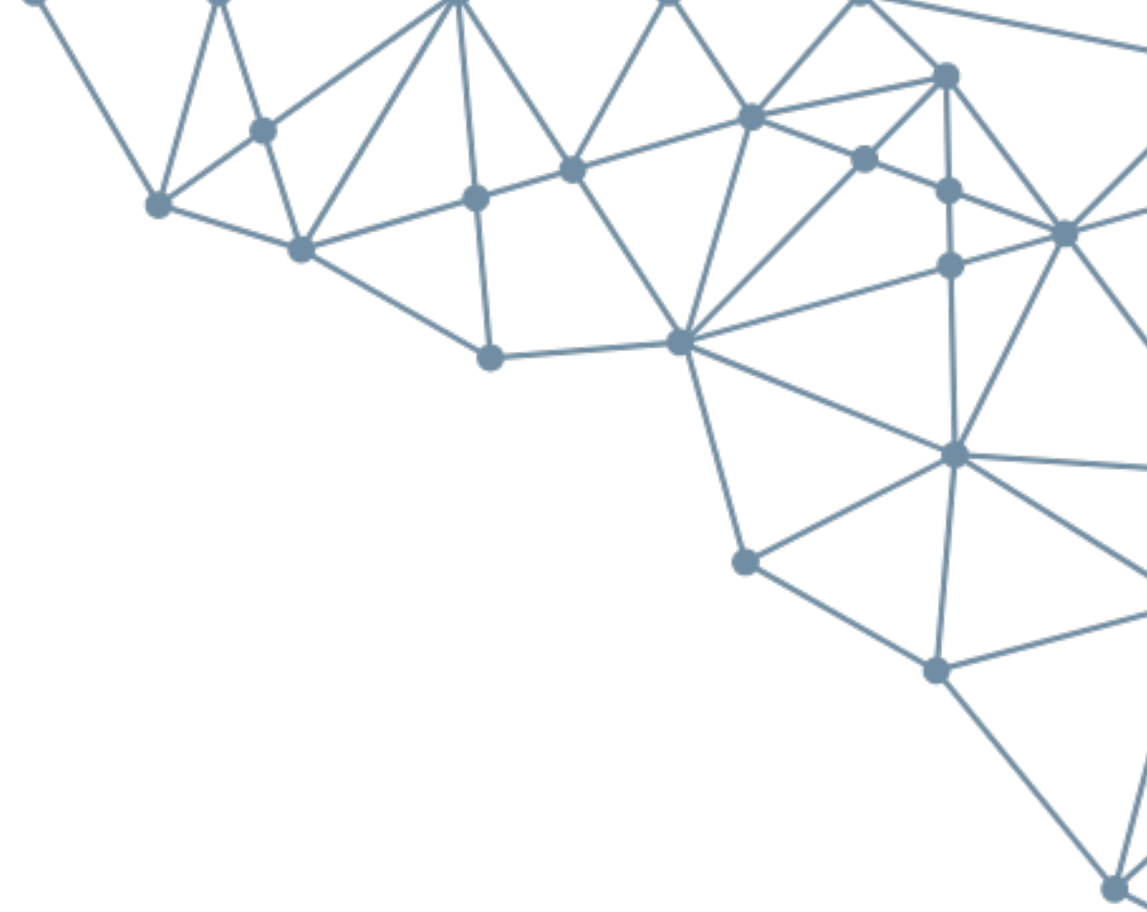


ABB - Session 6

AI Project Management

Shaw Talebi

Today's Session

1. **Housekeeping**
    1.1. Homework 5
    1.2. Course Recap
2. **AI Project Management**
    2.1. 5-step Project Framework
    2.2. Case study: YT Semantic Search Tool
    2.3. 4 Tips for Builders
3. **Epilogue**
    3.1. Note on LLM Loss Function
    3.2. Multimodal AI
    3.3. AI Agents

Homework 5

Shoutouts 🎉

AI Text Detector
Claudia Ng

Support Ticket Classifier
Serg

Stock Analysis Tool
Sangeeta Bahri

What we've learned…

AI = a computer's ability to solve problems and make decisions

Software 1.0 = rules are explicitly programmed into computers

- Python

Software 2.0 = computers learn programs by example (data)

- Machine learning, data engineering

Software 3.0 = adapting generic models for specific tasks

- LLMs, prompt engineering, RAG, text embeddings, fine-tuning

Meta Llama 3

What we've built...

Software 1.0

1. Replicating Maven Broadcasts
2. Automated Report Builder and Emailer

Software 2.0

3. ETL of Survey Data
4. Training an email classifier

Software 3.0

5. Summarizing Research Papers with GPT-4o
6. Text Classification with GPT-4o-mini
7. Local Document QA with Ollama
8. Analyzing Survey Data with Embeddings
9. Semantic Search with Embeddings
10. Blog QA Assistant with RAG
11. Fine-tuning BERT for Text Classification
12. Fine-tuning a LinkedIn Post Writer

File Organizer
Vladimir Belony

Resume Matcher
Saijai Osika

Report Image Extractor
Deborah Shutt

Automated Birthday Email Sender
Mathew Olajide

Personalized Mortgage Rate Emailer
Kalyan Mutyala

Football SuperLeague Leaderboard
Peirluigi Chiusolo

Bulk File Copier
Ronnie Rampersad

Radio Station Watcher
Sangeeta Bahri

Arxiv Paper Retriever
Fahad Ebrahim

News Feed Aggregator
Adam Rosenkoetter

Chess Grandmaster Ranking
Ludovic H

Python Project Summarizer
Bryce

iTunes Library Analysis
Rod Morrison

LI Internship Scraper and Emailer
Ewa Gros

Stock Picker
Vladimir Belony

Predicting Housing Prices
Rod Morrison

Support Ticket Classification
Serg

Predicting Natural Gas Prices
Sangeeta Bahri

Fraud Detection
Dario Zandolin

Credit Risk Modeling
Kalyan Mutyala

Reddit Post Classifier
Sanjeev NC

Cover Letter Enhancer
Ali Jnifen

Geopolitical Market Analyzer
Vladimir Belony

Sanjeev NC
Lecture Notes App

Stock Sentiment Analysis
Rod Morrison

Resume Summarizer
Juan Ignacio Beiroa

Arxiv Extractor & Summarizer
Fahad Ebrahim

Student Assessment Tool
Sangeeta Bahri

AI Accent Coach
Claudia Ng

YT Video Summarizer
Pierluigi Chiusolo

STNet Paper QA
Serg

Recipe Recommender System
Vladimir Belony

Applicant Matching System
Juan Ignacio Beiroa

Bluesky QA Bot
Claudia Ng

Stock Analysis Tool
Sangeeta Bahri

Foundational RAG System
Rod Morrison

Research Paper QA Bot
Serg

AI Text Detector
Claudia Ng

Support Ticket Classifier
Serg

Stock Analysis Tool
Sangeeta Bahri

Small projects are not enough.

While this is a great way to learn…

it breaks down as the stakes (and scale) go up.

I have a PDF I want to chat with

No problem.

I have 1000 PDFs I want to chat with

Uh…

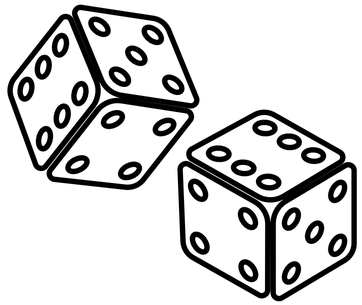
Solution: A structured framework to tackle any AI project

AI Project Management

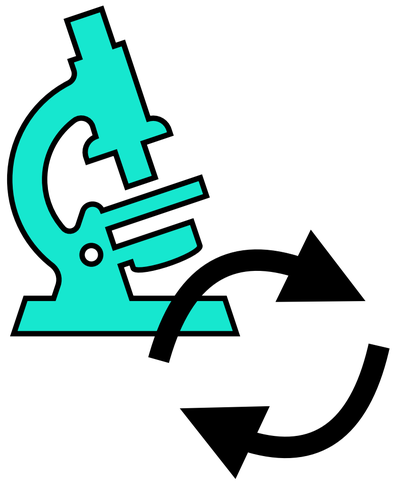
Why is AI different?

3 key differences from traditional software

1) Data are central



2) Probabilistic



3) Experimentation required

[1]

AI Project Management

A 5-step Framework

Step 1

Problem Definition & Scope

Step 2

Data Engineering

Step 3

Solution Development

Step 4

Solution Deployment

Step 5

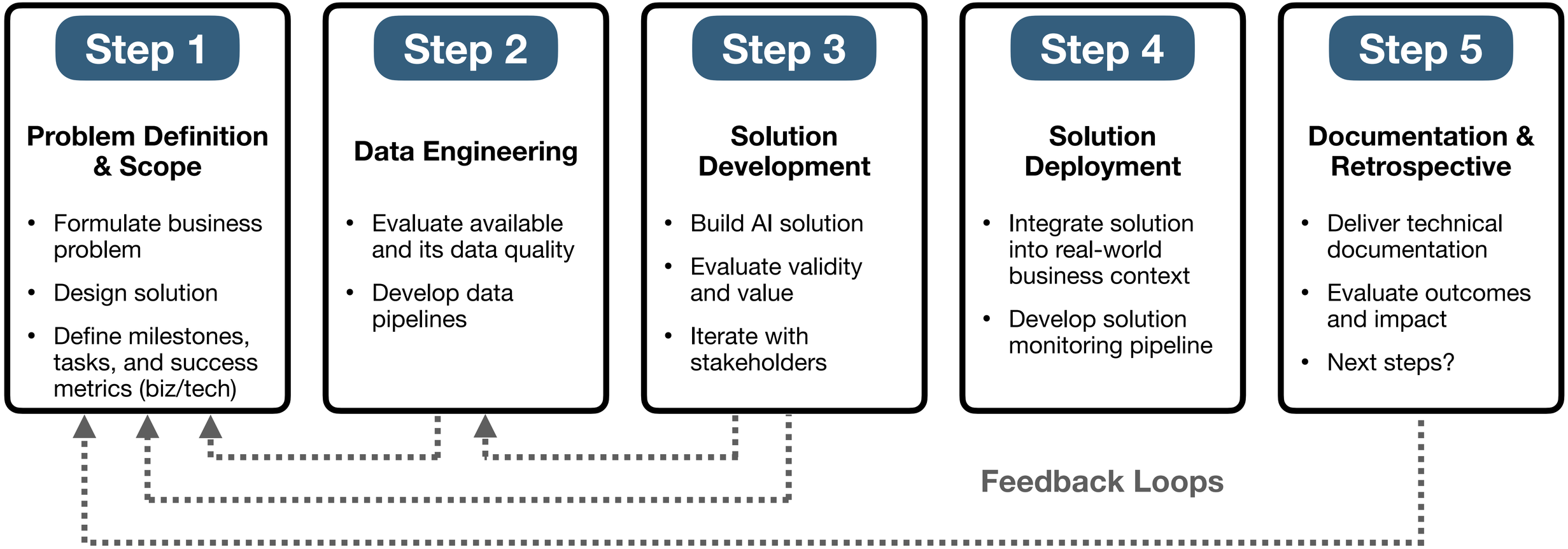
Documentation & Retrospective

Feedback Loops

[2]

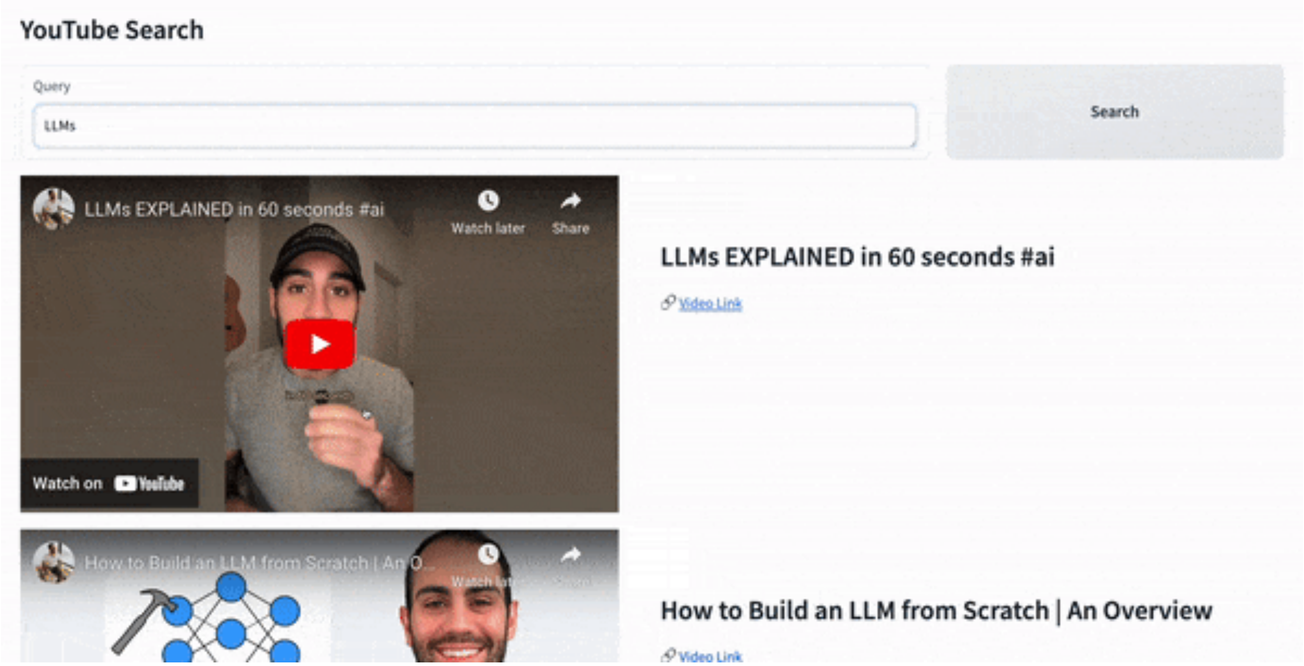
AI Project Management

A 5-step Framework



[2]

Case Study: YT Semantic Search Tool

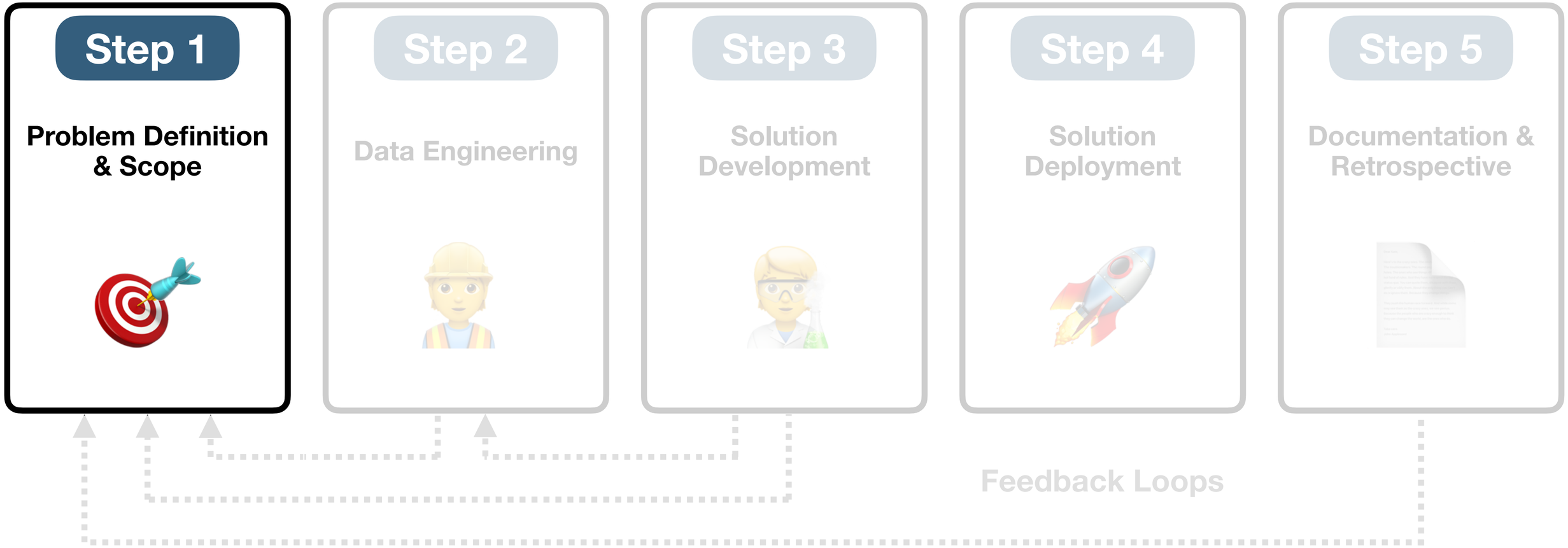


https://github.com/ShawhinT/yt-search

[3, 4]

Step 1: Problem Definition & Scope

A 5-step Framework



Step 1: Problem Definition & Scope

Formulating the business problem

What problem are you trying to solve?

Across YouTube, Medium, and GitHub I have about 300 technical resources for learners and builders. Although this is a robust set of resources, effectively navigating it is a challenge since these platforms don't talk to each other.

Difficult for audience to navigate all my content

Why is navigating your content a challenge?

While the search function on YouTube is pretty good, Medium and GitHub are horrible. Basically, people have to manually scroll and skim through dozens of items to find what they are looking for.

Why are the search functionalities on these platforms so bad?

Step 1: Problem Definition & Scope

Formulating the business problem

Why are the search functionalities on these platforms so bad?

Medium doesn’t have a way to search over an individual writer’s content, only search over the whole platform. GitHub has basic search, but it doesn’t support searching repo contents only the title and description.

Why hasn’t an integrated search solution like this been developed?

Why hasn’t an integrated search solution like this been developed?

Building such a solution is major time and effort investment, which may not be justified by the potential upside of the solution.

Why wouldn’t the investment be justified?

Step 1: Problem Definition & Scope

Formulating the business problem

Why wouldn't the investment be justified?

Quantifying the value creation is difficult because it's not clear how an improved search experience would impact other business metrics.

Why is the impact of improved search unclear?

Why is the impact of improved search unclear?

Generally, I can imagine that easier search would lead to more engagement, a larger audience, and eventually more course sales. However, I have no baseline to compare this to.

The lack of a centralized, user-friendly search system across YouTube, Medium, and GitHub makes it difficult for the audience to efficiently discover and engage with my content, and the unclear ROI of solving this issue has delayed investment in a solution.

Step 1: Problem Definition & Scope

Designing the (AI) solution

Create polls on YouTube and LinkedIn to gauge interest in tool

Create mockup of idea and post it on socials

Create curated PDF/blog guides for specific avatars (track traffic/downloads)

Create landing page with mock up and have people sign up for notifications

Do 1:1 interviews with audience members

Manually create Notion or Google Sheet with all content in one place

No AI or coding required

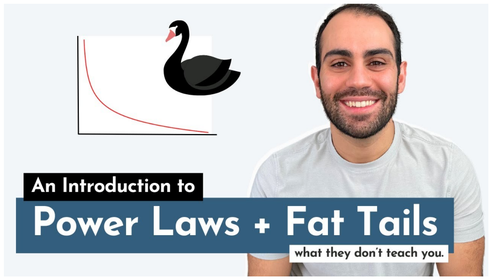
Build small web app where users can search and filter content **Software 1.0**

Build semantic search app prototype for content **Software 2.0/3.0**

Step 1: Problem Definition & Scope

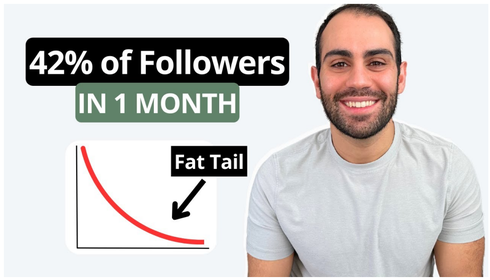
Designing the (AI) solution—Front-end

What are fat tails?



Pareto, Power Laws, and Fat Tails—what they don't teach you in STAT

🔗 Link

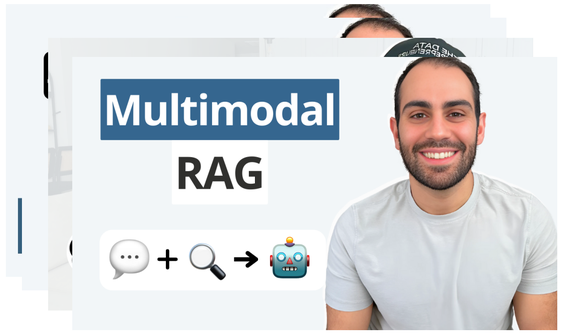


4 Ways to Measure Fat Tails with Python (+ Example Code)

🔗 Link

Step 1: Problem Definition & Scope

Designing the (AI) solution—Back-end



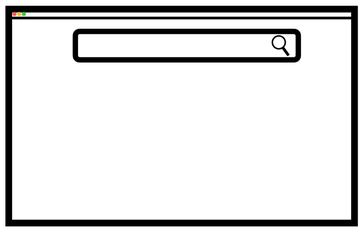
YouTube

Transcripts

Embeddings

Similarity Search

Output Results



Front-end

"Multimodal RAG"

Input Query

Embedding

Step 1: Problem Definition & Scope

Implementation Plan - Project Requirements

Project Requirements	
Roles	• Project Manager • Data Engineer • Data Scientist • ML Engineer
Data	• **Evaluation Dataset**: Table of 50 query-video pairs to evaluate the quality of the search feature. Fields: query, YouTube video ID
Infrastructure	• Google Cloud Run (est: $5/mo)
Technologies	• Python, YouTube API, YouTube-transcript-api, polars, Sentence Transformers, FastAPI, Docker, Gradio • GitHub repository for project code and documentation • YouTube API key

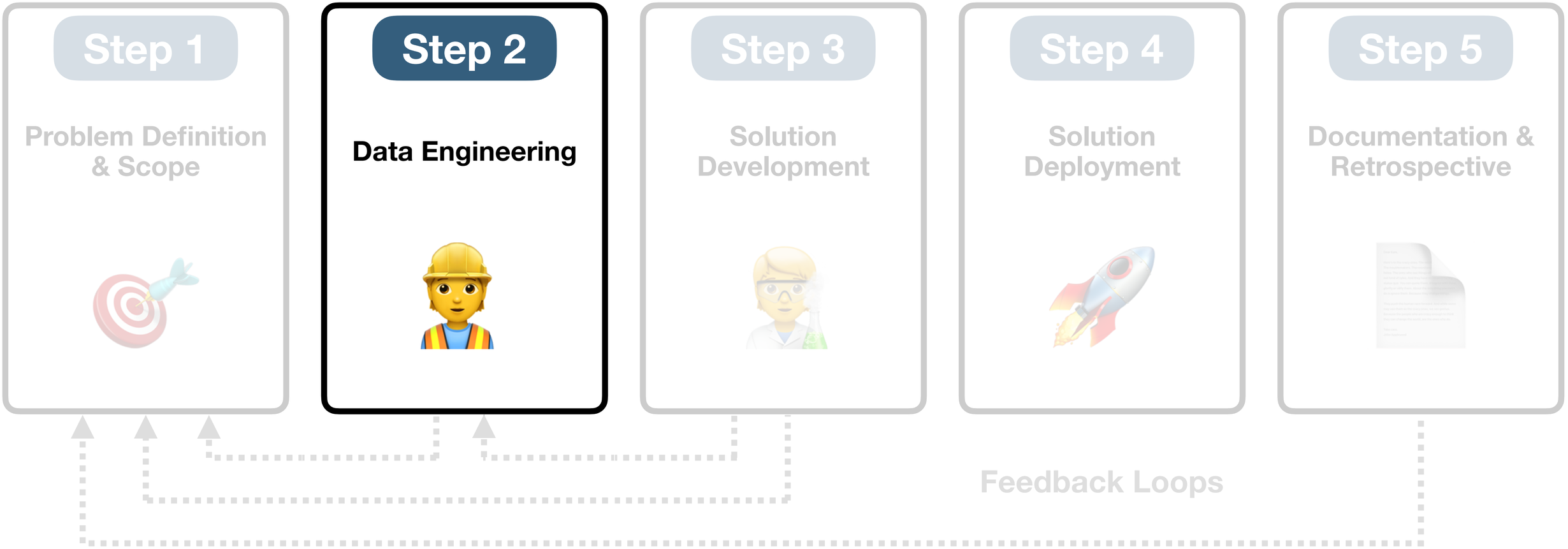
Step 1: Problem Definition & Scope

Implementation Plan - Project Requirements

Phase	Task	Role	Due Date
Phase 1	Extract list of YouTube video IDs and links from my channel using the YouTube API	DE	5/4
	Download automatically generated captions for all YouTube videos given their video IDs using the YouTube-transcript-api Python library	DE	5/4
	Create index containing YouTube video ID, video captions, and video link (parquet file)	DE	5/4
	Append video captions to evaluation dataset and save as parquet file.	DE	5/4
Phase 2	Use multiple candidate models to compute text embeddings for all video captions	DS	5/11
	Develop search function which takes a user query and performs similarity search over video captions. Input: natural language query. Output: video IDs of top search results.	DS	5/11
	Compare the performance of multiple open-source text embedding models using evaluation dataset. Performance metrics: numerical ranking of correct result, binary in-top-k flag.	DS	5/11
	Append text embeddings to video index (fields: YouTube video ID, video captions, video link, and caption text embeddings).	DS	5/11
	Create search function that works directly with video index	DS	5/11
Phase 3	Create API for search function using FastAPI	MLE	5/18
	Containerize search function and index using Docker	MLE	5/18
	Create Gradio UI that can interact with search API (for local validation)	MLE	5/18
	Deploy Docker container on Google Cloud Run	MLE	5/18
	Validate deployment using Gradio UI	MLE	5/18
Phase 4	Create technical documentation on GitHub	PM	5/25
	Project retrospective	PM	5/25

Step 2: Data Engineering

A 5-step Framework



Step 2: Data Engineering

Training and Evaluation Data

Training Data

n/a

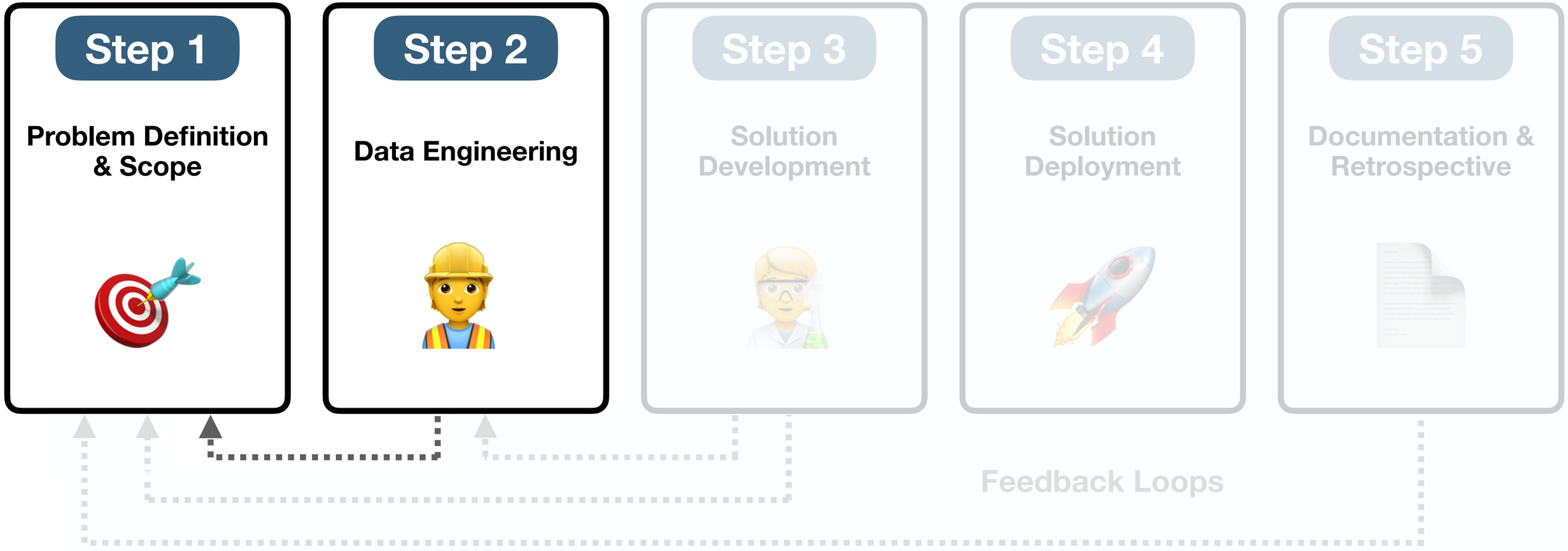
Evaluation Data

query	video_id
ai consulting	INICLmWlojY
fine tuning llm	eC6Hd1hFvos
When do you recommend fine tunning and when do you recommend vector database?	eC6Hd1hFvos
llm from scratch	ZLbVdvOoTKM
What if you could make a small language model, that maybe only understand english, can understand code, and is easy to run?	ZLbVdvOoTKM
gmail signature	NjMD1bGBNqw

Manually created no DE needed

If DE needed for train/eval data…

One may discover something that requires returning to Step 1.

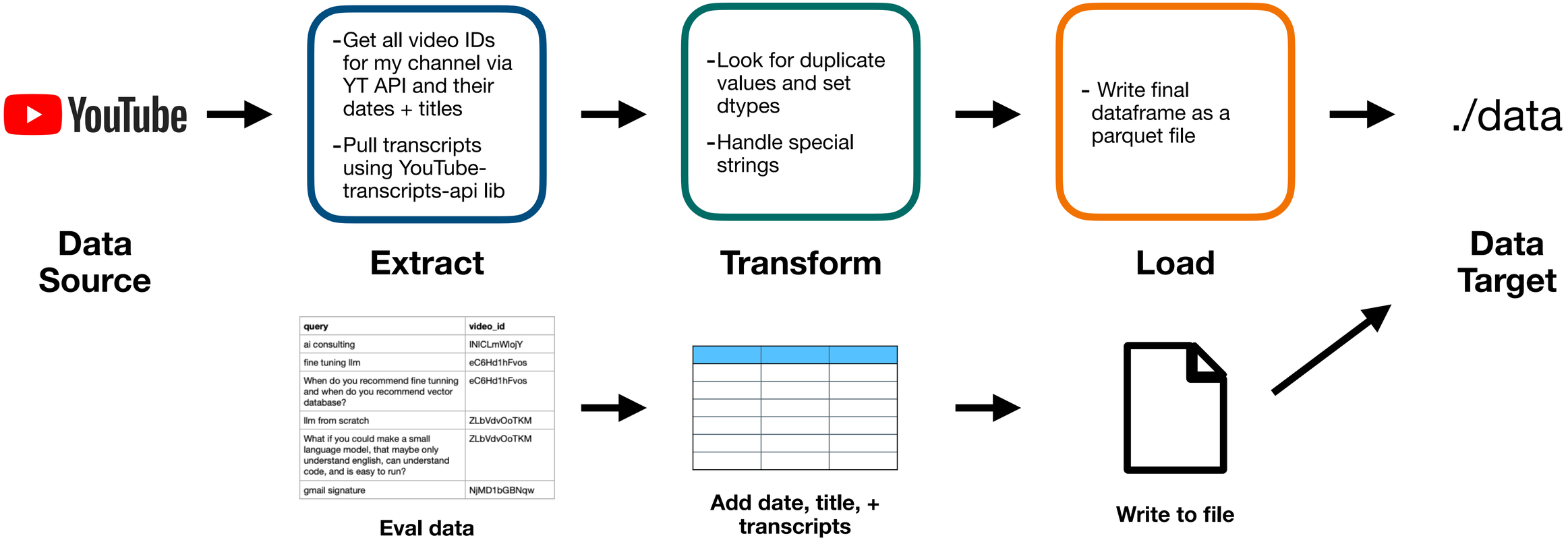


Step 2: Data Engineering

Developing Data Pipeline

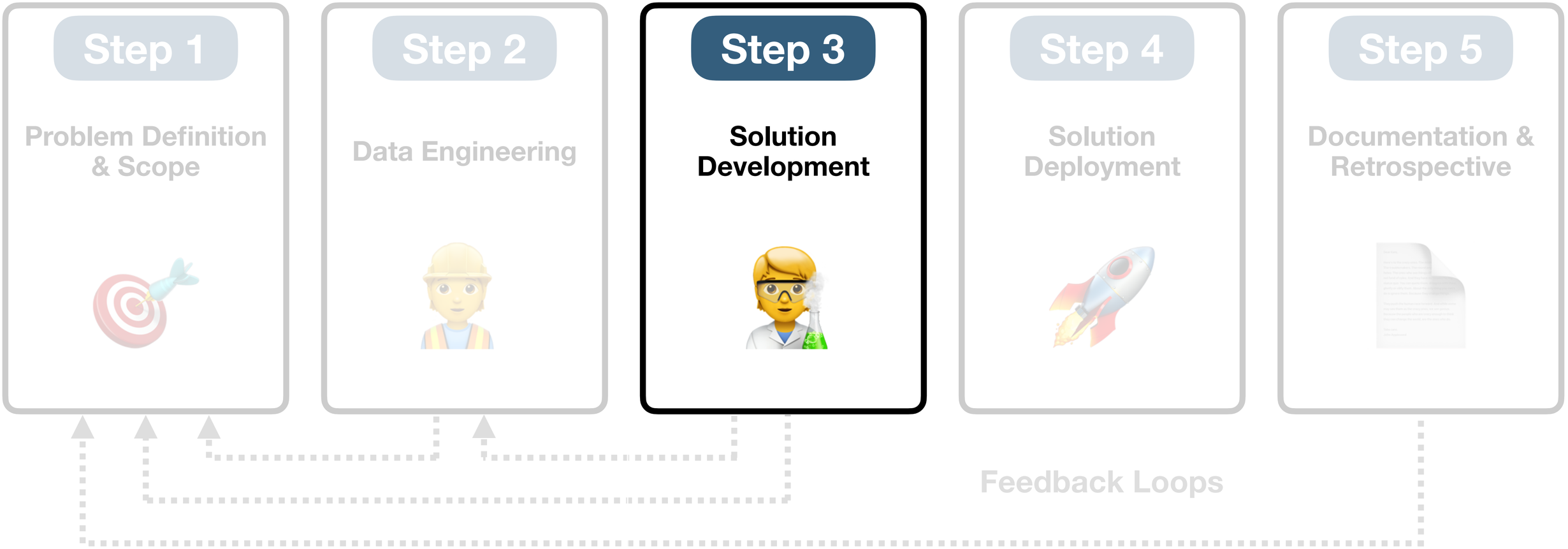
Step 2: Data Engineering

Developing Data Pipeline



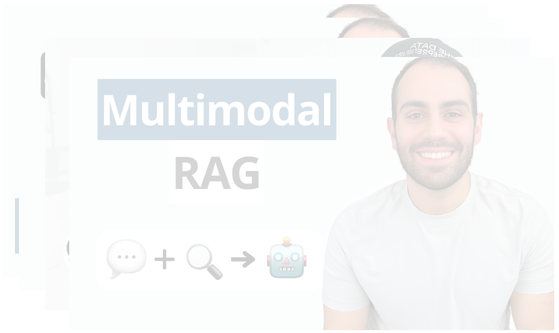
Step 3: AI Development

A 5-step Framework



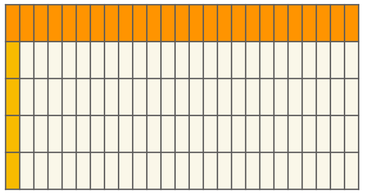
Step 3: AI Development

Build AI Solution



YouTube

Transcripts



Embeddings

Similarity Search

Output Results

Front-end

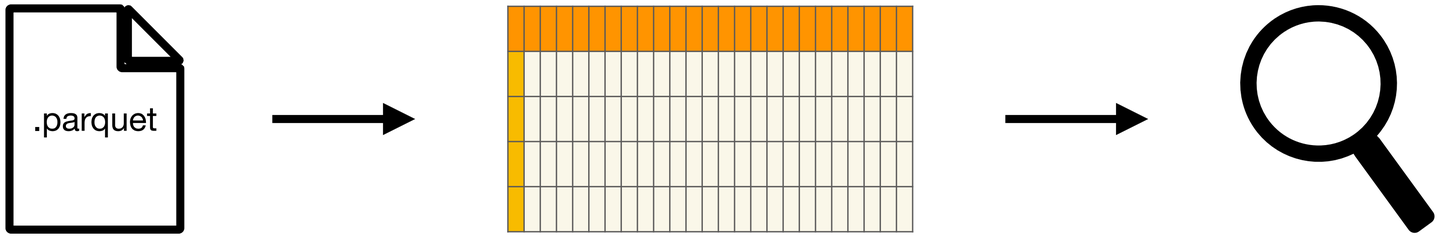
"Multimodal RAG"

Input Query

Embedding

Step 3: AI Development

Build AI Solution



Video Data
(Video ID, Title, Transcript)

Embeddings

Similarity Search

1) Title, Transcripts, or both (3 options)

2) Embedding Model? (3 options)

3) Similarity/Dist Metric? (5 options)

3 × 3 × 5 = 45 configurations!

Think like a scientist!

Step 3: AI Development

Experimentation and Evaluation

3 × 3 × 5 = 45 configurations!

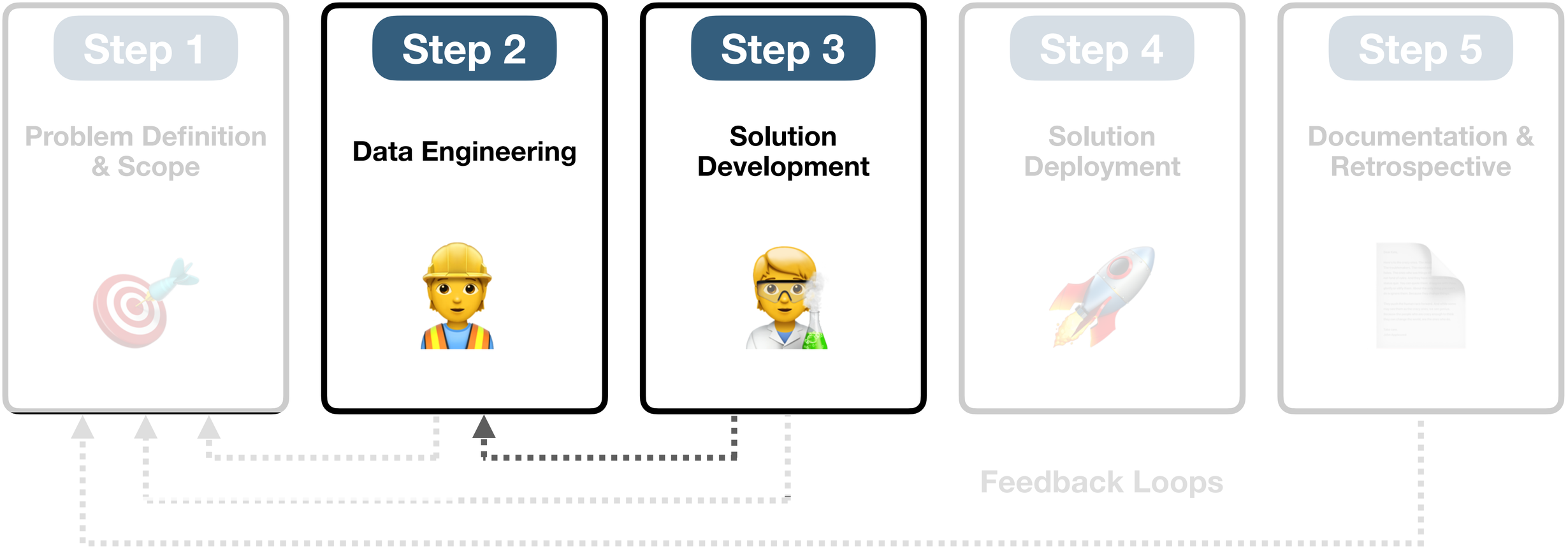
Run each on eval data

#	Text	Embedding Model	Similarity/Dist Metric	Mean Rank of ground truth	# Ground Truth in Top 1	# Ground Truth in Top 3
1	Title only	Model 1	Metric 1	2.37	31	39
2	Transcript only	Model 1	Metric 1	5.13	25	35
3	Title + Transcript	Model 1	Metric 1	0.54	48	50
4	Title only	Model 2	Metric 1	5.0	25	34
5	Transcript only	Model 2	Metric 1	3.14	35	39
6	Title + Transcript	Model 2	Metric 1	4.77	29	34
7	Title only	Model 3	Metric 1	2.08	37	45
8	Transcript only	Model 3	Metric 1	5.67	22	28
9	Title + Transcript	Model 3	Metric 1	1.64	41	49
…	…	…	…	…	…	…

[1]

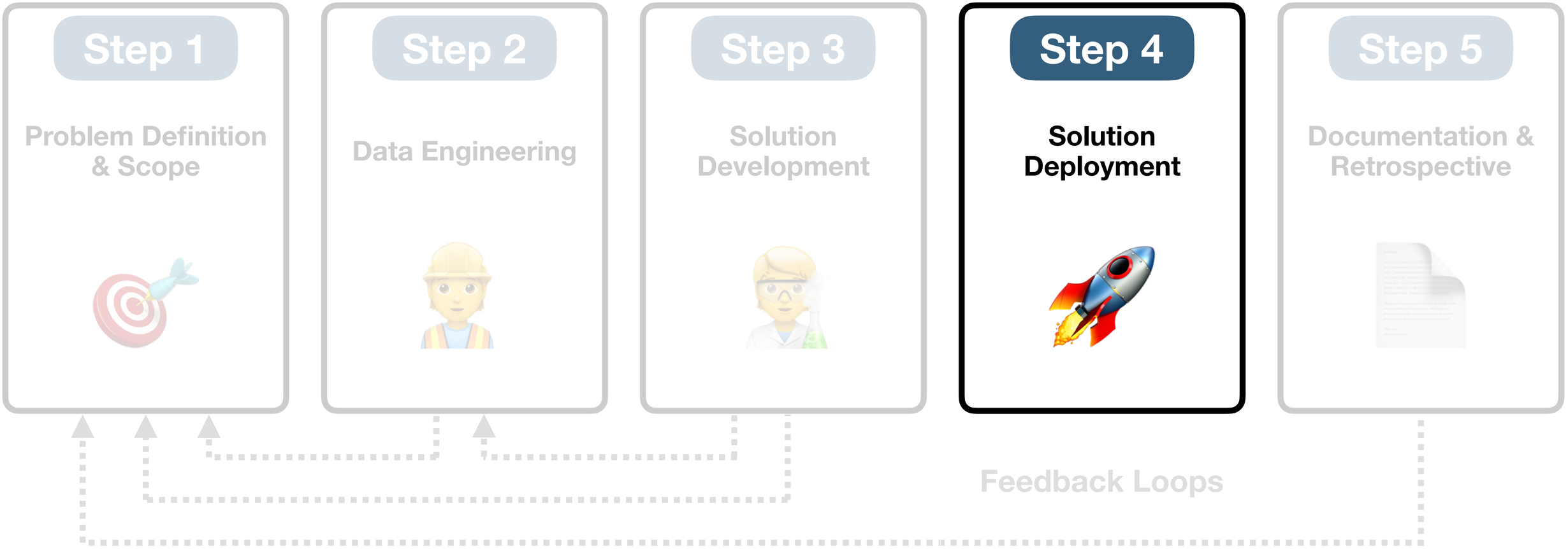
Update Data Pipeline…

Add embeddings to pipeline



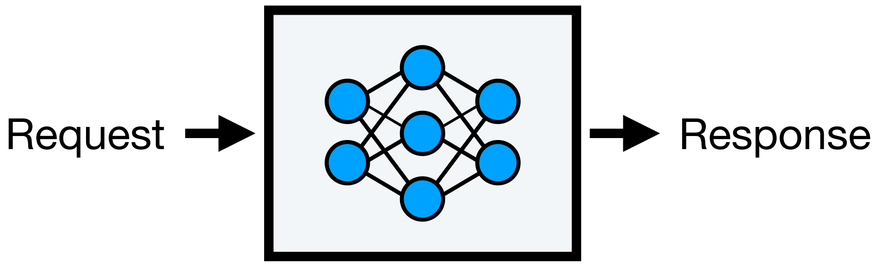
Step 4: AI Deployment

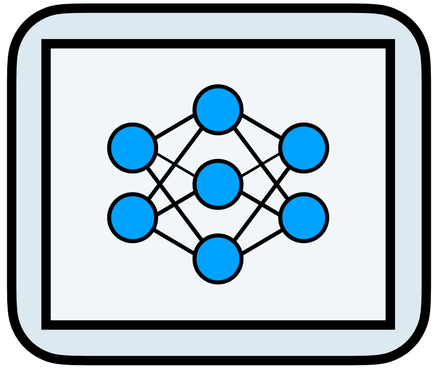
A 5-step Framework

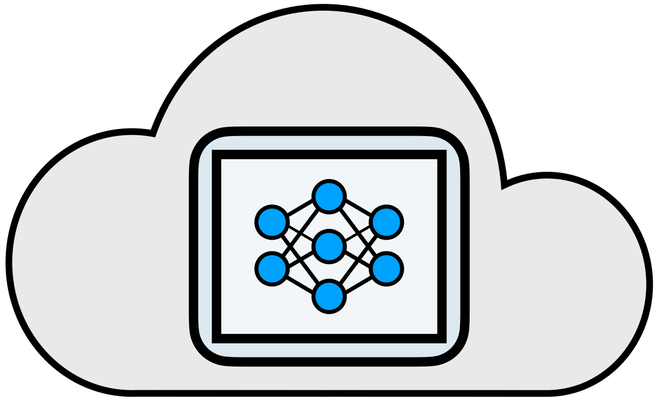


Step 4: AI Deployment

A simple 3-step deployment strategy







1) Create API

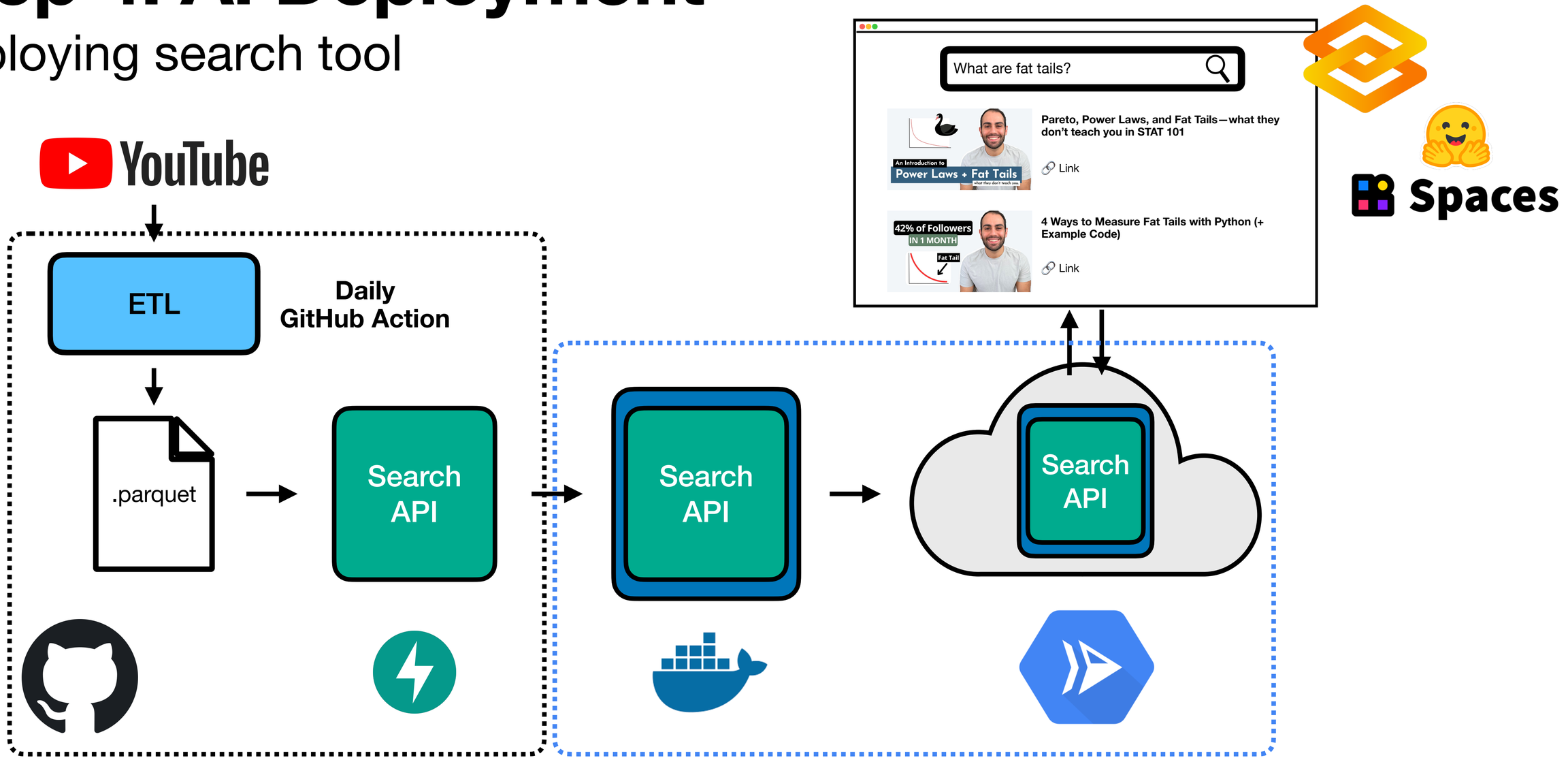
2) Containerize

3) Deploy

[7]

Step 4: AI Deployment

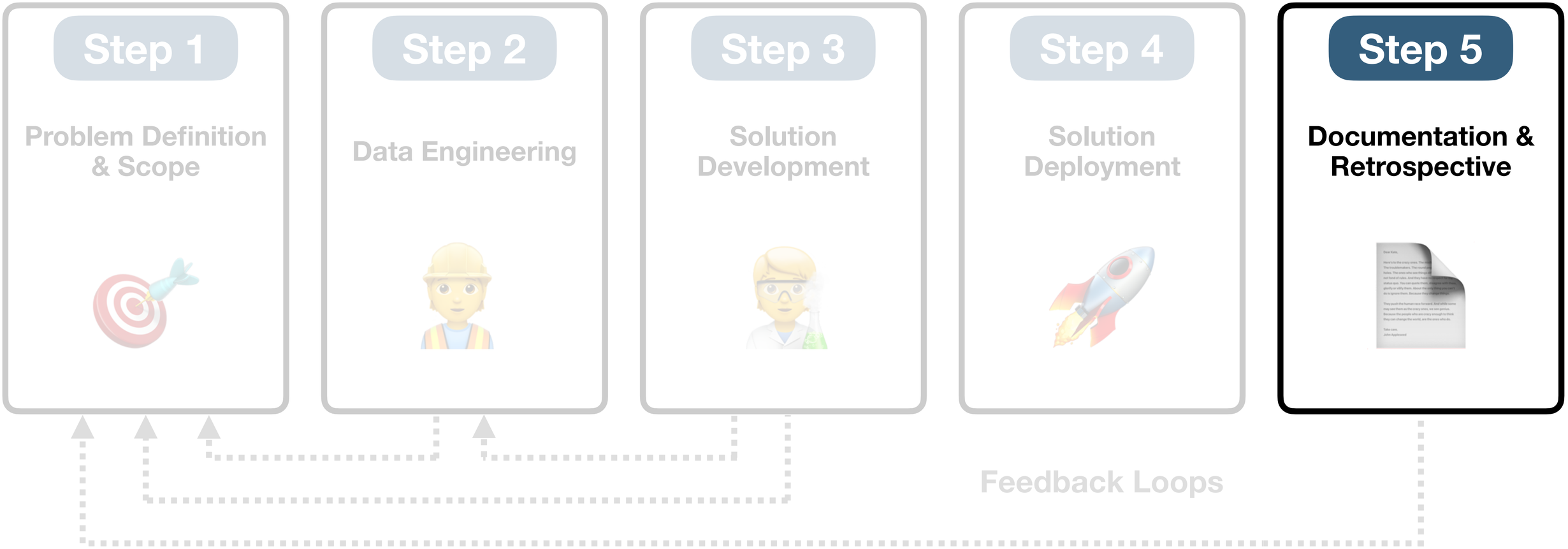
Deploying search tool



[6, 7]

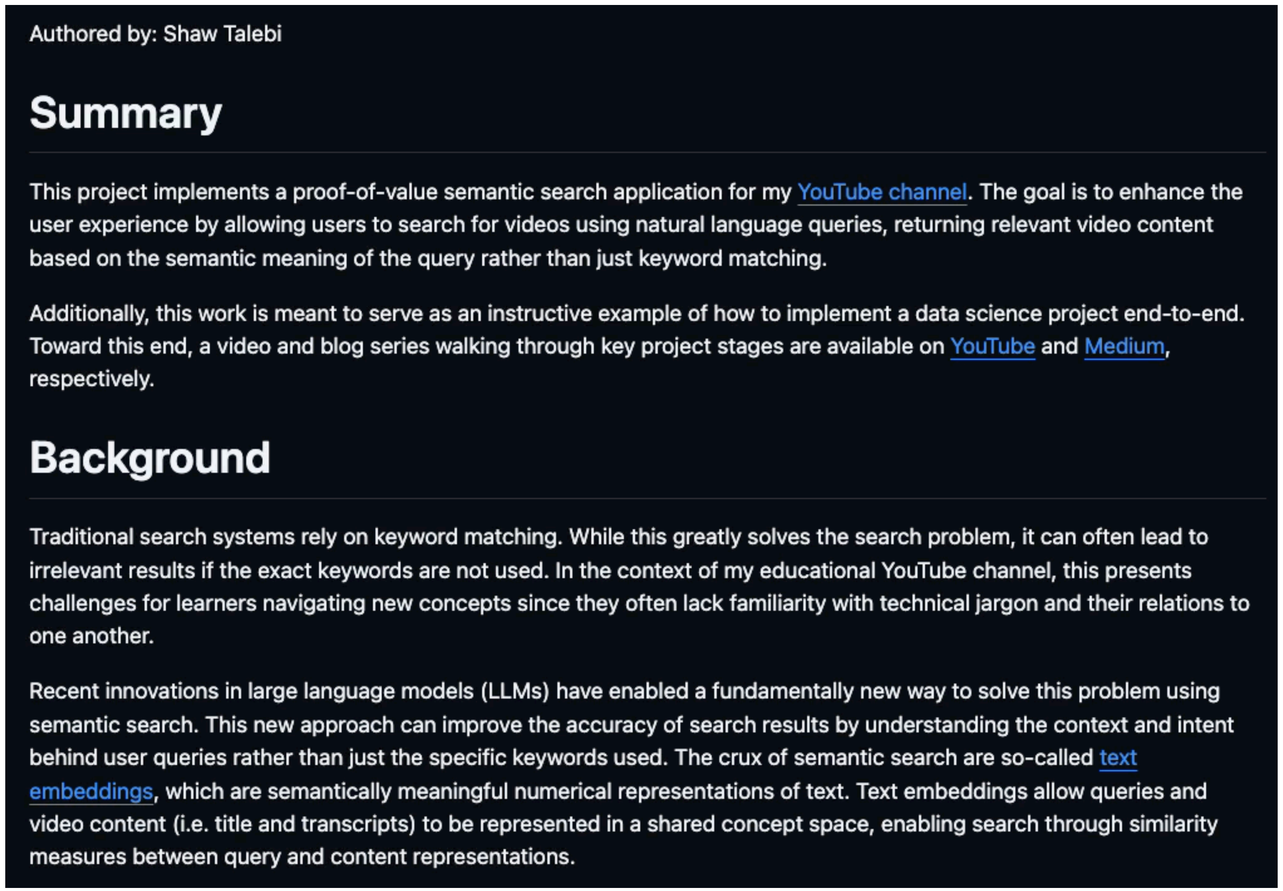
Step 5: Retrospective & Doc

A 5-step Framework



Step 5: Retrospective & Doc

Technical documentation



Authored by: Shaw Talebi

Summary

This project implements a proof-of-value semantic search application for my YouTube channel. The goal is to enhance the user experience by allowing users to search for videos using natural language queries, returning relevant video content based on the semantic meaning of the query rather than just keyword matching.

Additionally, this work is meant to serve as an instructive example of how to implement a data science project end-to-end. Toward this end, a video and blog series walking through key project stages are available on YouTube and Medium, respectively.

Background

Traditional search systems rely on keyword matching. While this greatly solves the search problem, it can often lead to irrelevant results if the exact keywords are not used. In the context of my educational YouTube channel, this presents challenges for learners navigating new concepts since they often lack familiarity with technical jargon and their relations to one another.

Recent innovations in large language models (LLMs) have enabled a fundamentally new way to solve this problem using semantic search. This new approach can improve the accuracy of search results by understanding the context and intent behind user queries rather than just the specific keywords used. The crux of semantic search are so-called text embeddings, which are semantically meaningful numerical representations of text. Text embeddings allow queries and video content (i.e. title and transcripts) to be represented in a shared concept space, enabling search through similarity measures between query and content representations.

Why?

- Helps others (and your future self) build on top of work
- Gives you clarity about what you built

Pro Tip: Doc as you go (little and often)

[3]

Step 5: Retrospective & Doc

Evaluate outcomes and next steps

Impact & Outcomes

- I learned a lot
- Video series viewed by 50k people + gained 1200 new subs.
- Little usage of tool may indicate this isn't a major pain point

What went well?

Search function works well despite little optimization

GitHub Actions and HF Spaces provides no-cost, easy to use deployment solutions

What can be improved?

First two queries are slow to generate since container spins up and down

Low traffic (622 lifetime users) on HF Spaces

What's next?

Take a step back and investigate possible pain points through audience polls and 1:1 interviews.

Integrate search tool into higher traffic page like my website and track usage with Google Analytics

Improve search function by adding content type and topic tag filtering

Improve search function by incorporating other search dimensions such as recency and popularity.

4 Tips for Builders

1) Focus on problems not technologies 🤔 > 🔨

2) Start small and iterate

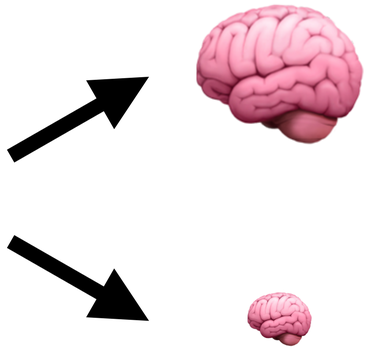
3) Move fast and experiment

4) 1 failure doesn’t erase 9 successes ❌ ≠ ✅✅✅✅✅✅✅✅✅

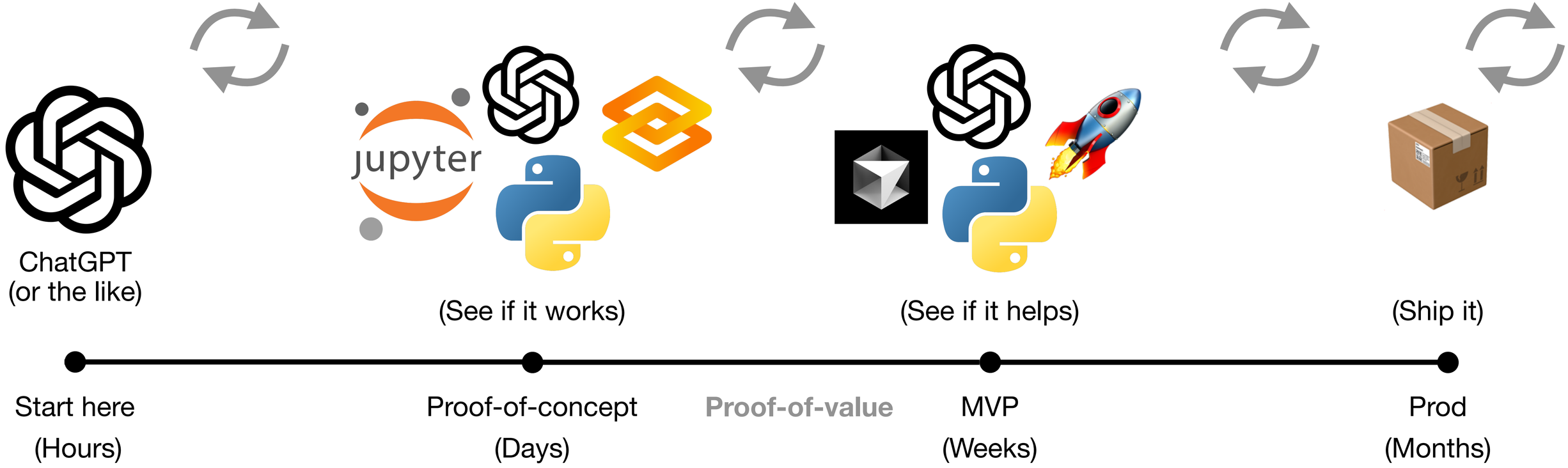
Tip 1: Focus on Problems (not tech)

“When you have a really nice hammer, everything looks like a nail”

The lack of a centralized, user-friendly search system across YouTube, Medium, and GitHub makes it difficult for the audience to efficiently discover and engage with my content, and the unclear ROI of solving this issue has delayed investment in a solution.



Tip 2: Start Small and Iterate

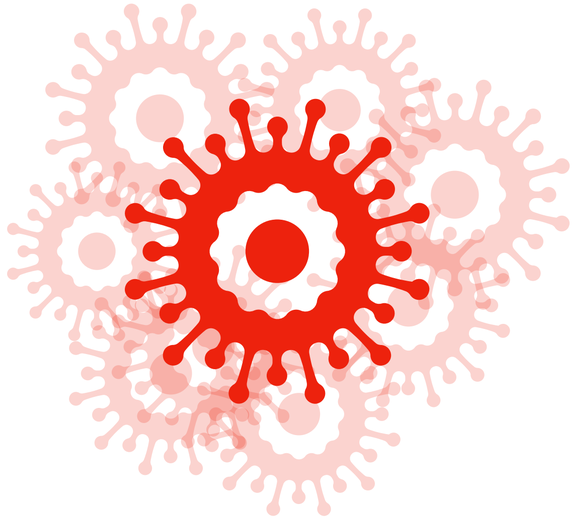


Don’t worry about scaling or optimizing until you have validation

Tip 3: Move fast and experiment

Iteration beats intelligence.

- Ray Dalio

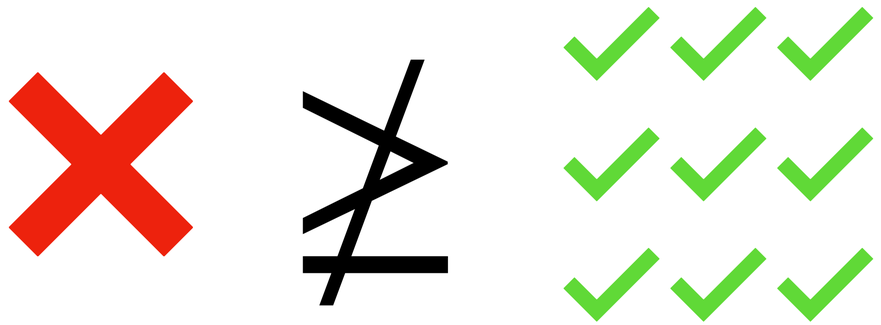


vs

[9, 10]

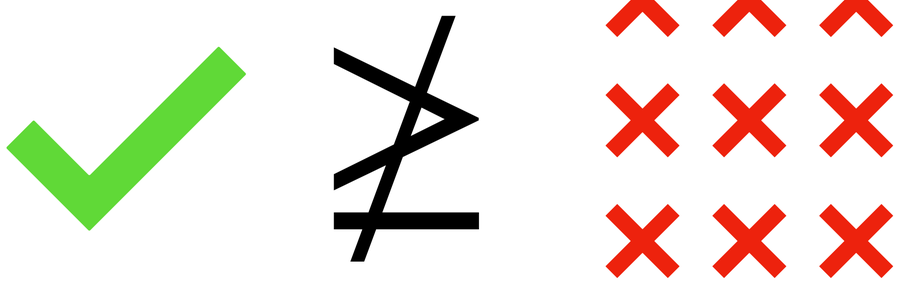
Tip 4: 1 Failure ≯ 9 Successes

Avoid this!



Cancer Doctor

Pursue this!



Cancer Researcher

Epilogue

LLM Loss Function

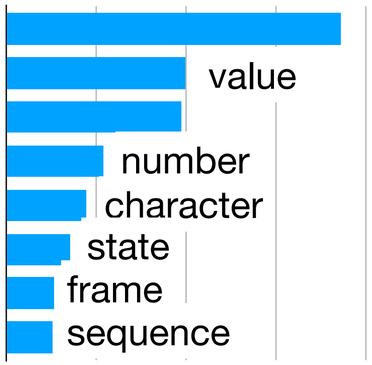
The math behind next-token prediction

Inference:

predict the next → LLM → word

Going deeper:

predict the next → LLM →



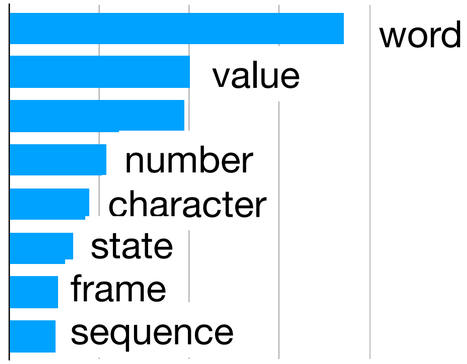
Generation = Sampling from this distribution

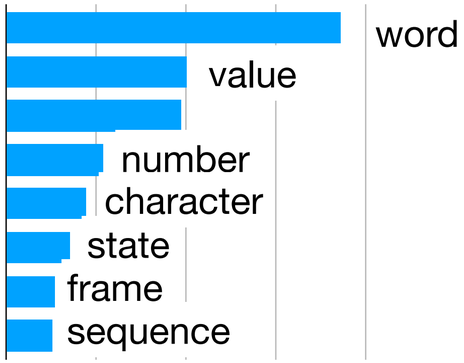
LLM Loss Function

The math behind next-token prediction

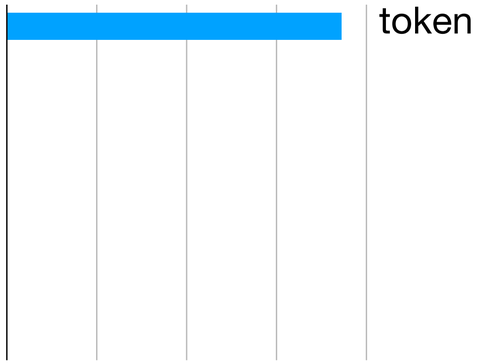
Training:

predict the next → LLM →





Model prediction



Ground truth

$$\mathscr{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

y_i = ground truth label for i^{th} token

$\hat{y}_i$ = predicted logit for i^{th} token

N = num tokens in vocabulary

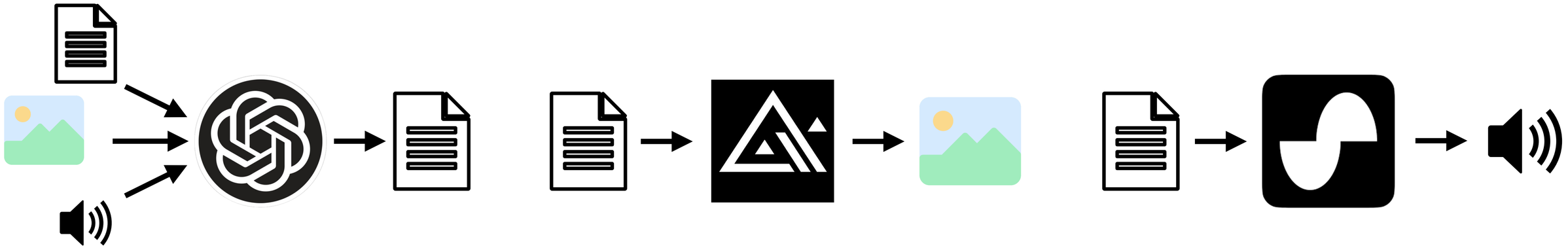
For hard labels $\Longrightarrow$

$$\mathscr{L} = -\log(\hat{y}_t)$$

Where, t = index of ground truth token

Multimodal AI

Models that can process multiple data types



GPT-4o

X-to-text

FLUX

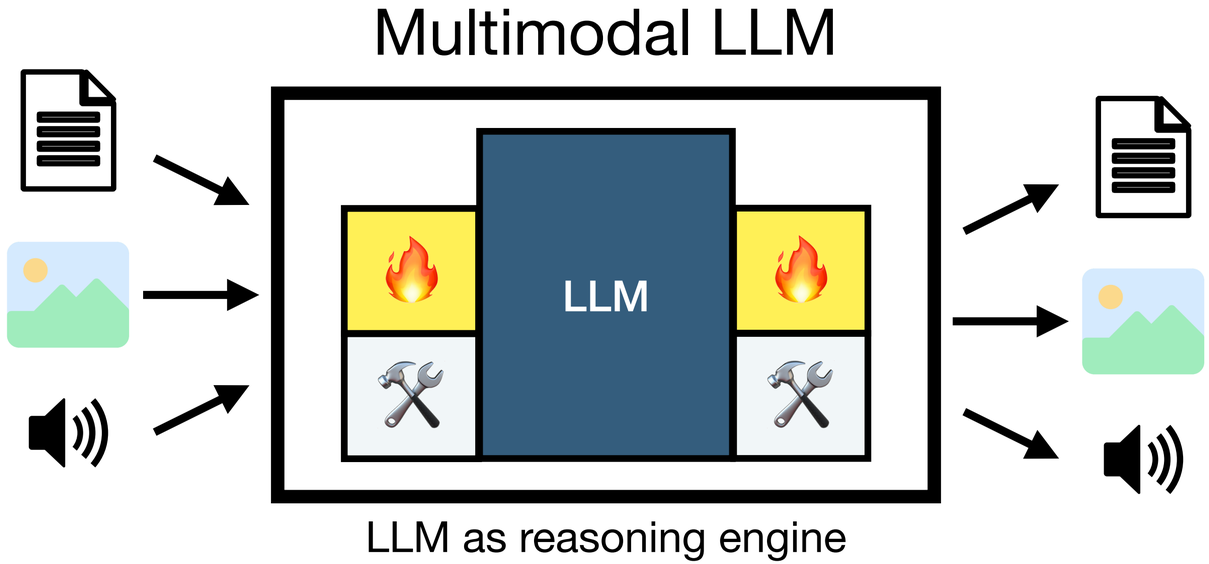
Text-to-image

Suno AI

Text-to-music

Multimodal (Large) Language Models

Multimodal models built on top of LLMs



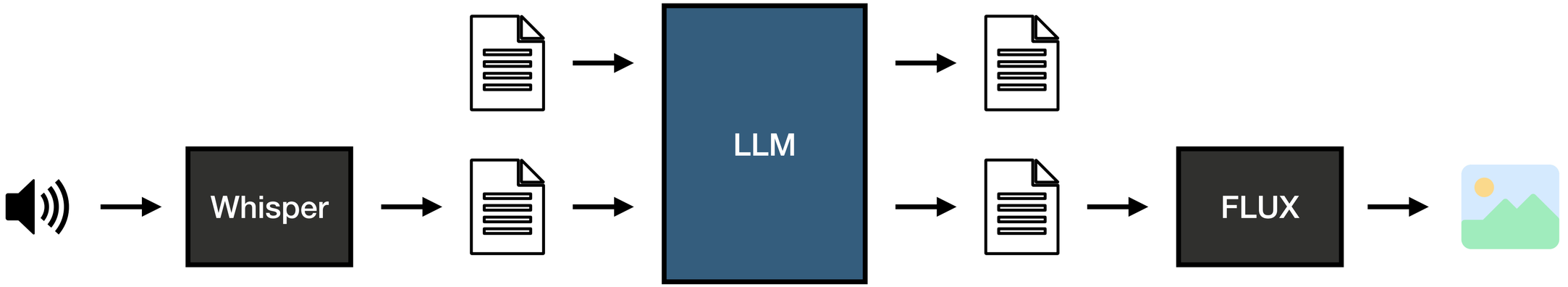
Why?

LLMs have strong ability to acquire world knowledge

0-shot capabilities

Path 1: LLM + Tools

Add external modules to do X-to-text or text-to-X



Pros

Simple to implement

No training data needed

Cons

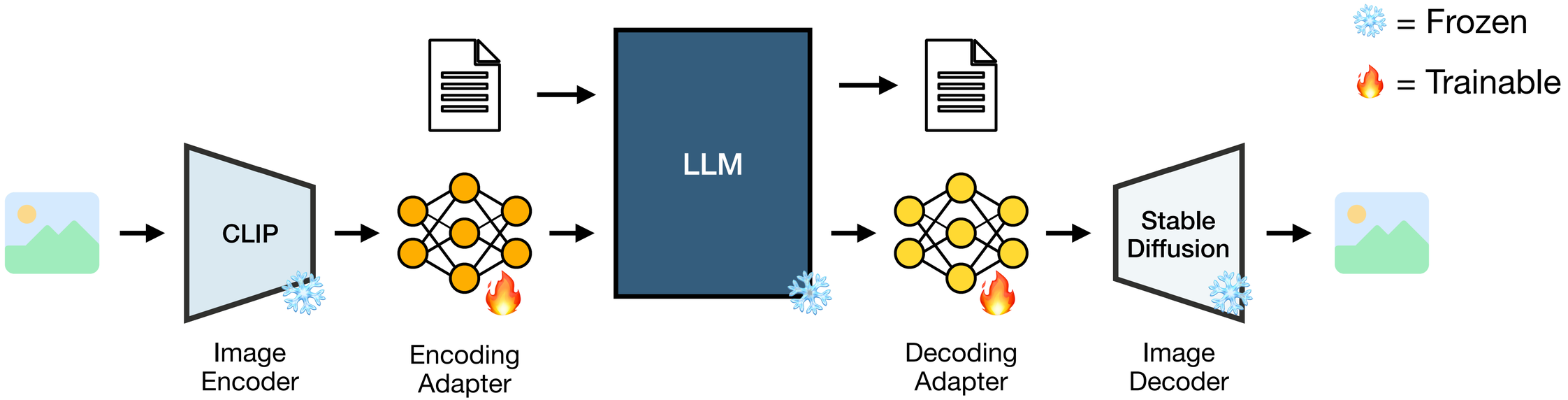
Limited capabilities

Difficult to customize

[11]

Path 2: LLM + Adapters

Add encoders/decoders which are aligned via fine-tuning



Pros

Better customization

Data efficiency

Cons

Requires training data

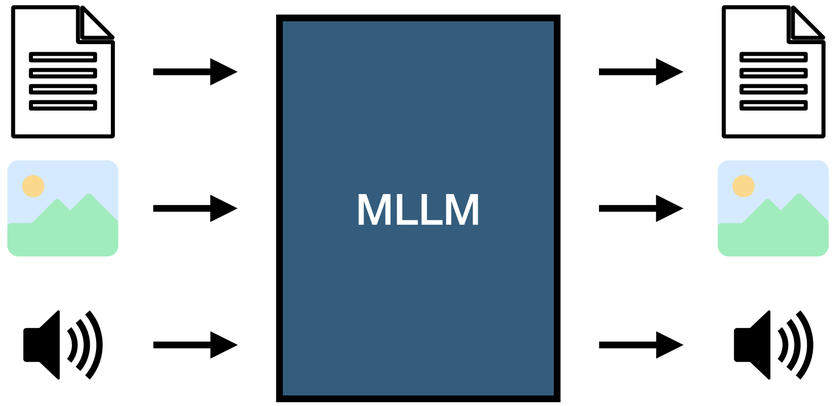
Technically sophisticated

Examples: LLaVA, LLaMA 3.2 Vision, Flamingo, MiniGPT4, Janus, Mini-Omni2, and IDEFICS

[11]

Path 3: Unified Models

Mixing modalities at pre-training (i.e. training from scratch)



Pros

Seemless modality integration

Faster inference times

Cons

Advanced technical challenges

Required massive data + compute

Examples: GPT-4o, Gemini, Emu3, BLIP, and Chameleon

[11]

AI Agents

Systems that can perform self-guided actions (my definition)

Other definitions…

Automation with fuzzier instructions

AI system that can perform tasks like a human

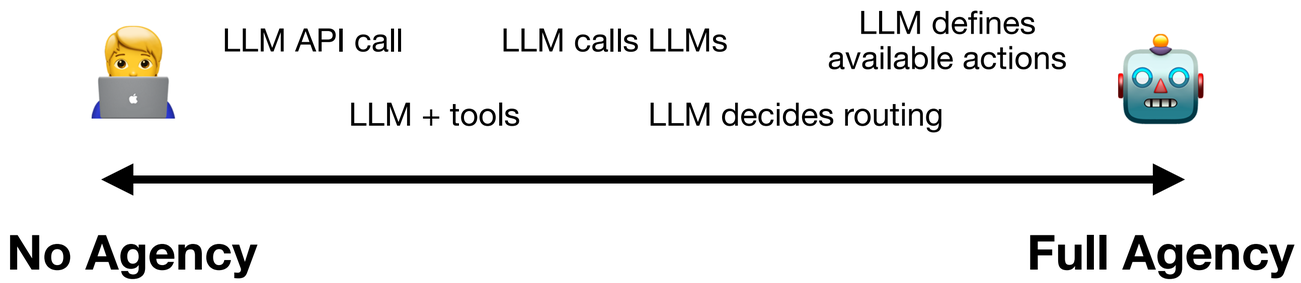
LLM that can prompt itself

LLM that can use tools

LLM that can talk to other LLMs

LLM decides control flow of app

Agentic Systems: agency is on a spectrum

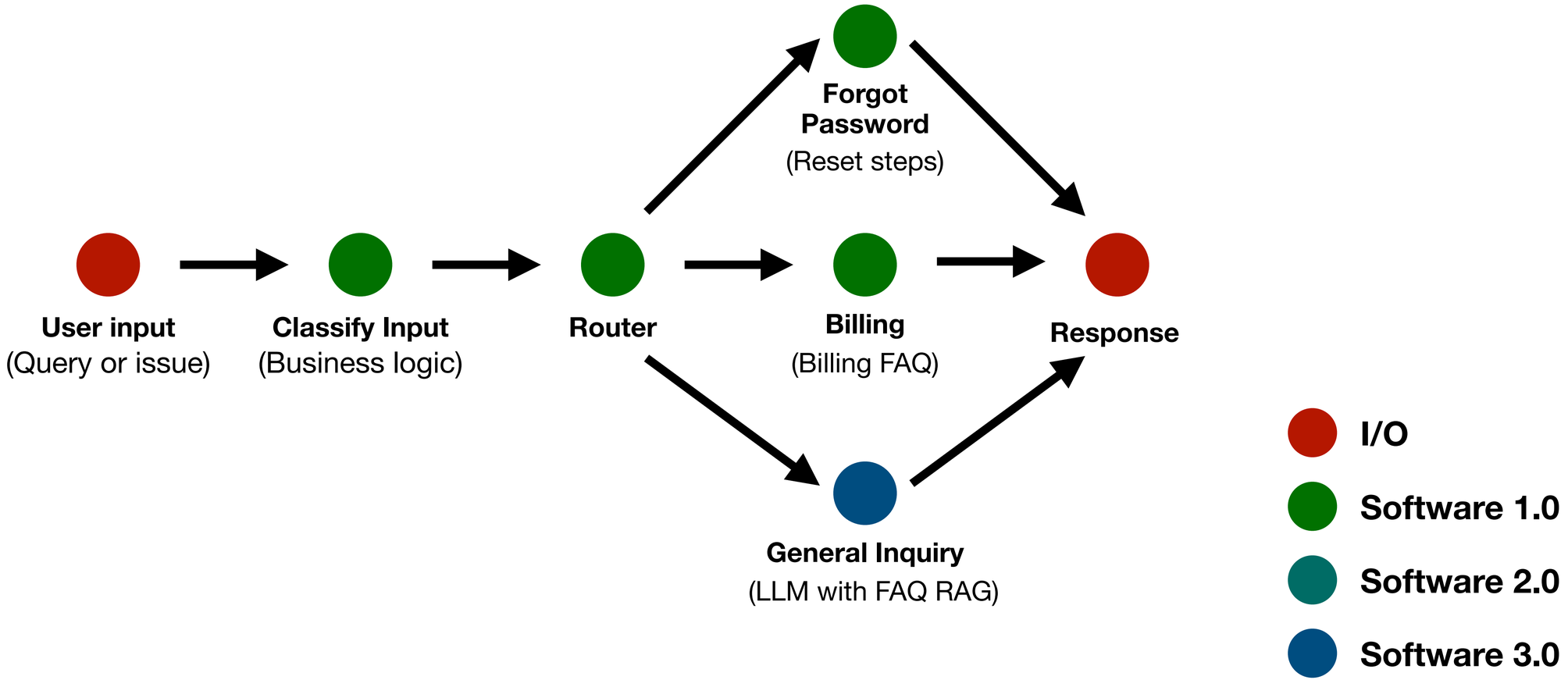


[12, 13]

AI Agents

Systems that can perform self-guided actions (my definition)

Vanilla Support Agent (low agency)

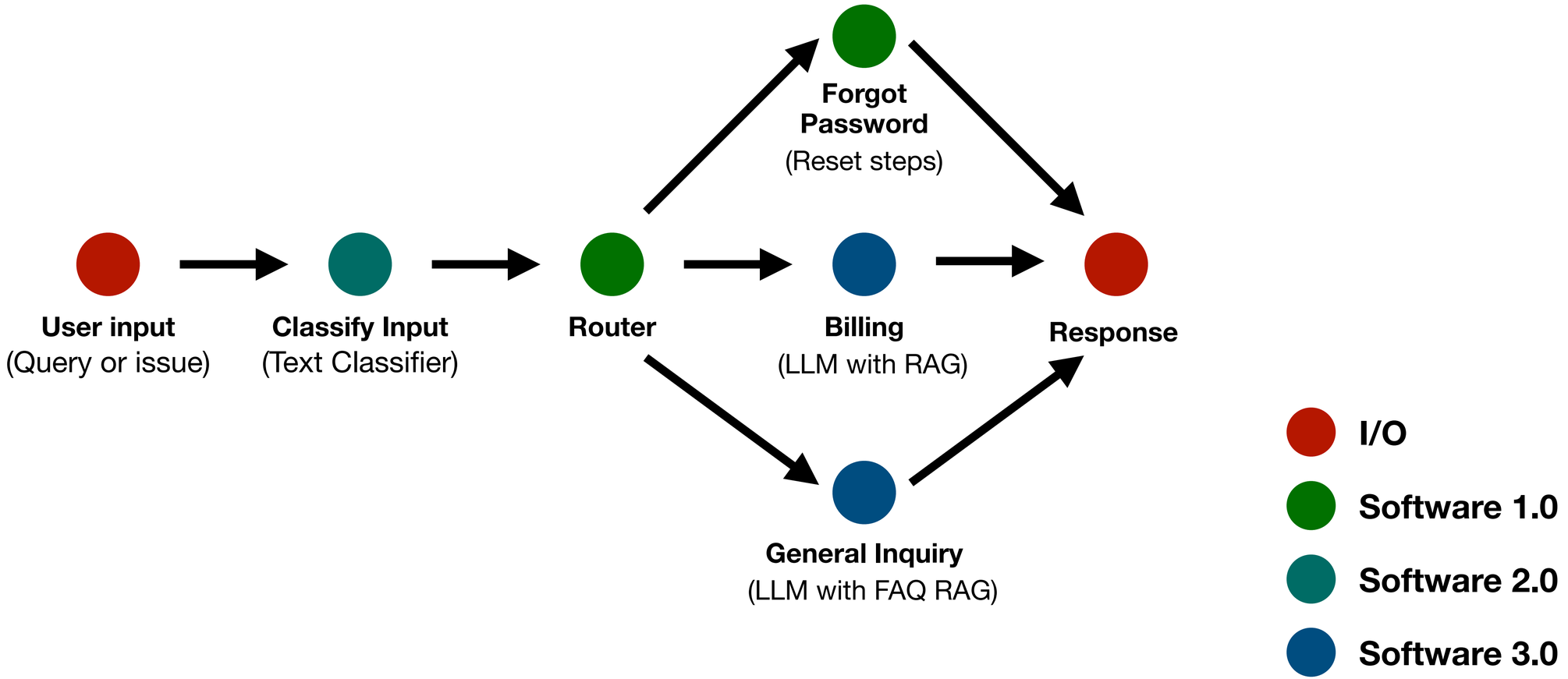


[12, 13]

AI Agents

Systems that can perform self-guided actions (my definition)

Vanilla Support Agent (moderate agency)

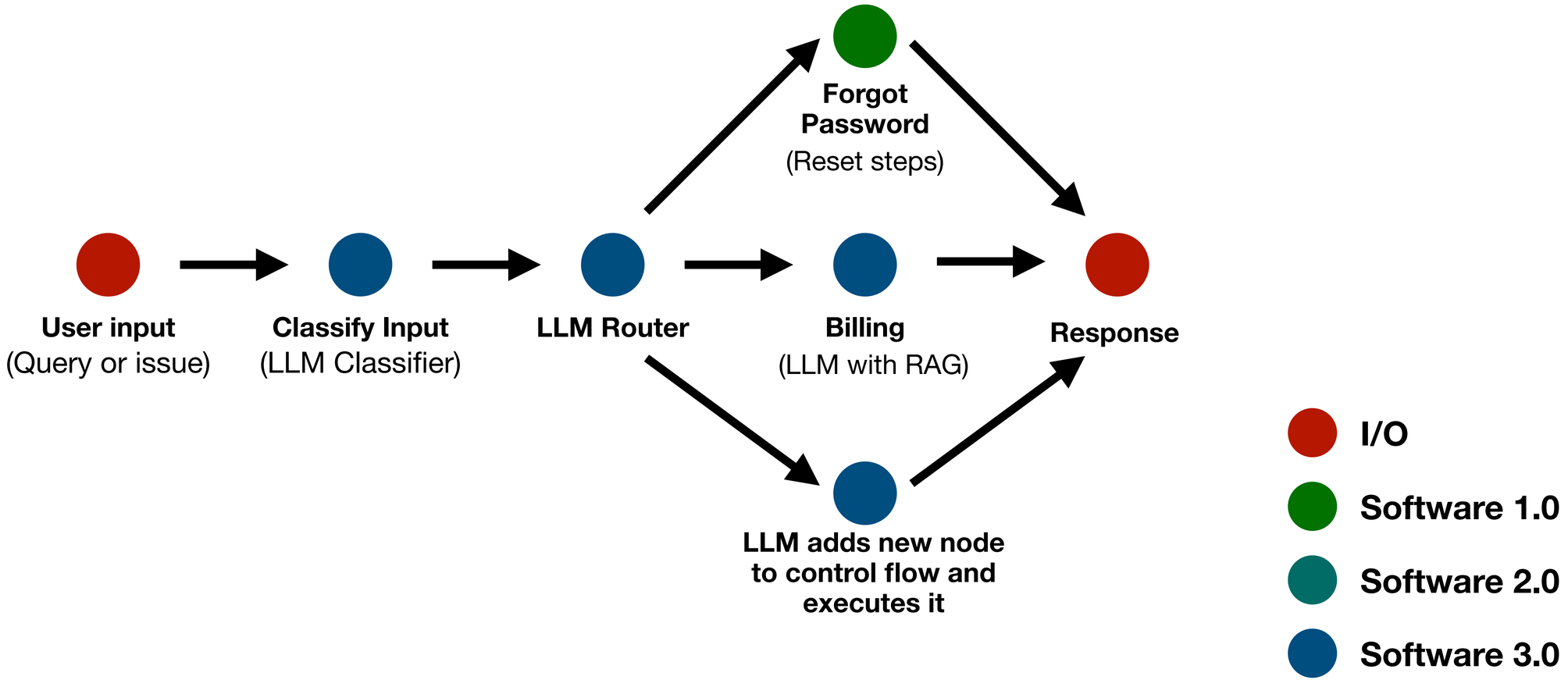


[12, 13]

AI Agents

Systems that can perform self-guided actions (my definition)

Vanilla Support Agent (high agency)



[12, 13]

What’s Next?

Thank you 🙏

Free Enrollment in Cohort #2

Post-course Survey ✍️

Share your feedback and get a **free 25 min call** with me!

References

[1] How to Build ML Solutions

[2] How to Manage Data Science Projects

[3] yt-search

[4] Full Stack Data Science Series

[5] How to Build Data Pipelines for ML Projects

[6] Automating Data Pipelines with Python & GitHub Actions

[7] How to Deploy ML Solutions with FastAPI, Docker, & AWS

[8] AI for Business: A (non-technical) introduction

[9] Principles by Ray Dalio

[10] What Nature Can Teach Us About Business

[11] Multimodal AI: LLMs that can see (and hear)

[12] What is an AI agent? By Harrison Chase

[13] State of AI Agents (LangChain)