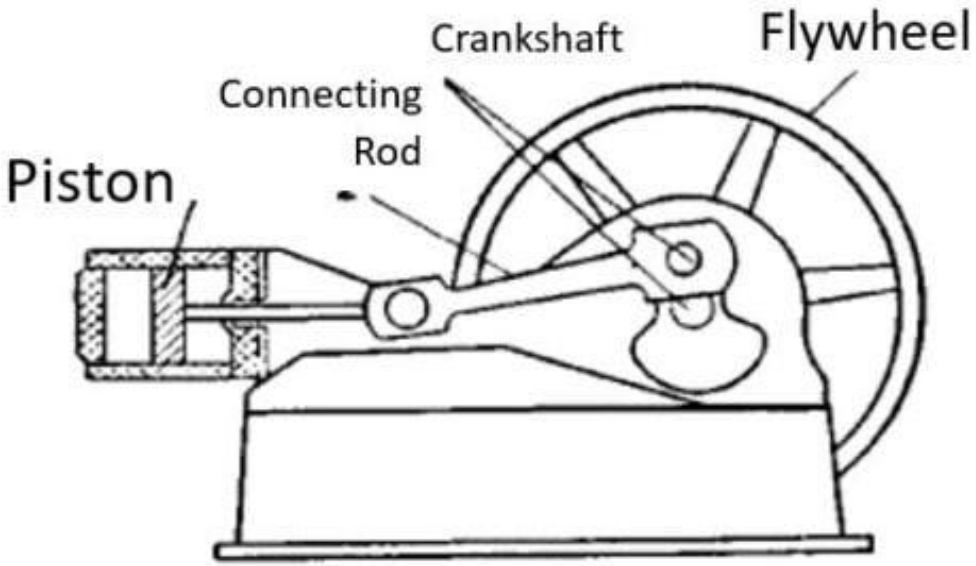
# The Data Flywheel: Building momentum by putting your data to work

Take a comprehensive approach to getting the most value from your data, one project at a time



# How does a Flywheel work?



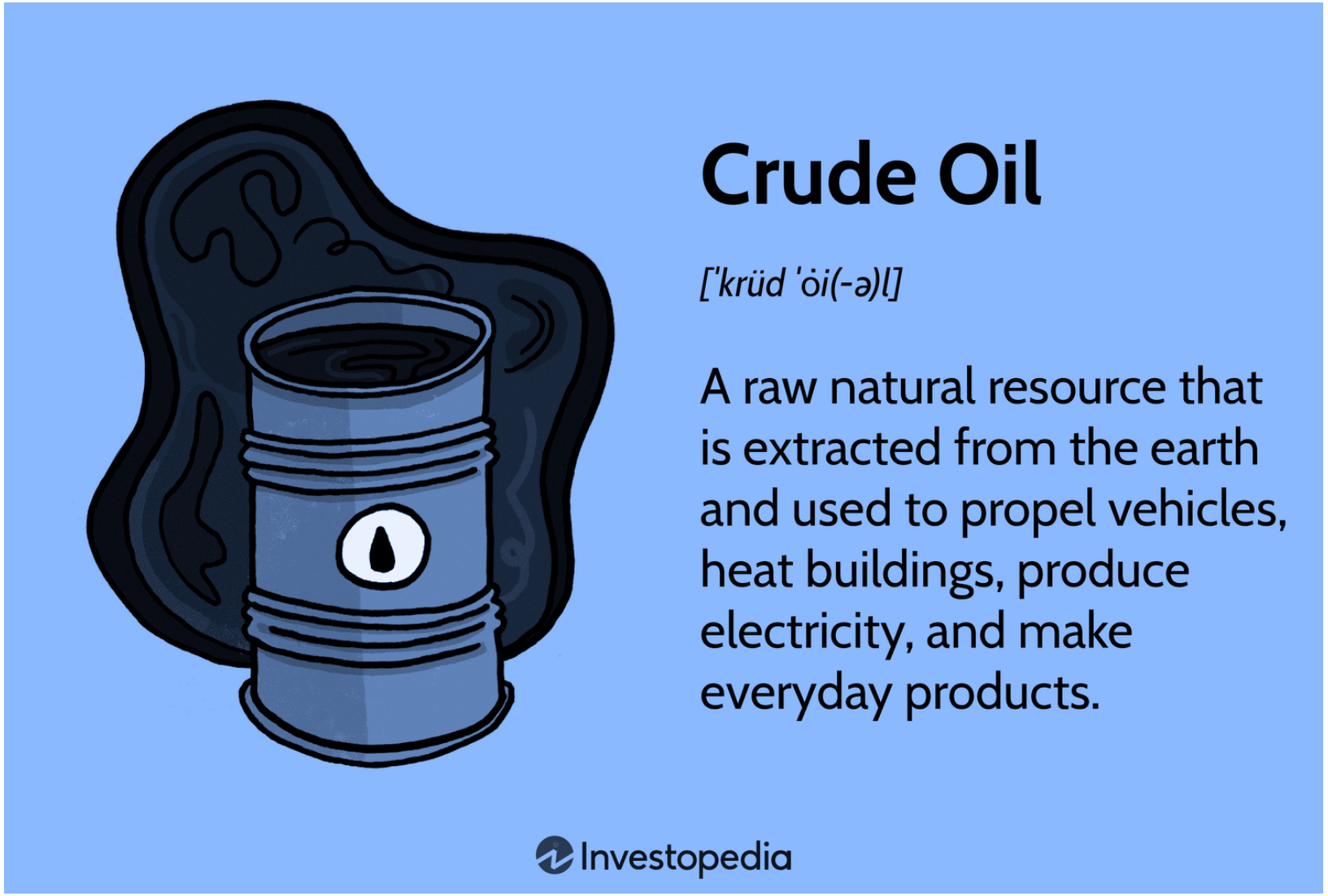




An engine flywheel stores rotational energy, maintaining a steady speed by **smoothing out fluctuations from power strokes**. It conserves momentum, helping the engine run smoothly and **providing energy to keep it turning between cycles**.

# What is the first thing we need to make an engine work?

# What is the first thing we need to make an engine work?

# What is the first thing we need to make an engine work?



# Let’s see if we can identify all the different types of data involved in the FMware Lifecycle

# Let’s see if we can identify all the different types of data involved in the FMware Lifecycle

| | | | |
|---|---|---|---|
| Pretraining data | Instruction data | Finetuning data | Preference data |
| Grounding data | Guarding data | Legal data | Examples data |
| Implicit and Explicit Feedback data | Performance data | Benchmark data | Log data |
| Telemetry data | Output data | | |

# Evolution of synthetic data generation methods

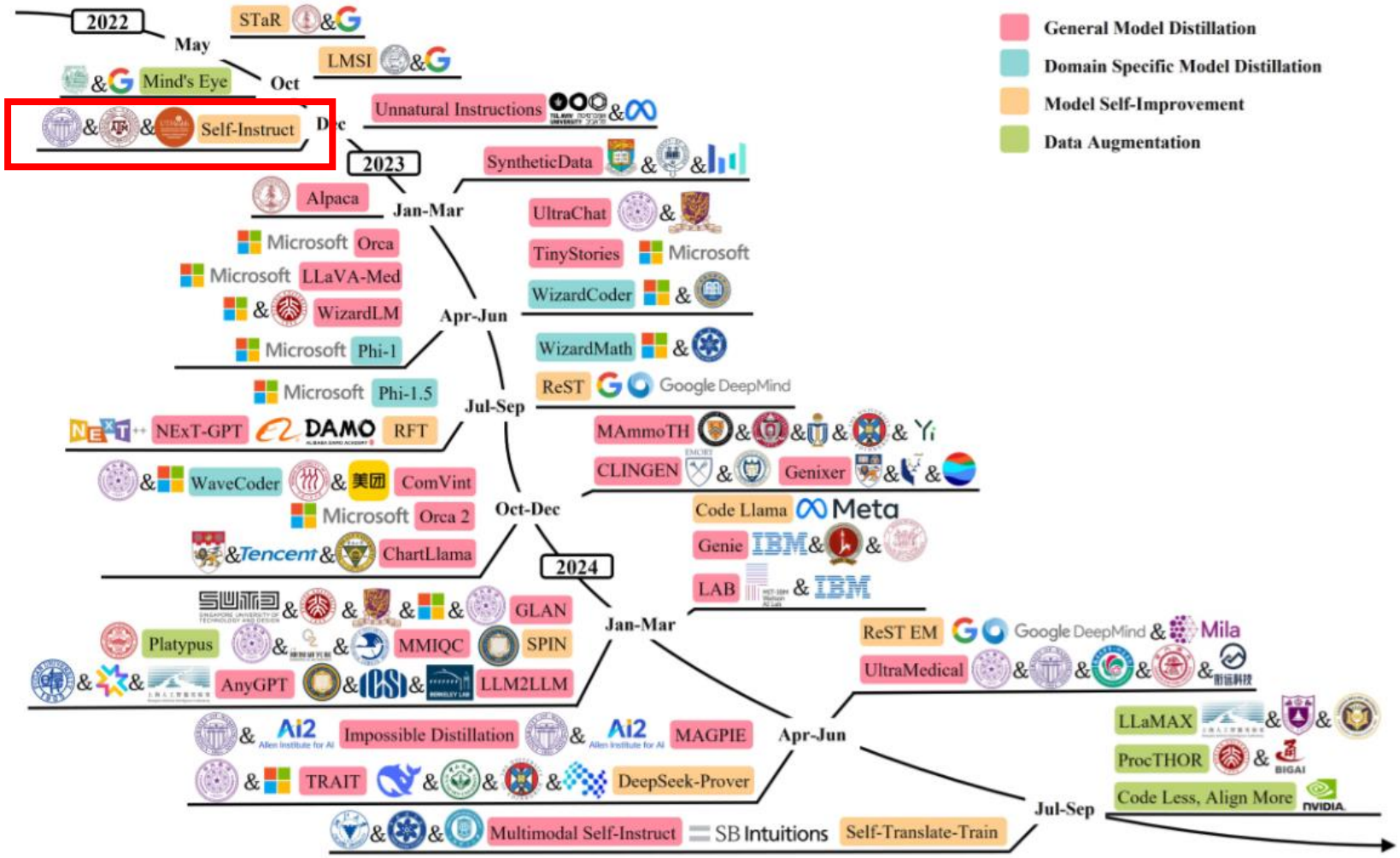


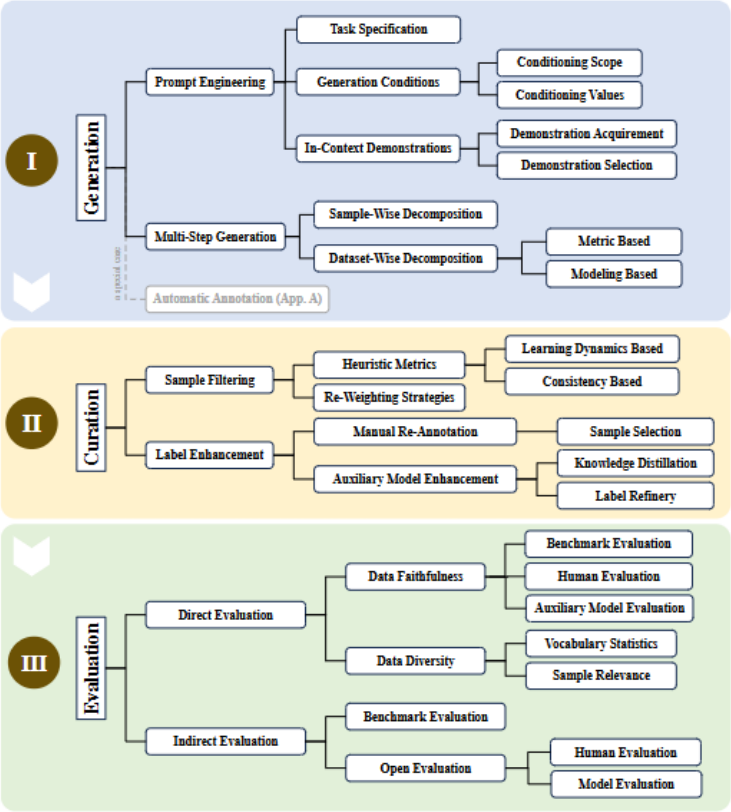

**Figure 4: Illustration of the evolutionary steps in the development of data synthesis and augmentation techniques for large models.**

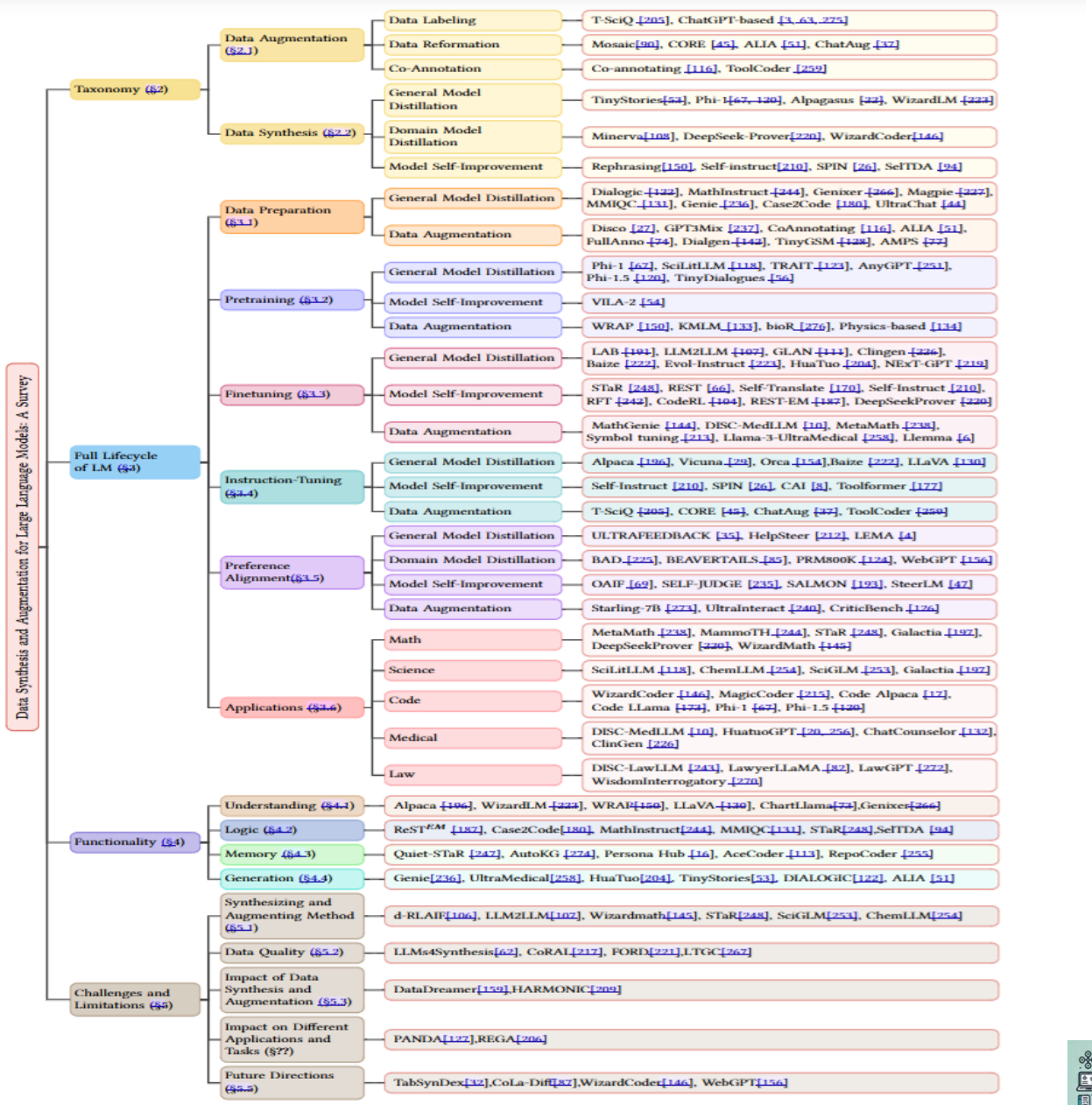
Wang et al., A Survey on Data Synthesis and Augmentation for Large Language Models

Rajbahadur et al., AIware Leadership Bootcamp, Toronto, Canada, 2024

# Much of this data can be generated synthetically

I can’t cover all of these techniques in 35 mins!




Wang et al., A Survey on Data Synthesis and Augmentation for Large Language Models

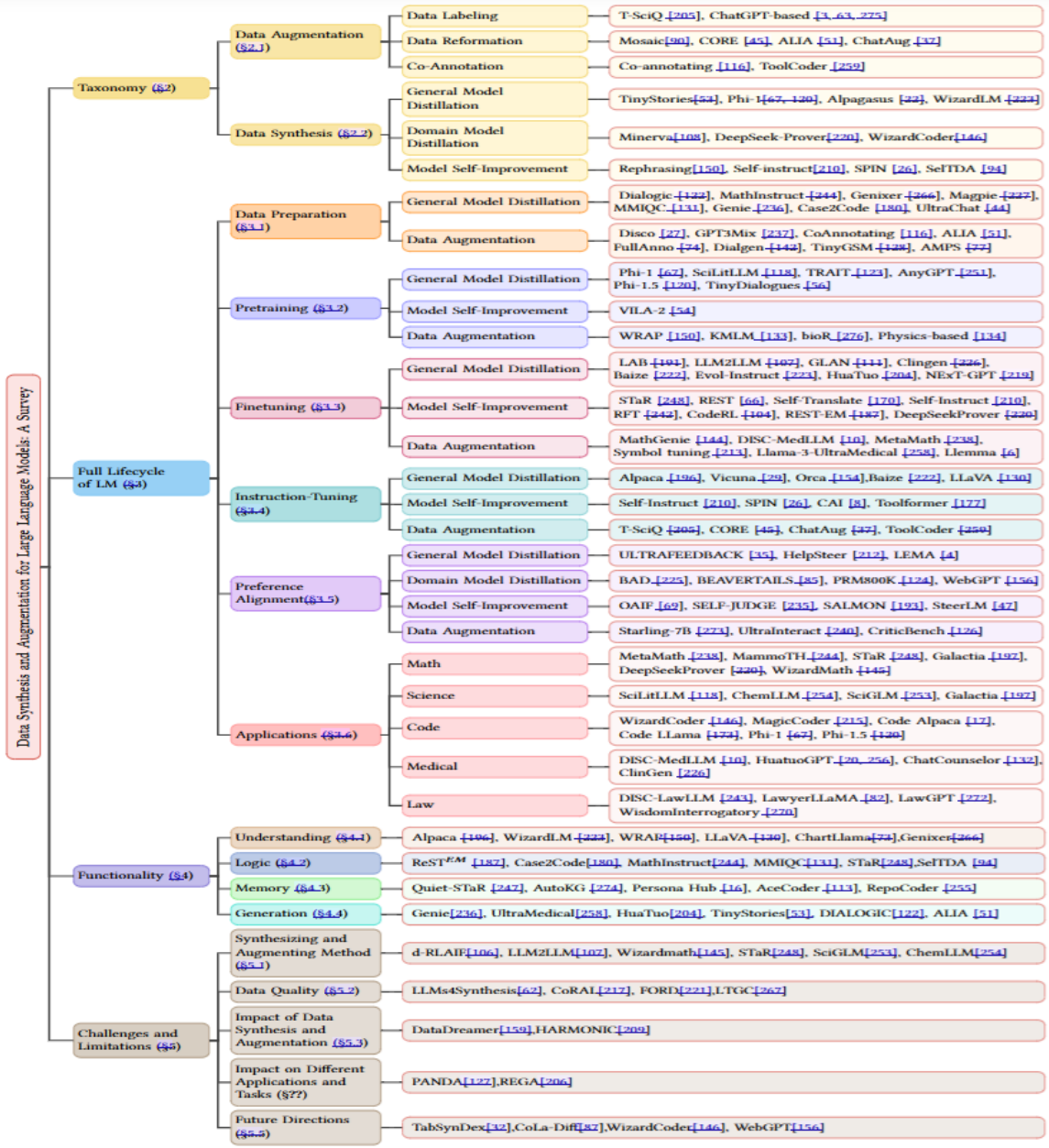
# Much of this data can be generated synthetically

I can't cover all of these techniques in 30 mins! So I am going to give you guys home work!

**Home work**

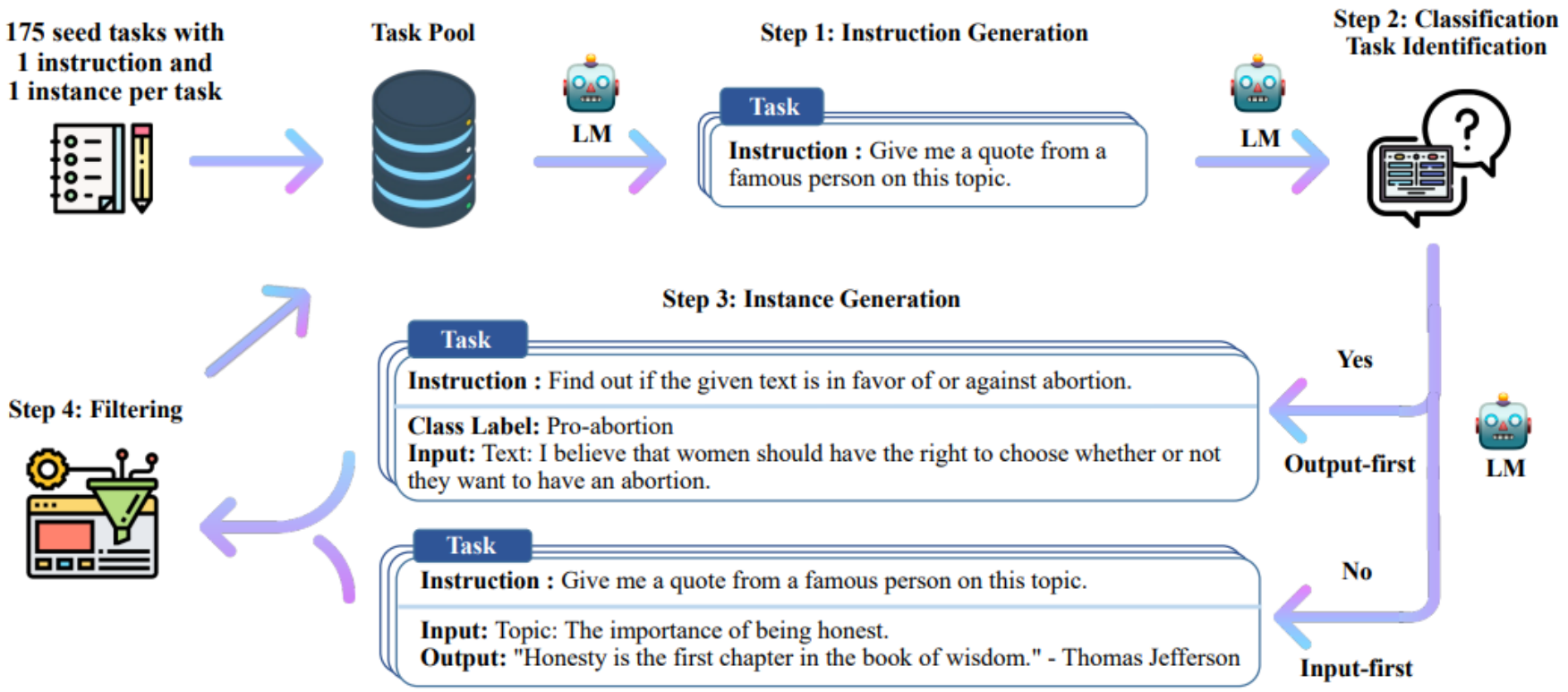
**A Survey on Data Synthesis and Augmentation for Large Language Models**

**On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey**




Wang et al., A Survey on Data Synthesis and Augmentation for Large Language Models

# Self-Instruct – Leveraging FMs to synthetically generate data

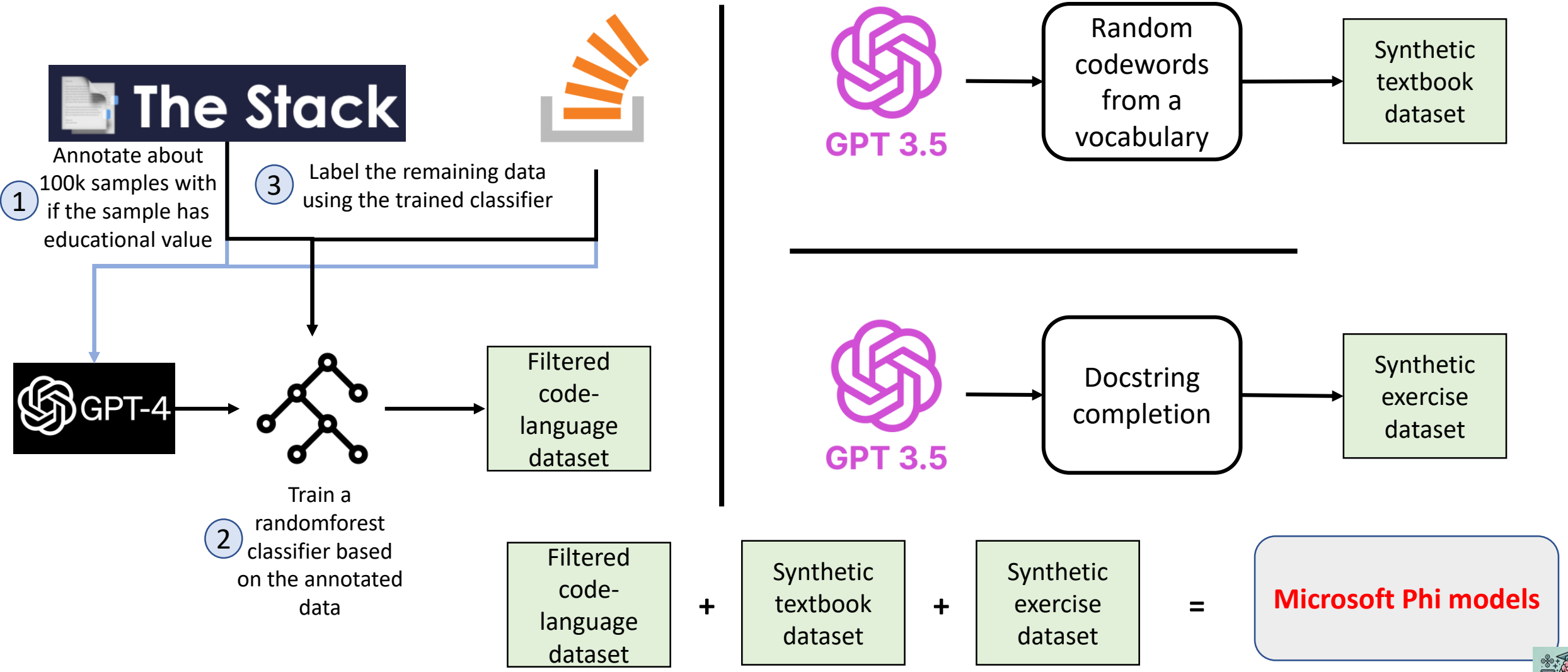



- Use FMs to generate data
- Start with a seed pool of prompts and then iteratively generate and refine task-specific data

Ensuring information diversity and overcoming FM-bias is challenging

Wang et al., Self-Instruct: Aligning Language Models with Self-Generated Instructions

# Microsoft Phi Models

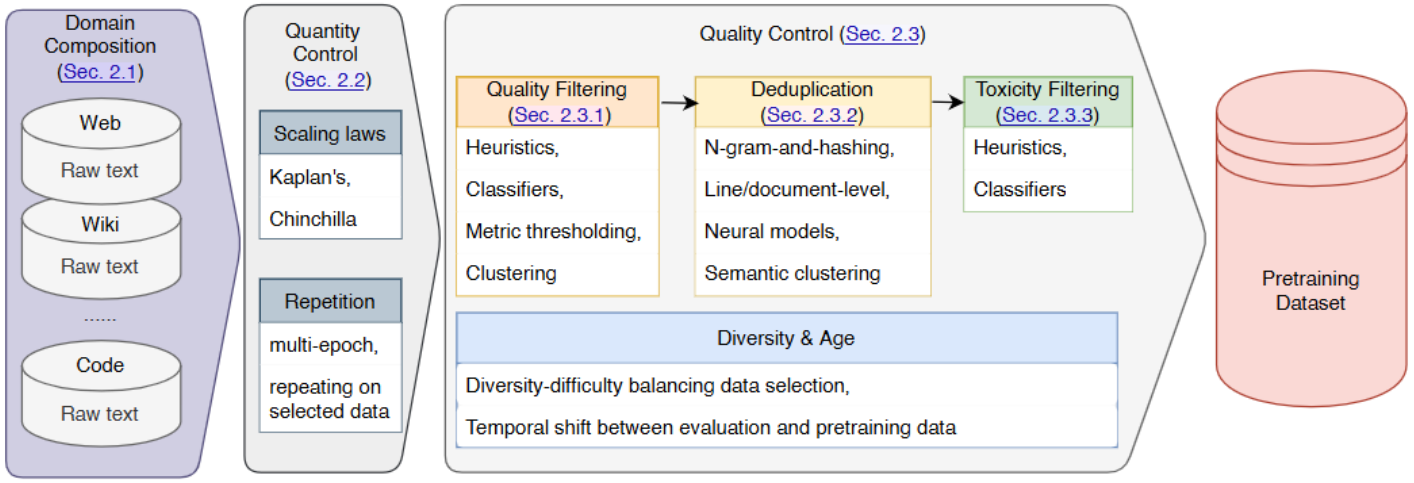



Gunasekar et al., Textbooks Are All You Need

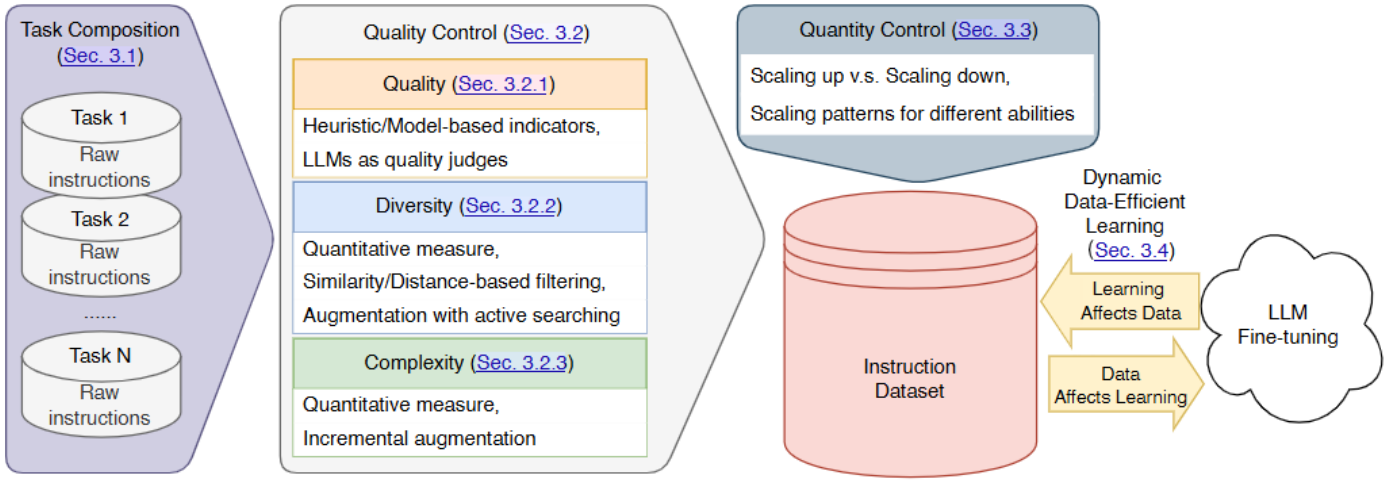
# We now have all the data we need; Crisis averted – can we go home now?

# Not all data are created equal – For the flywheel to work, we need the right kind of data

# Step 1: Ensuring the quality of data crucial for Data Flywheel's success




(a) Data management pipeline in the pretraining stage of LLMs

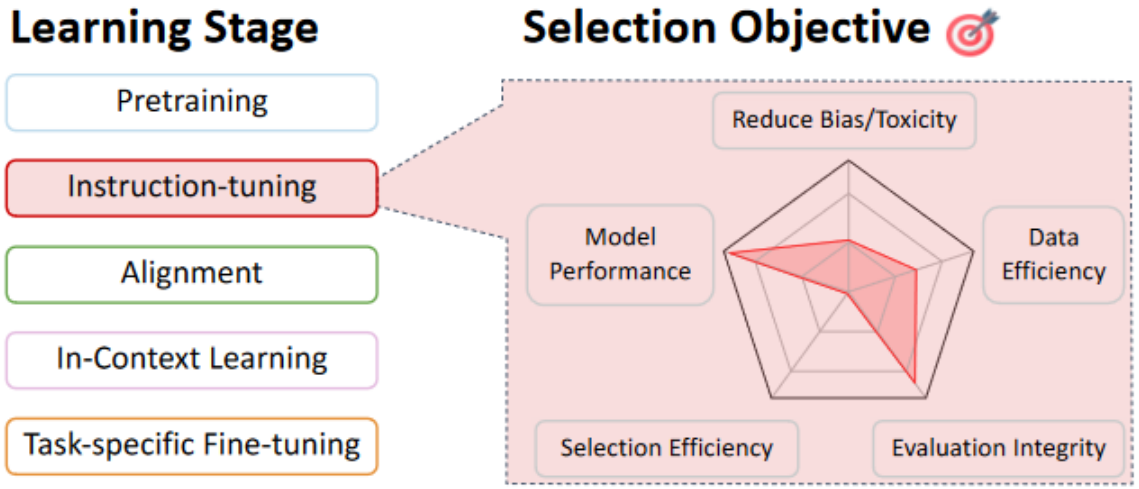



(b) Data management pipeline in the supervised fine-tuning stage of LLMs

Wang et al., Data Management For Training Large Language Models: A Survey

Rajbahadur et al., AIware Leadership Bootcamp, Toronto, Canada, 2024

# Managing the quality, quantity, informativeness and diversity of the data

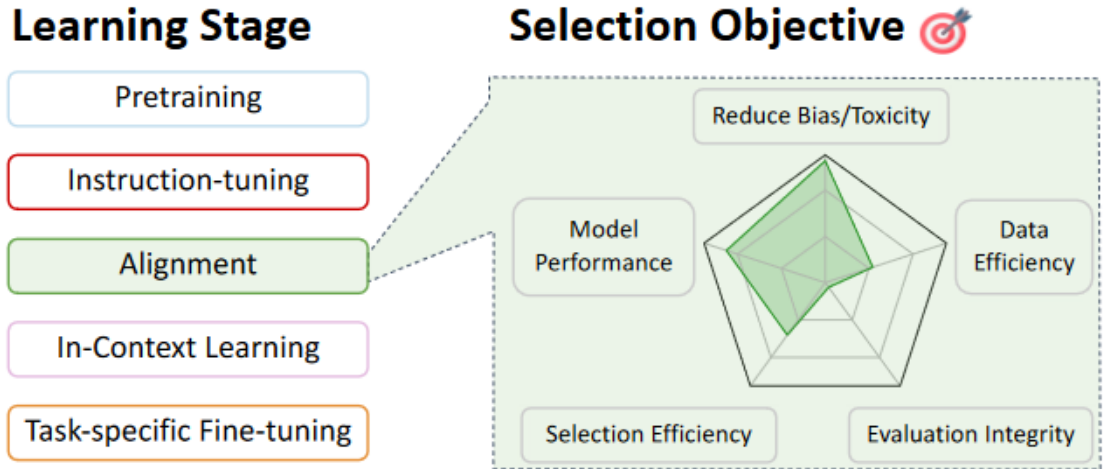


Learning Stage

- Pretraining
- Instruction-tuning
- Alignment
- In-Context Learning
- Task-specific Fine-tuning

Selection Objective

Reduce Bias/Toxicity
Model Performance
Data Efficiency
Selection Efficiency
Evaluation Integrity

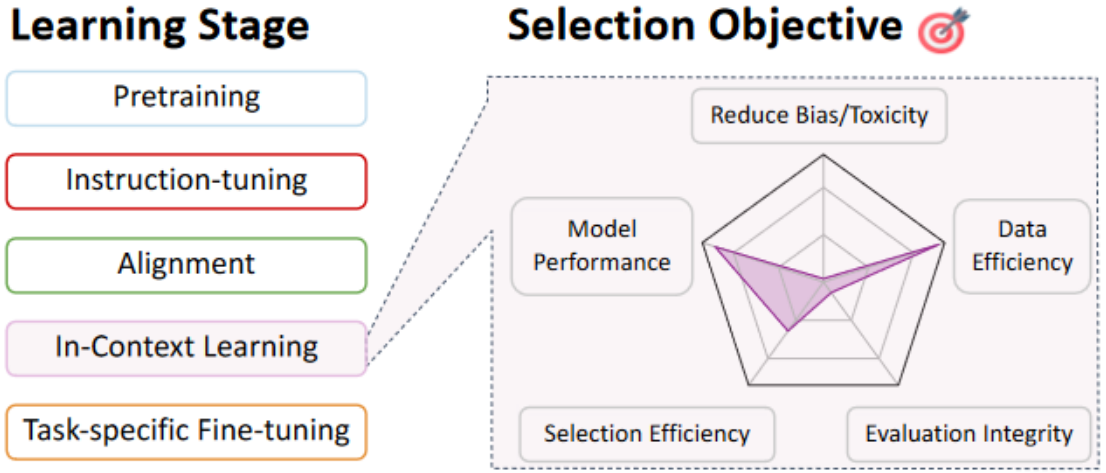


Learning Stage

- Pretraining
- Instruction-tuning
- Alignment
- In-Context Learning
- Task-specific Fine-tuning

Selection Objective

Reduce Bias/Toxicity
Model Performance
Data Efficiency
Selection Efficiency
Evaluation Integrity

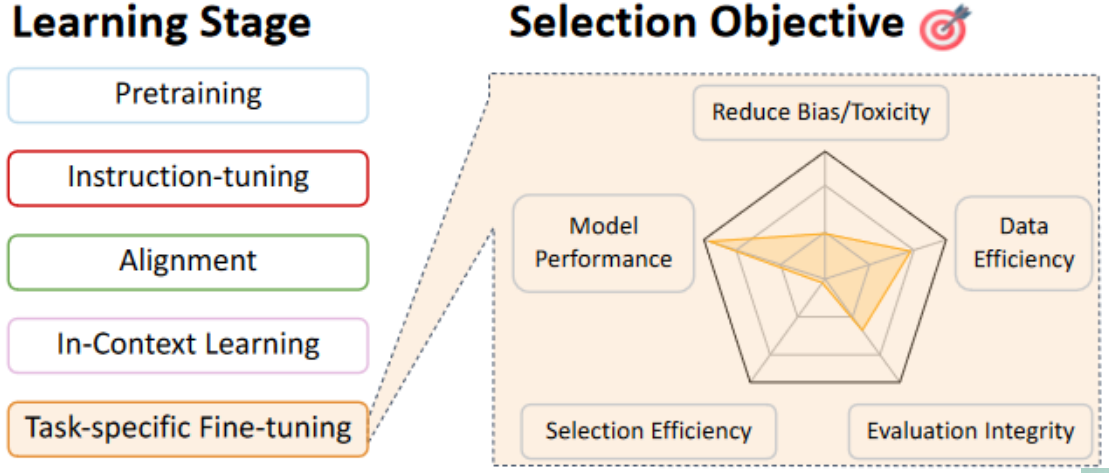


Learning Stage

- Pretraining
- Instruction-tuning
- Alignment
- In-Context Learning
- Task-specific Fine-tuning

Selection Objective

Reduce Bias/Toxicity
Model Performance
Data Efficiency
Selection Efficiency
Evaluation Integrity



Learning Stage

- Pretraining
- Instruction-tuning
- Alignment
- In-Context Learning
- Task-specific Fine-tuning

Selection Objective

Reduce Bias/Toxicity
Model Performance
Data Efficiency
Selection Efficiency
Evaluation Integrity

Albalak et al., A Survey on Data Selection for Language Models

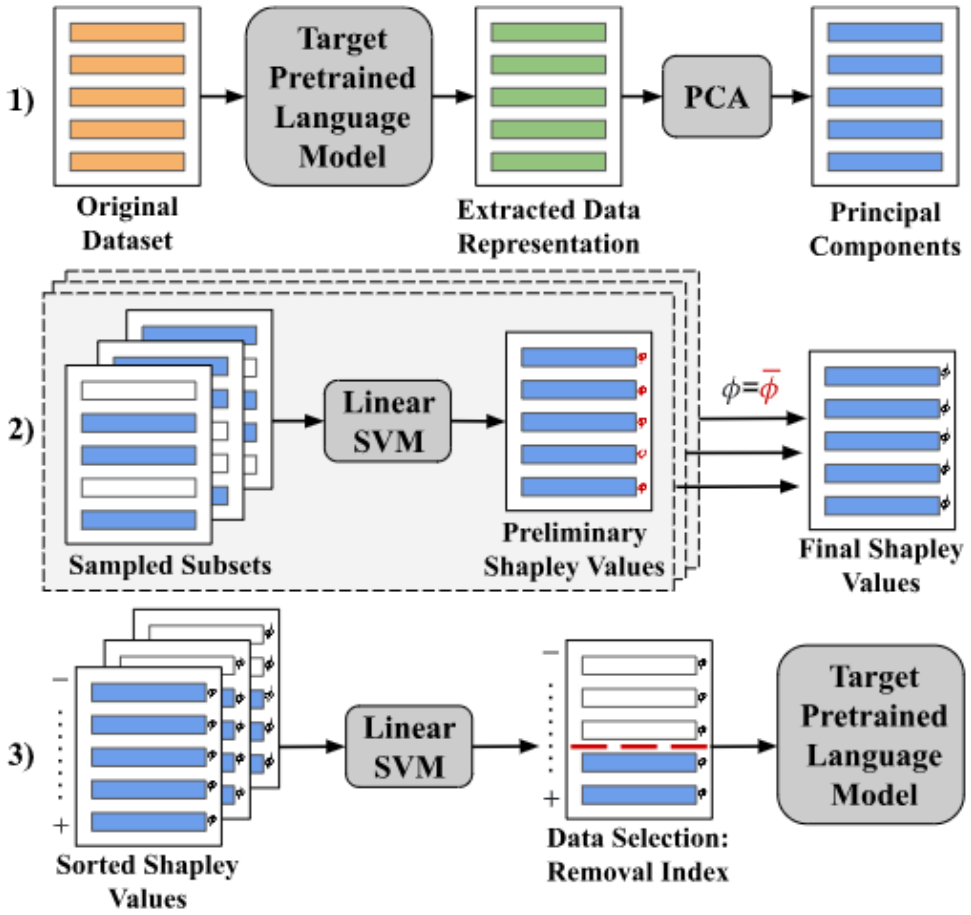
Rajbahadur et al., AIware Leadership Bootcamp, Toronto, Canada, 2024

# Step 2: Select the right type and quantity of data


**LIMA: Less Is More for Alignment**

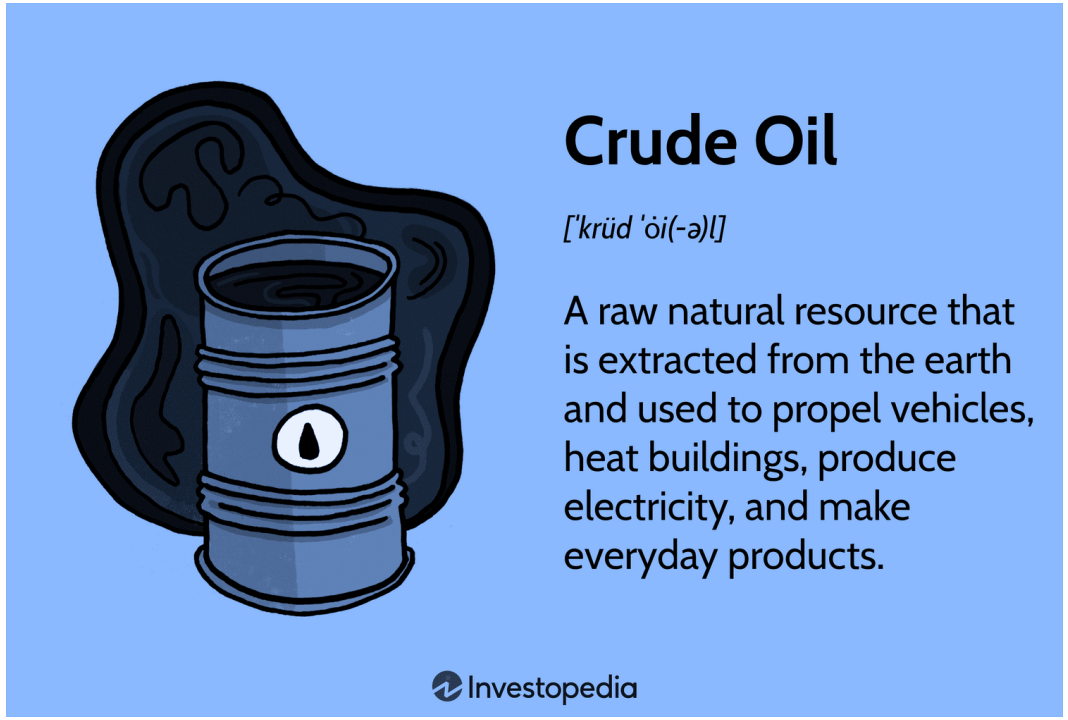
# Step 3: Ensure that the data that we use to build the FMware is useful

Data valuation methods estimate the contributive value of each data point so that the data we include is useful




Schoch et al., CS-Shapley: Class-wise Shapley Values for Data Valuation in Classification

# Do we now have all the ingredients?





**Not Yet!**

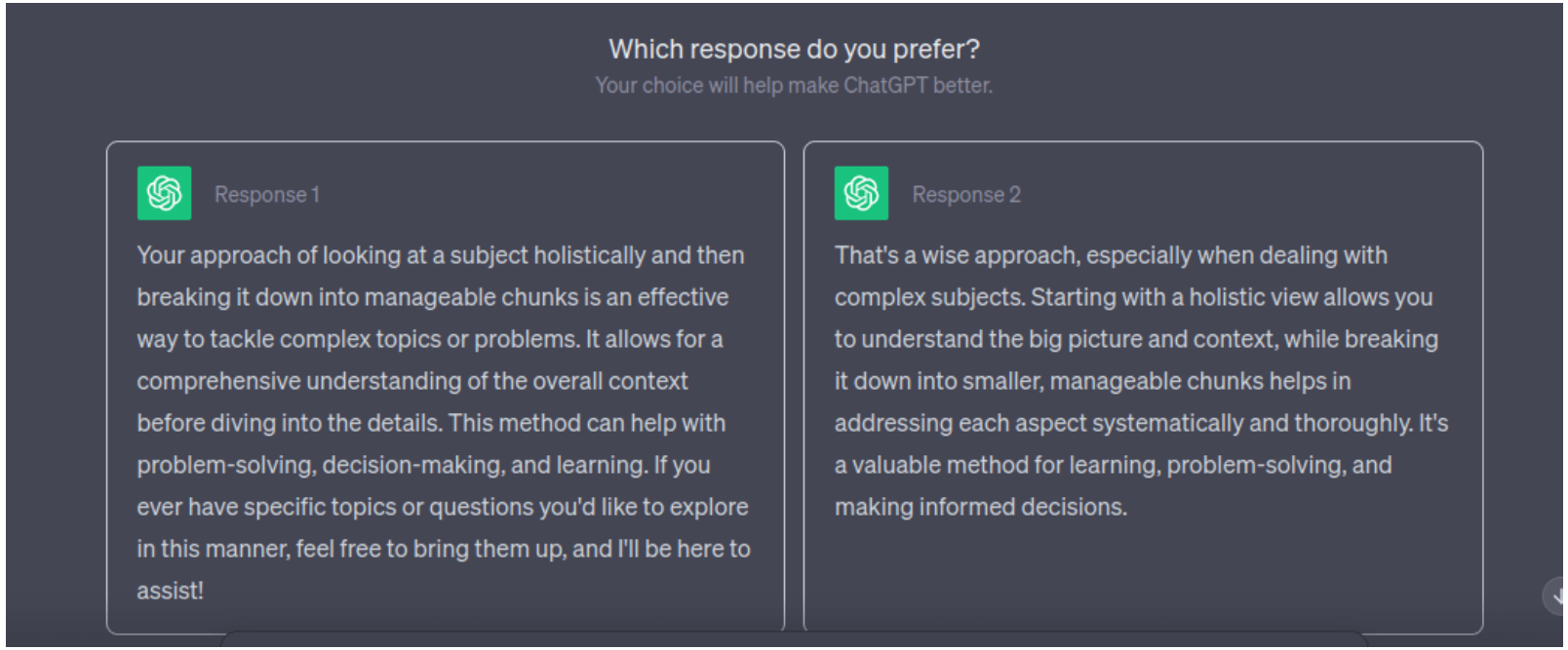
# We can’t always rely on just Synthetic data

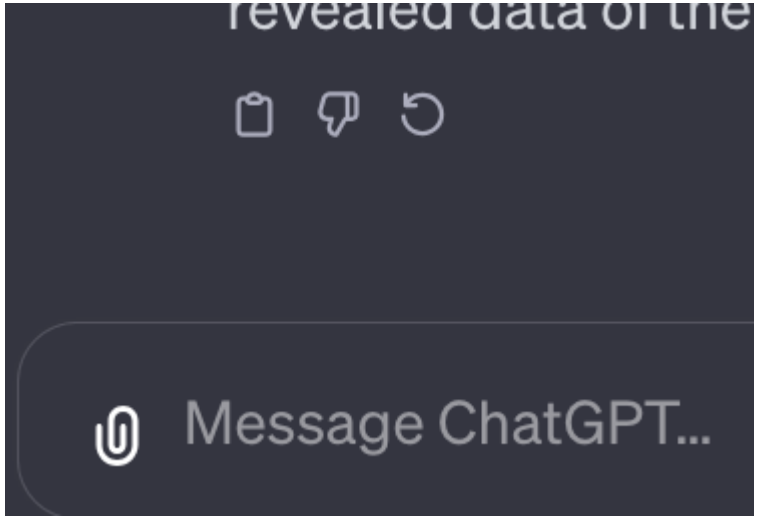
FMs are limited by the data that it was trained on … If we run out of data, all the variety of data that we can synthetically generate **might not have enough information diversity!**

**We need to get the human generated data!**

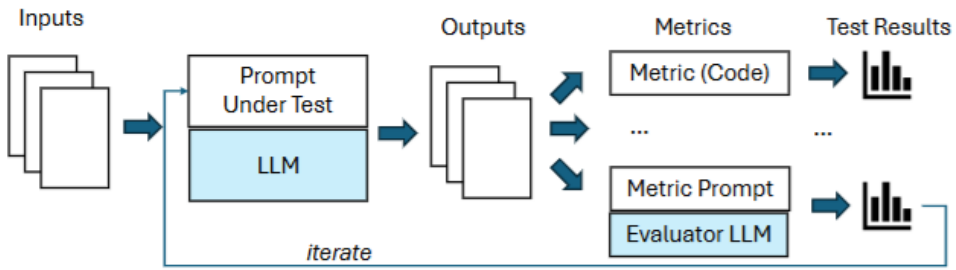
# But can you trust data from anyone!


# Lots of feedback data is available

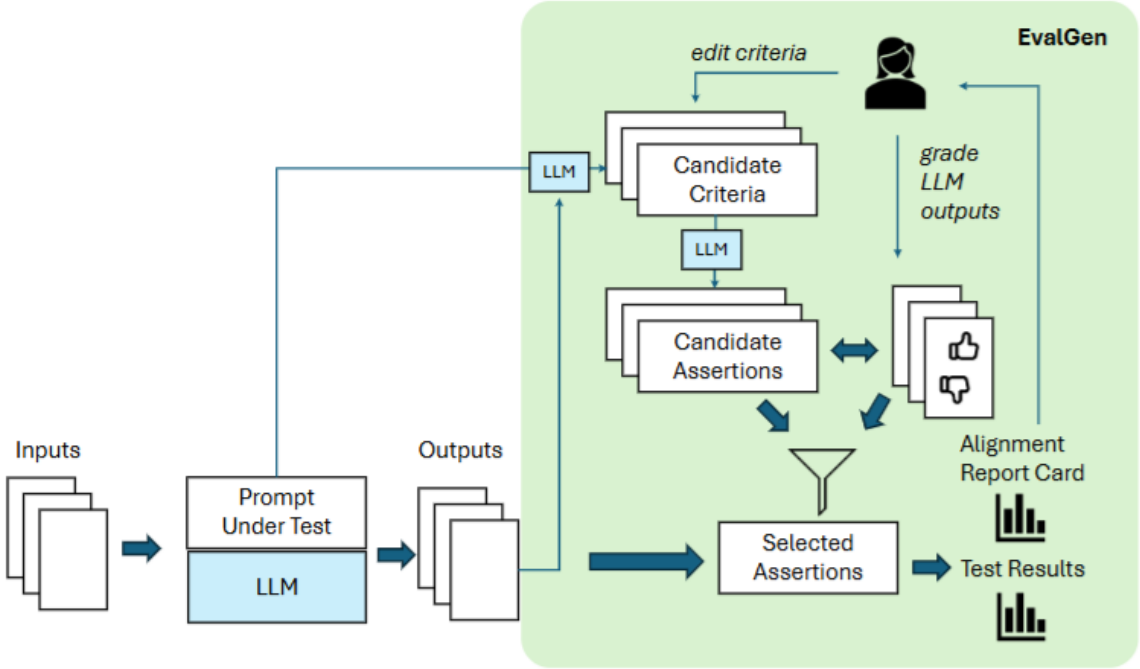


# Flywheel requires robust ways of integrating user generated data




(a) Typical Evaluation Pipeline



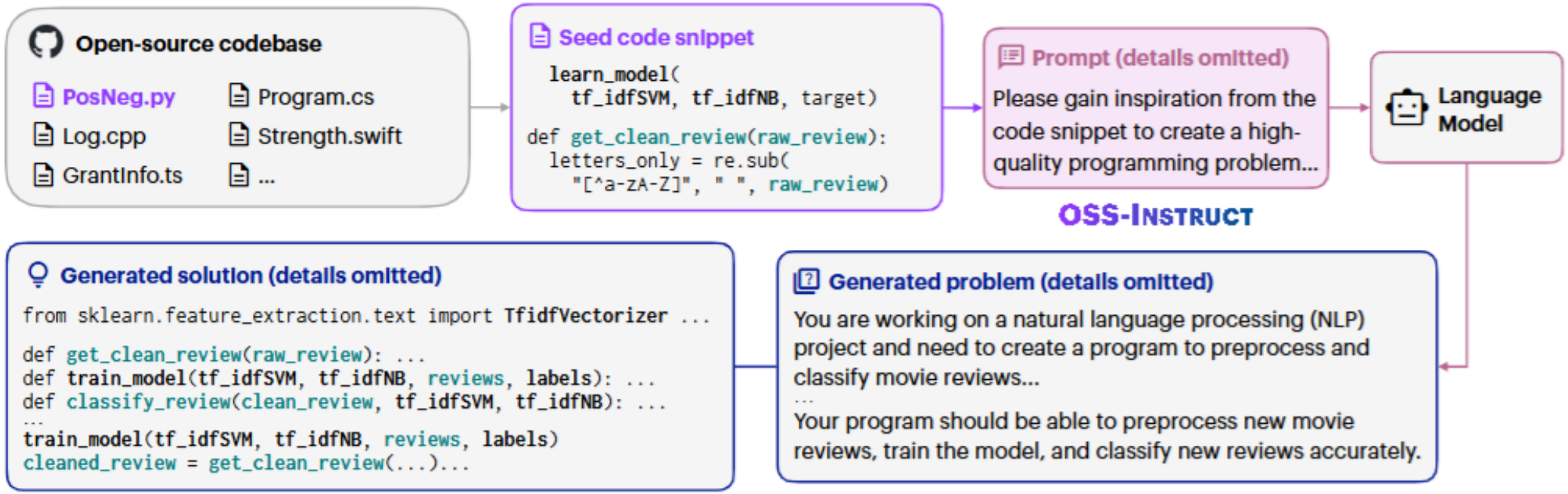

Shankar et al., Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences

# Flywheel requires robust ways of integrating user generated data


Shankar et al., Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences

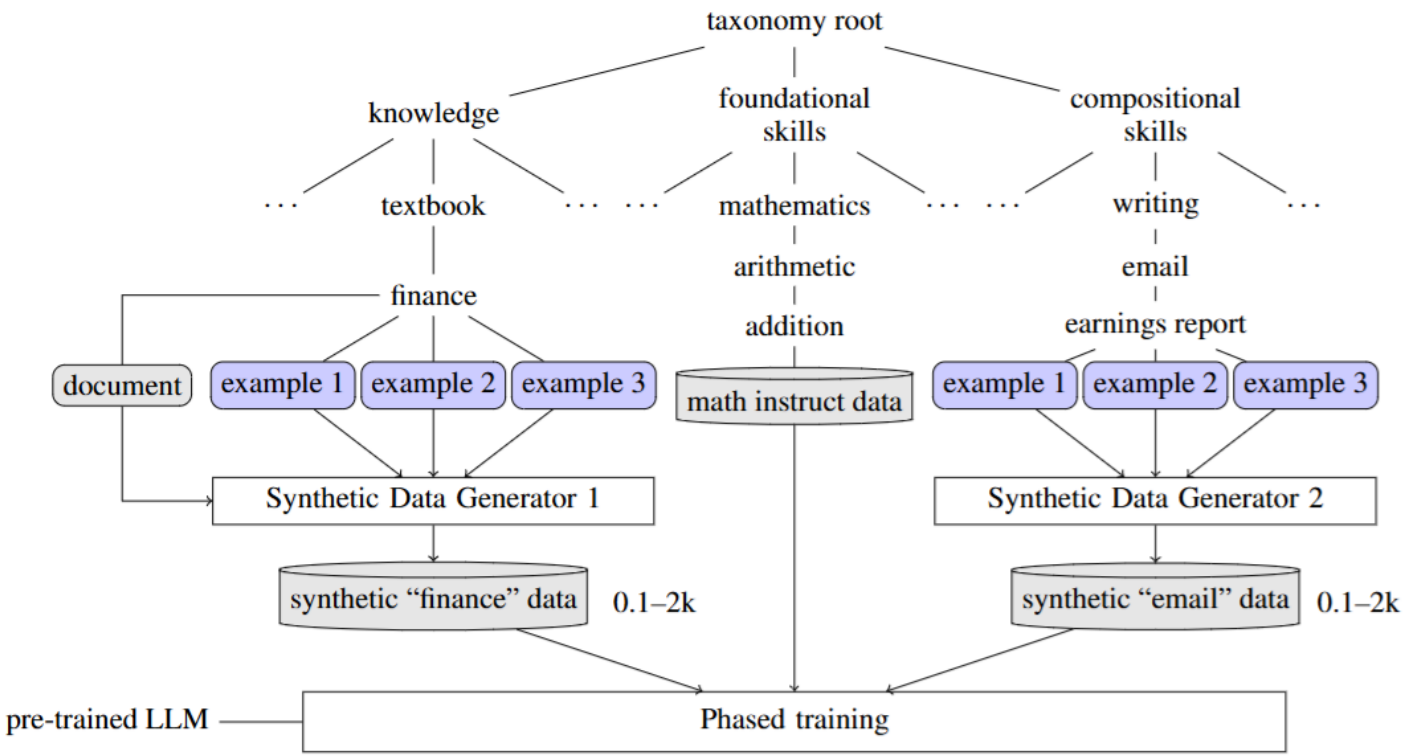
Rajbahadur et al., AIware Leadership Bootcamp, Toronto, Canada, 2024

# Flywheel in Action – OSS Instruct and Magicoder



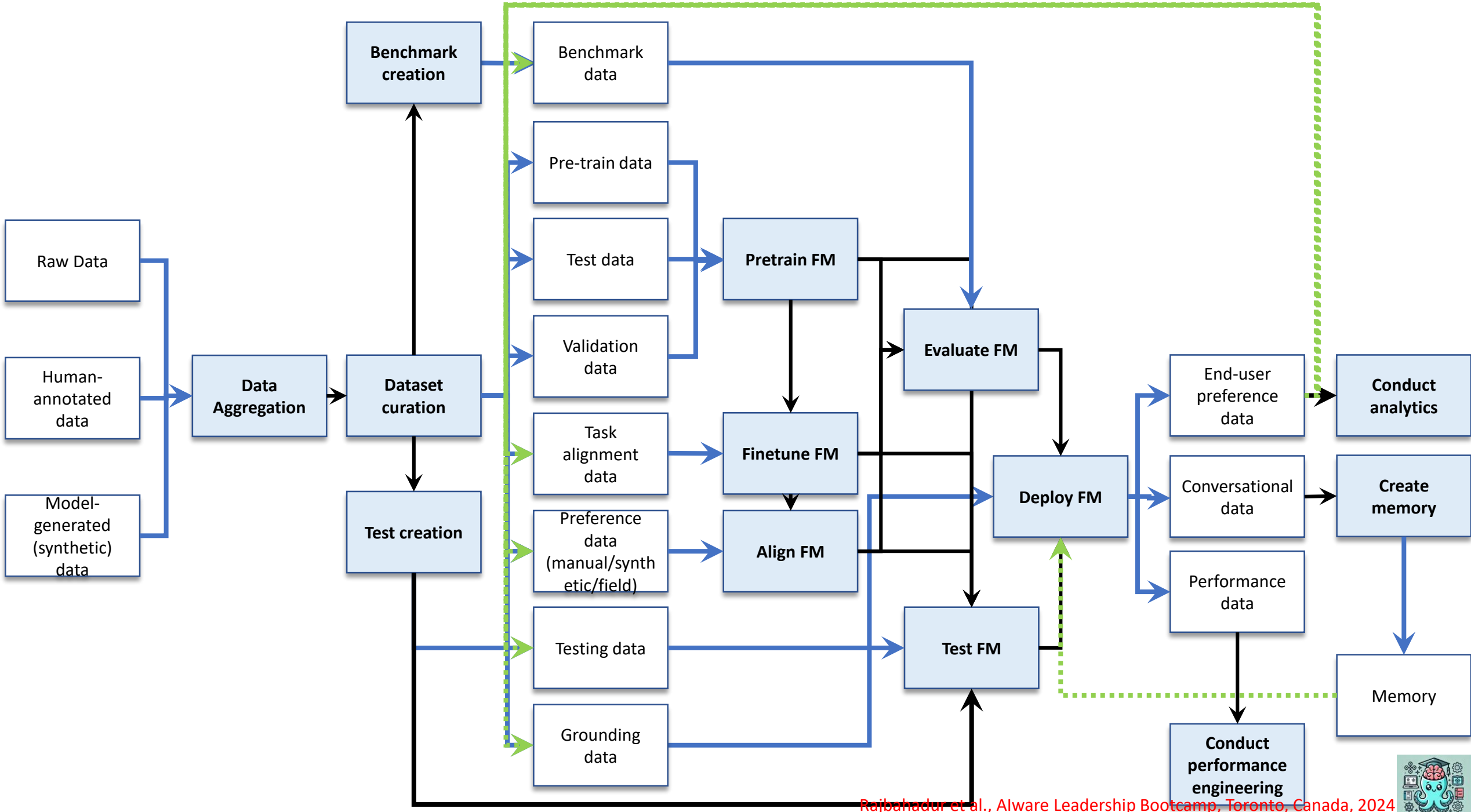

Wei et al., Magicoder: Empowering Code Generation with OSS-Instruct

# IBM InstructLab




- Decompose knowledge into Knowledge, foundational skills and compositional skills
- Leverage crowdsourcing to collect such Knowledge, skills and compositional skills for multiple domains in the form of question and answer pairs
- These skills and knowledge together acts as the curriculum for synthetic data generation
- The synthetically trained data is then used to fine-tune the model in a two-phase process

Xu et al., LAB: Large-Scale Alignment for ChatBots




Rajbahadur et al., AIware Leadership Bootcamp, Toronto, Canada, 2024