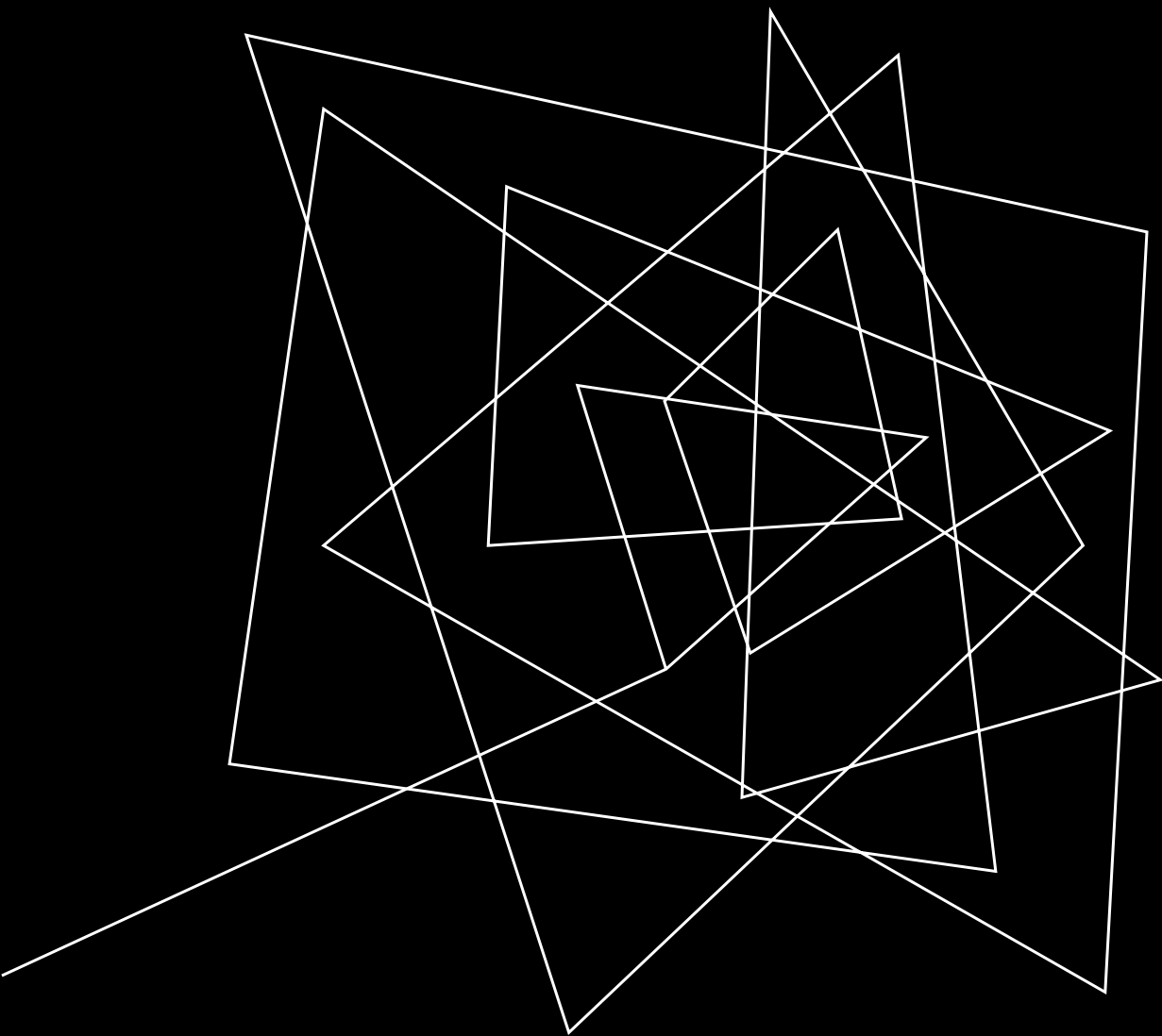




Recap

- Geschichte / Einordnung der KI
- NLP
- Embeddings

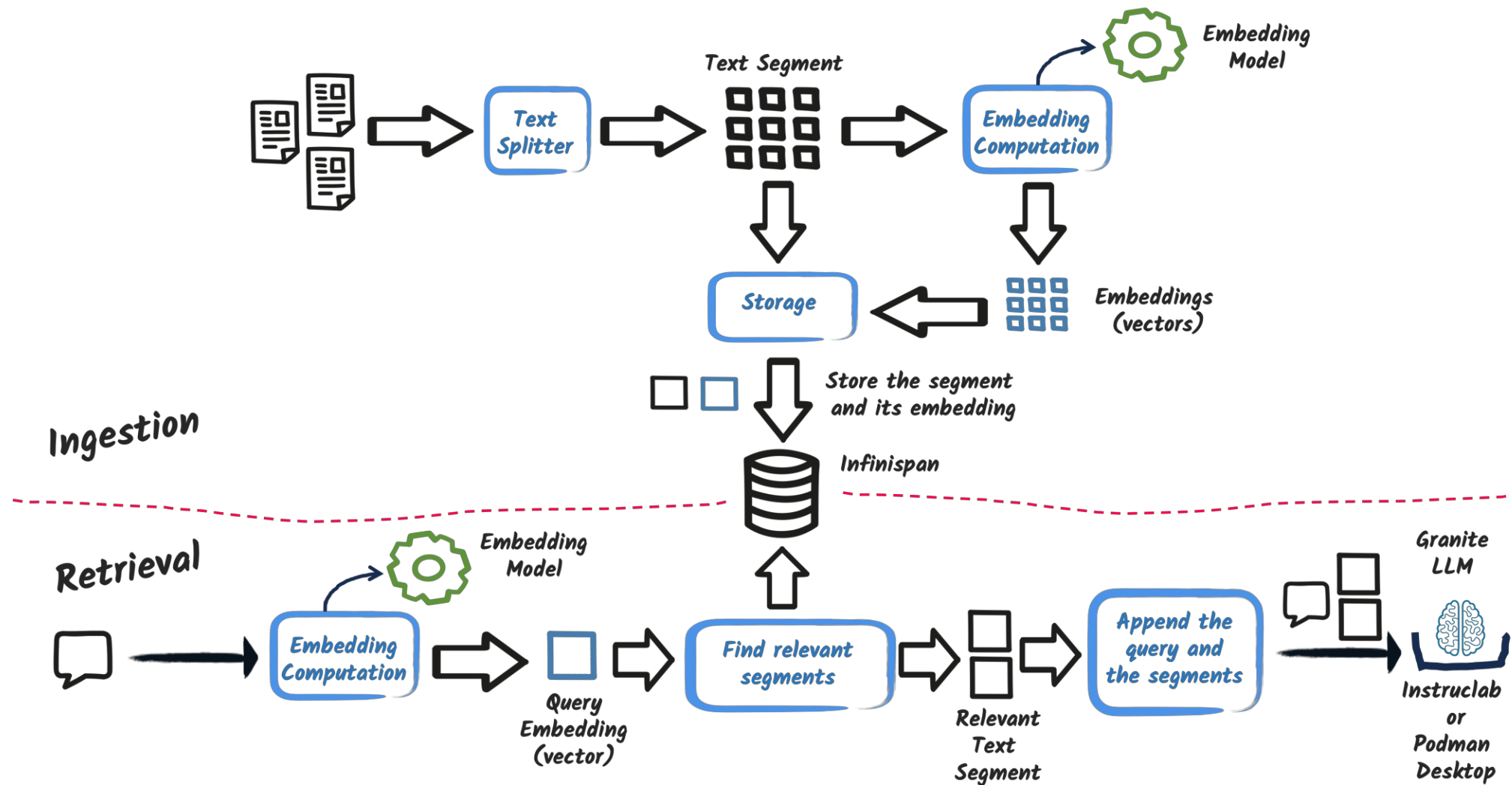


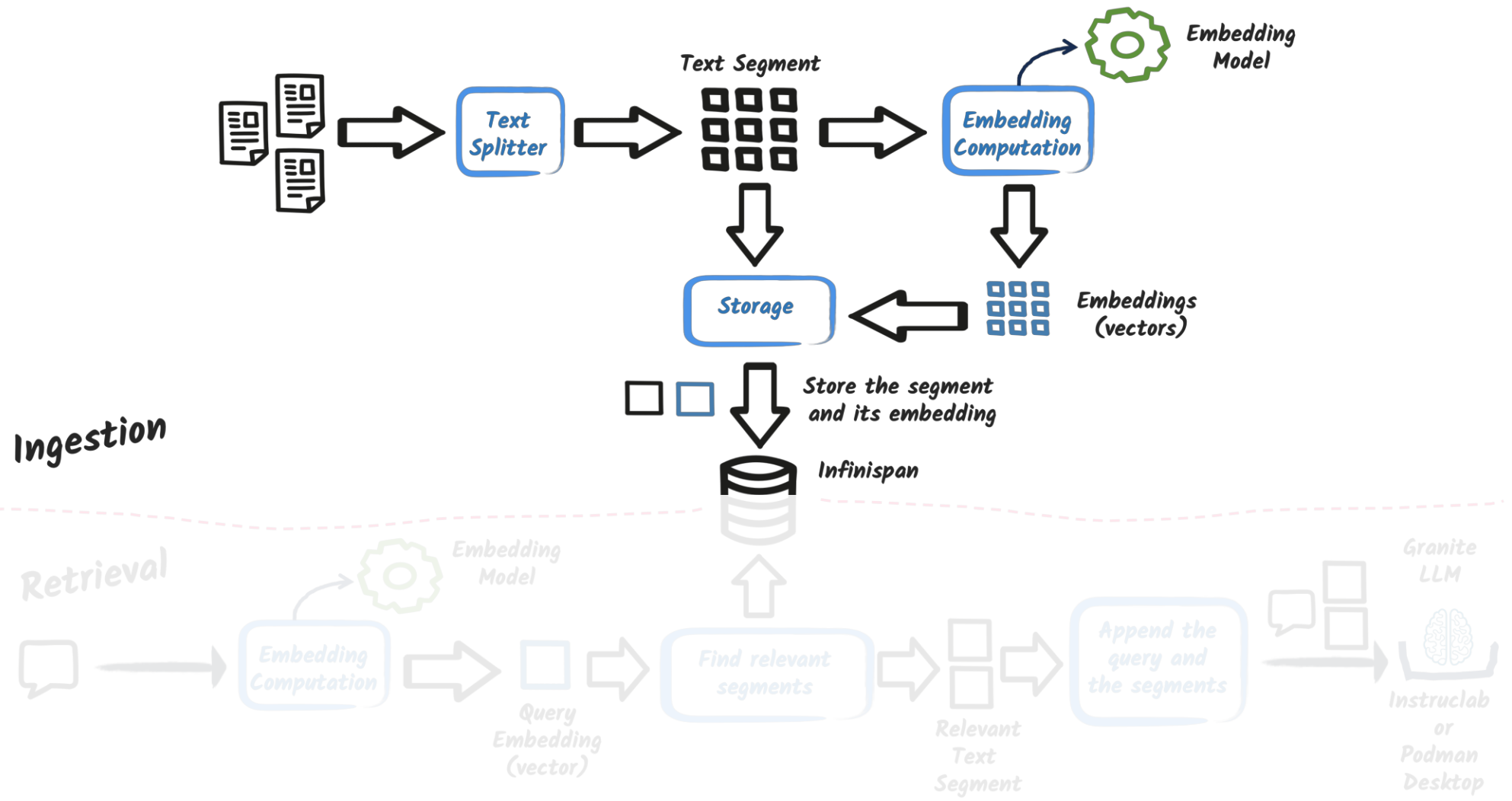
RAG

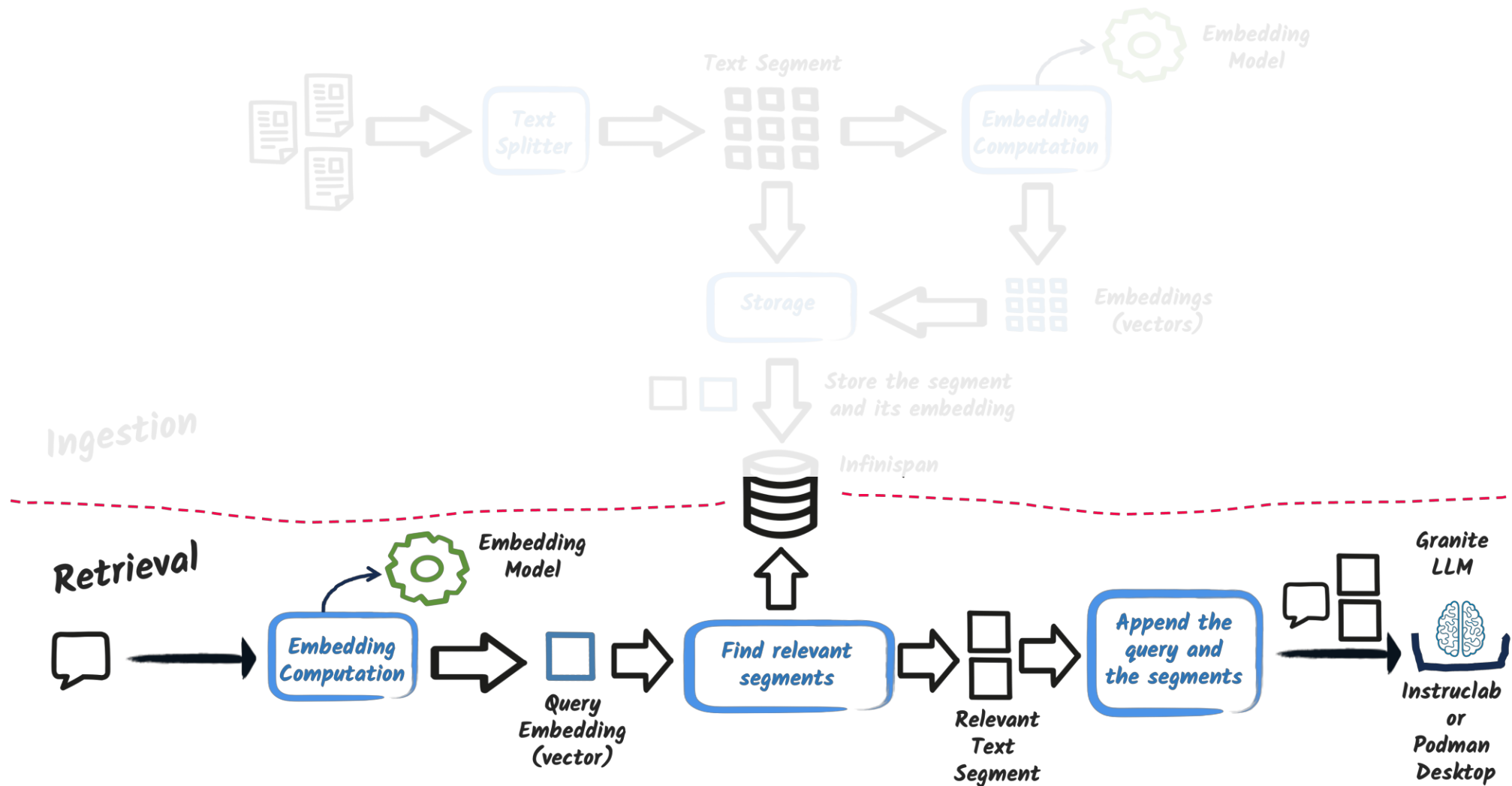
RAG

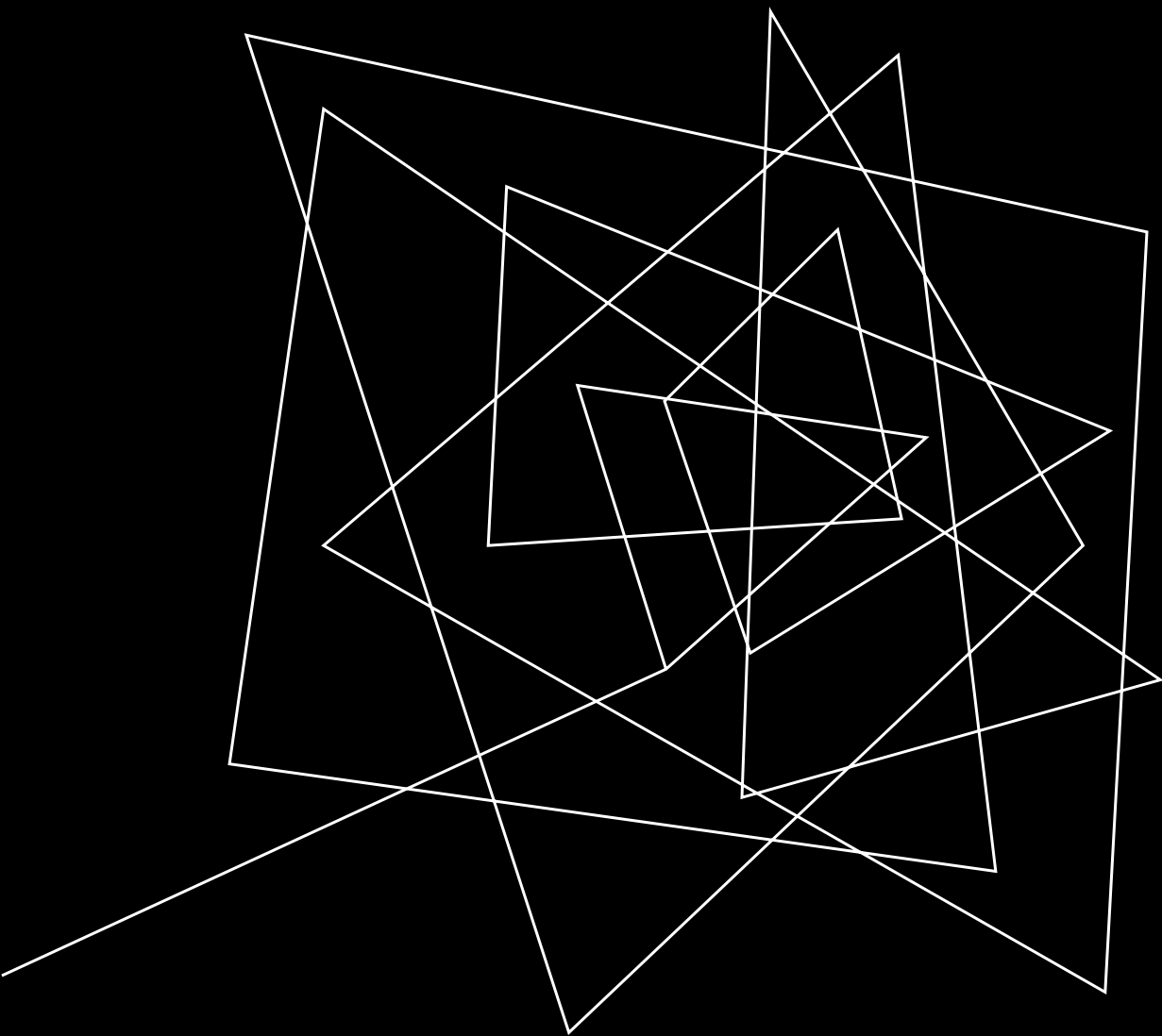
Retrieval Augmented Generation

- Neue Daten
- Private Daten
- Long Context
- Customization

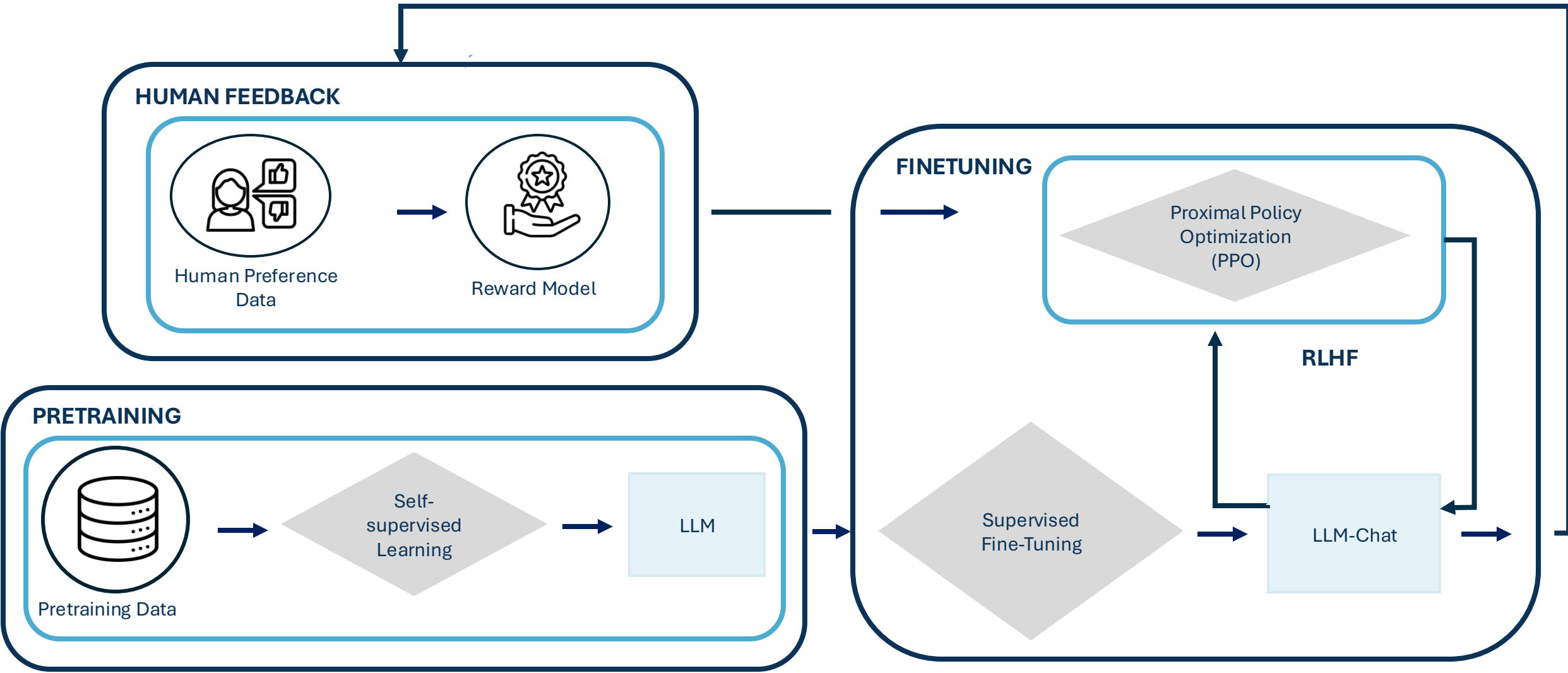


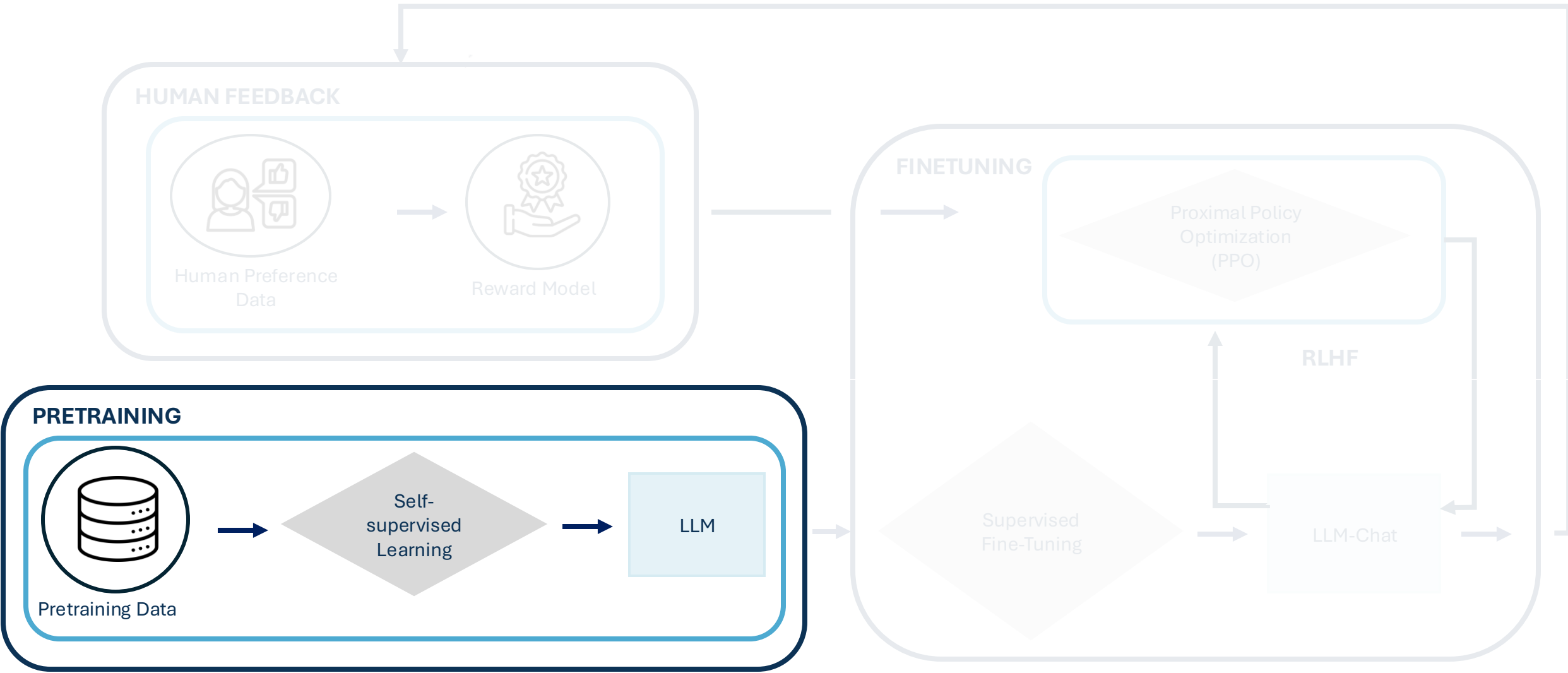


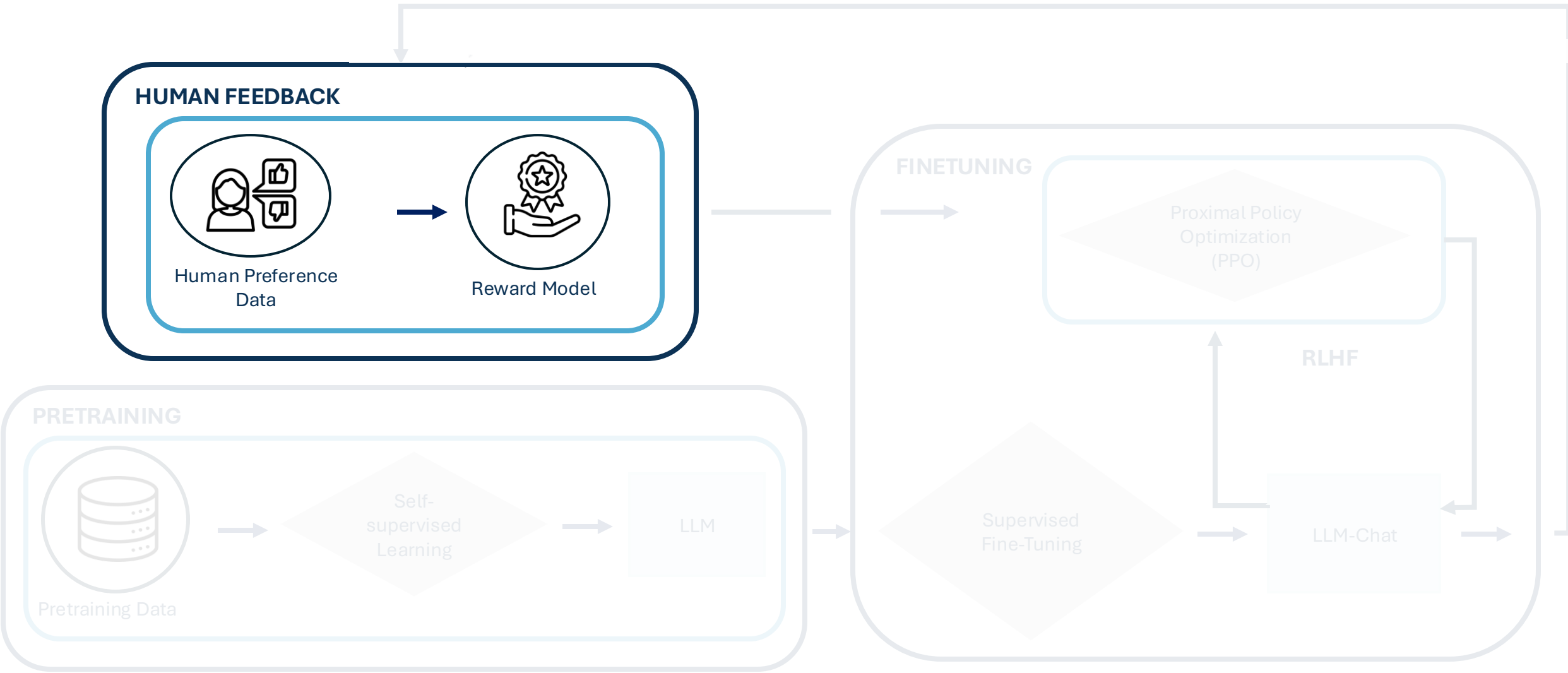


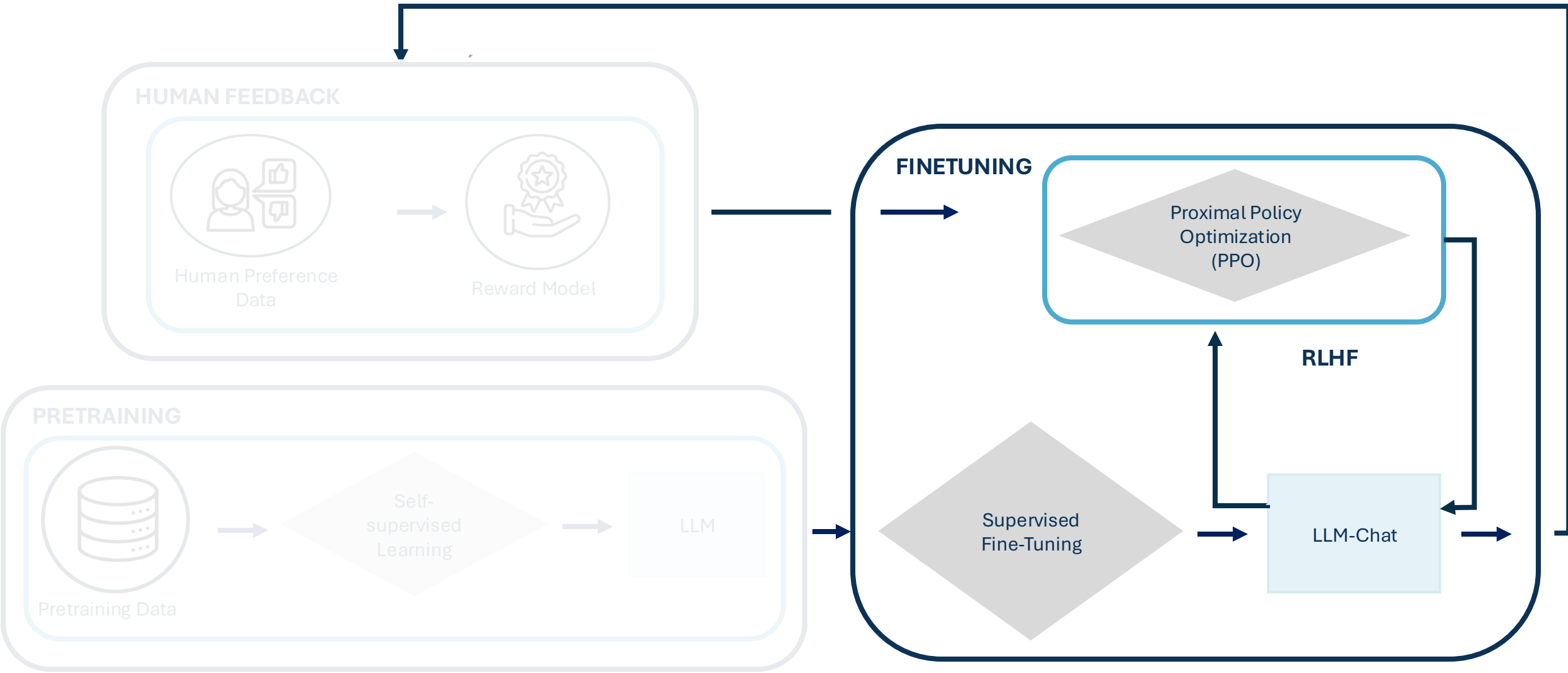


LLM





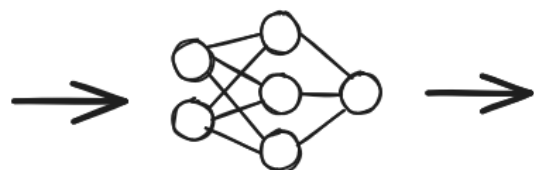
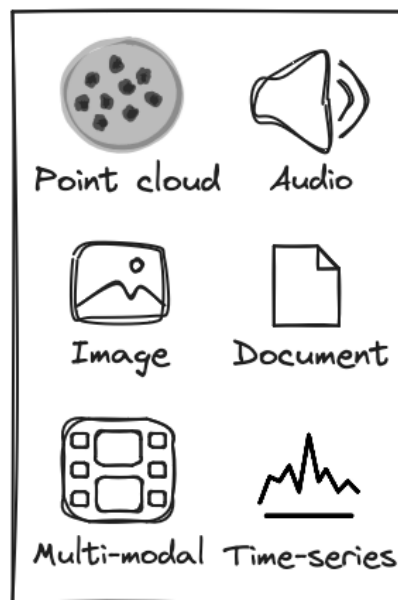






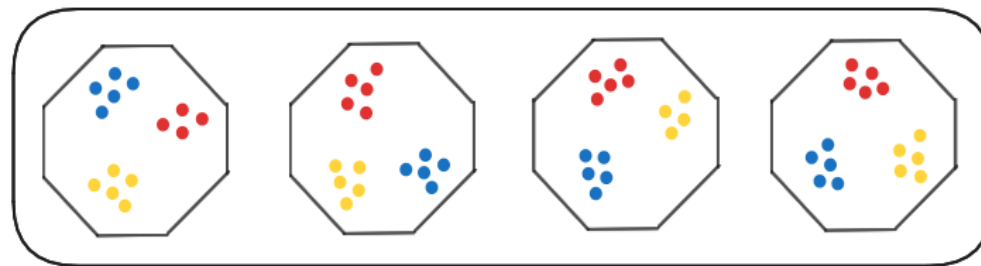
Vector DB

Data



Embedding
Model

Index



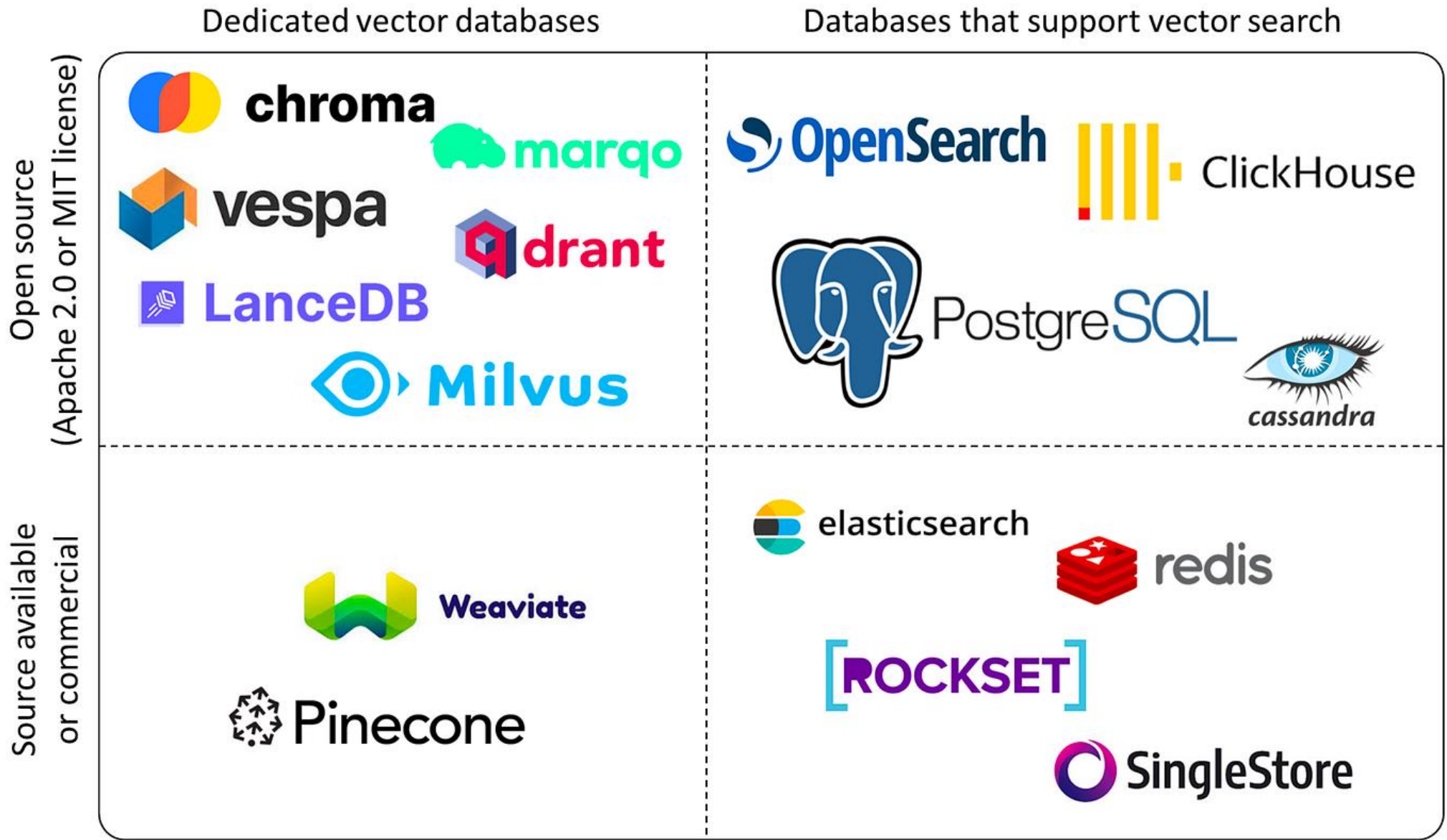
Embeddings



Search engine

Embeddings





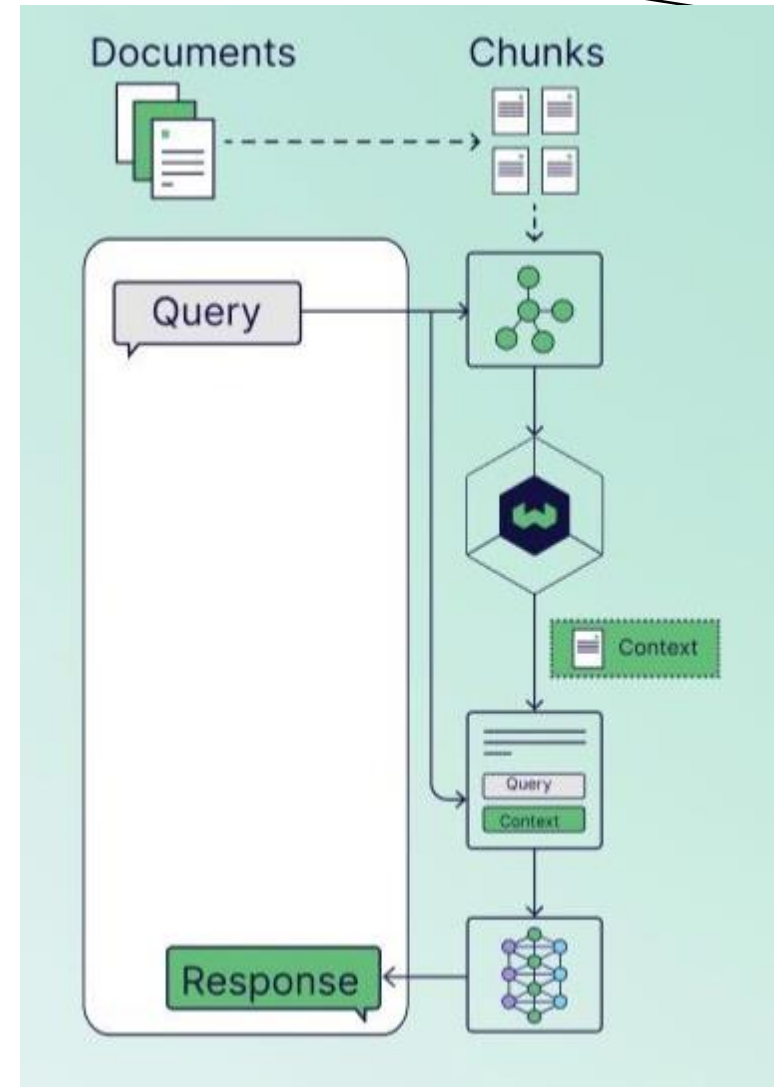
Abstract geometric lines in the top-left corner of the slide, consisting of several thin black lines forming overlapping, irregular polygons and triangles.

Praxisteil RAG

<https://github.com/enki-farm/MLCourse-I>

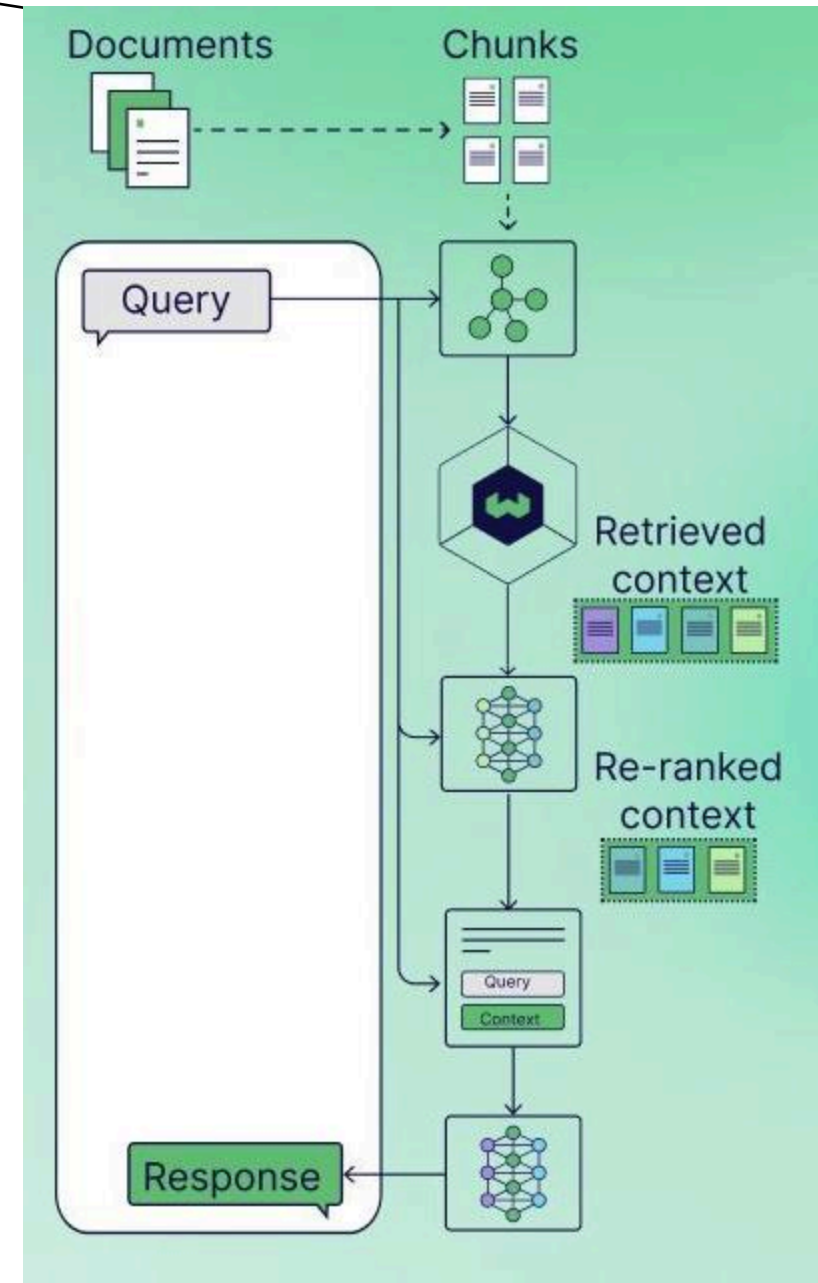
Naïve RAG

- Einfach zu bauen
- Manchmal zu simple



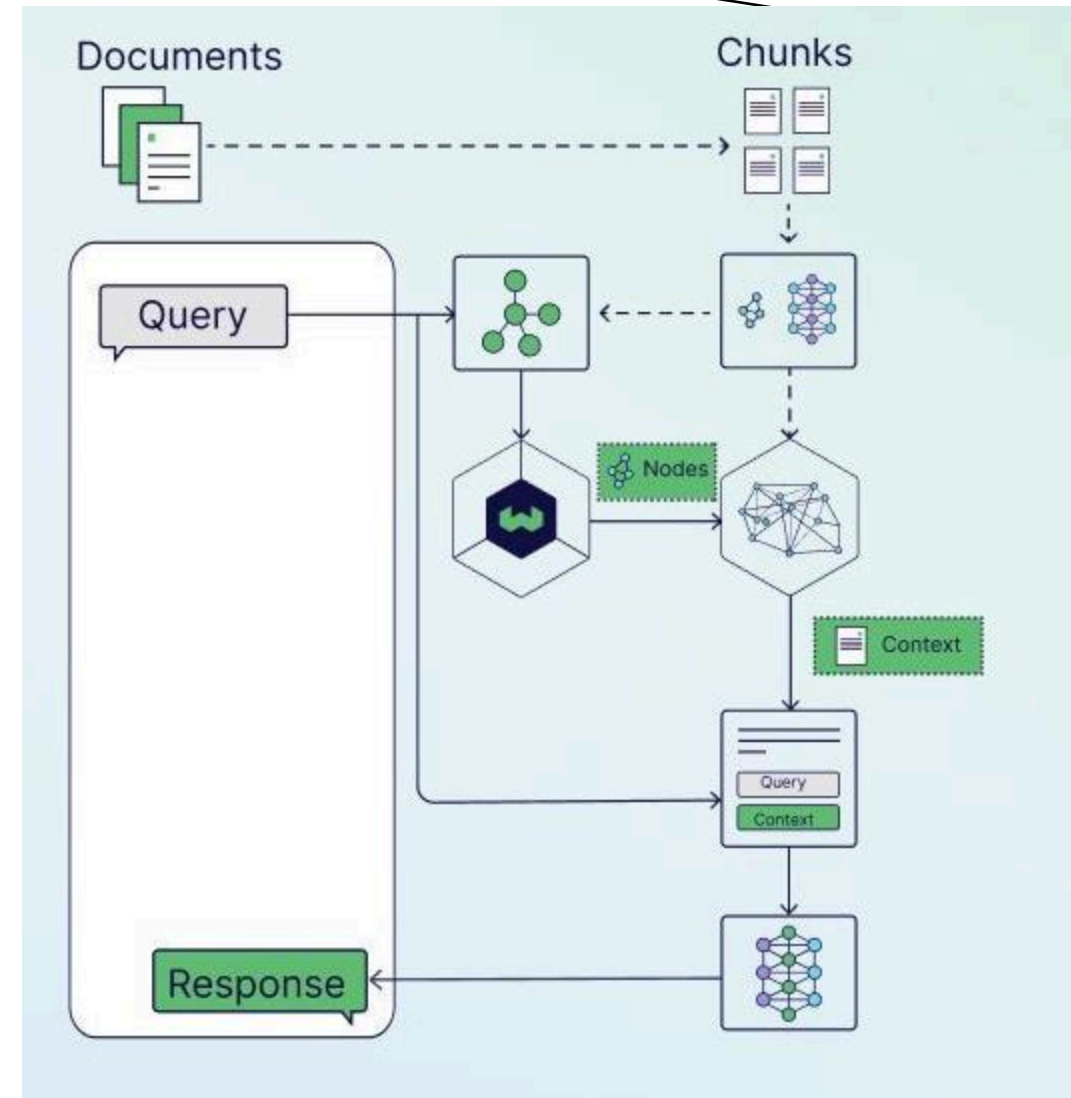
Retrieve-and-Rerank

- Eine Komponente mehr
 - Mehr Kontext abfragen und diesen Bewerten
- Genauere Ergebnisse
- Komplexer



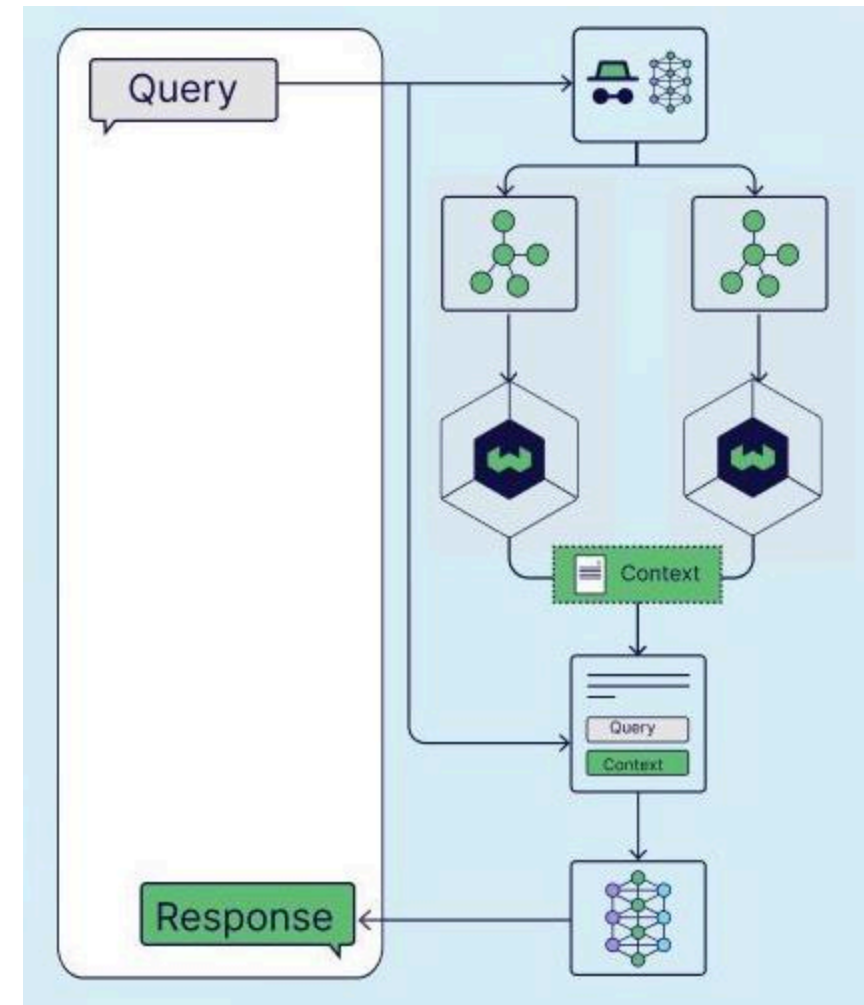
Graph RAG

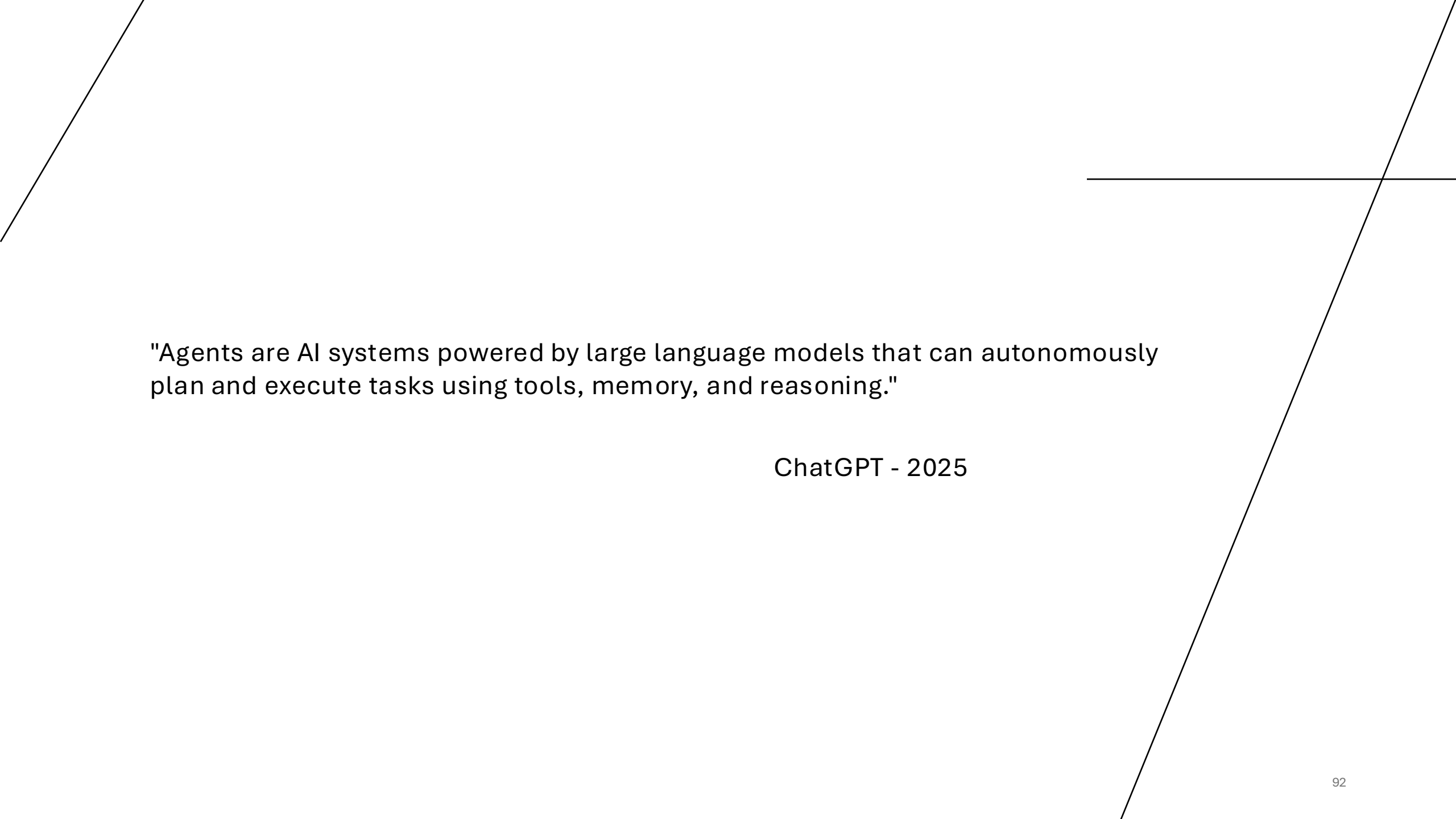
- GraphDB zusätzlich
- Nochmals genauere Ergebnisse
- Graph als schönes Nebenprodukt
- Einiges Komplexer
 - Entity Extraction nicht out-of-the-Box
- Teuer



Agentic RAG (Router)

- Agent (aka. LLM) wählt geeignete Source
- Adaptierbarer
 - Query pre-processing
 - Tools (aka. APIs)
- Nochmals komplexer
 - Schwierig zu debuggen
- Richtig Teuer





"Agents are AI systems powered by large language models that can autonomously plan and execute tasks using tools, memory, and reasoning."

ChatGPT - 2025



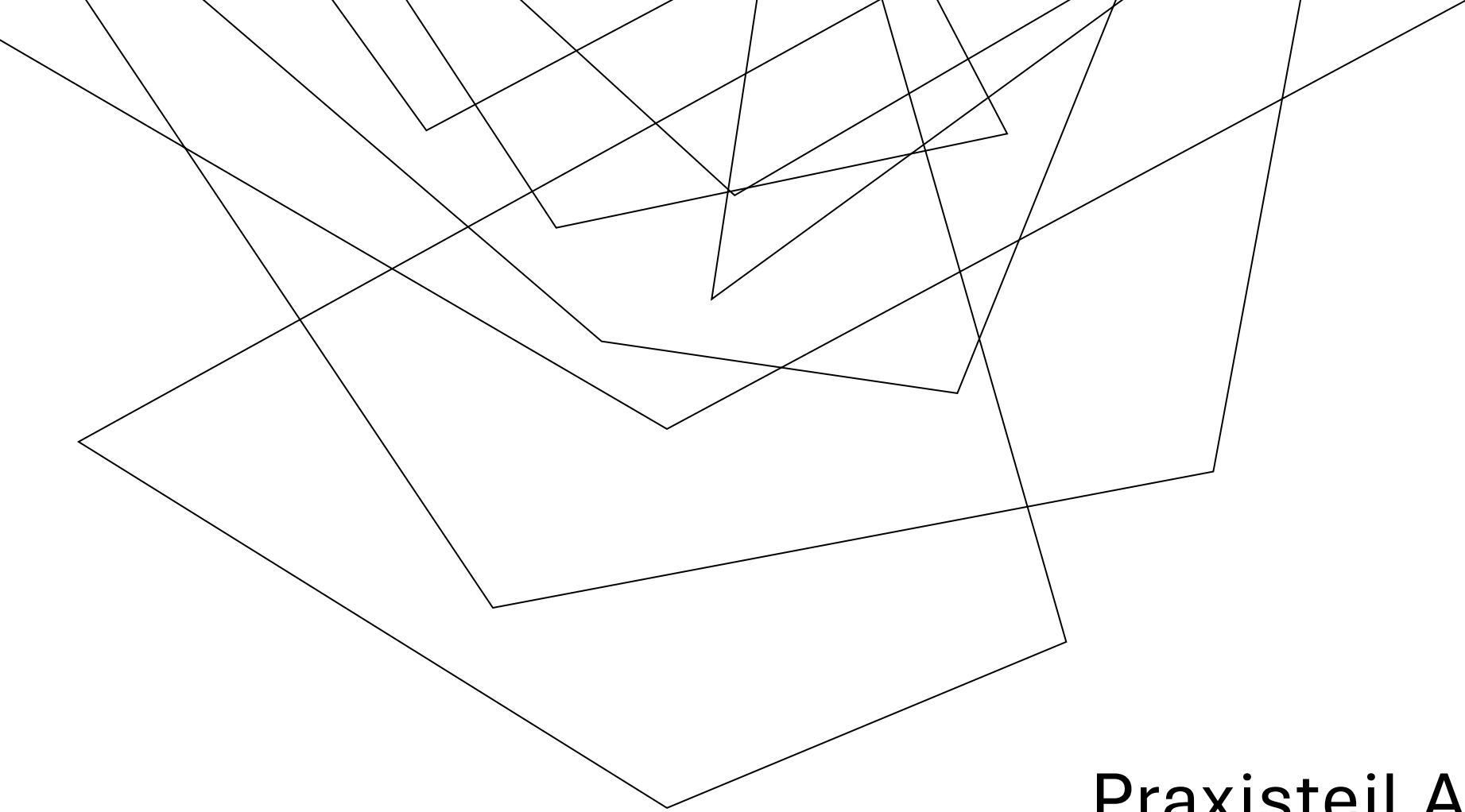
ganz nice

wege bisch mit autogen gange?

Verschiedeni Gründ. (Ned nach wichtigkeit sortiert)

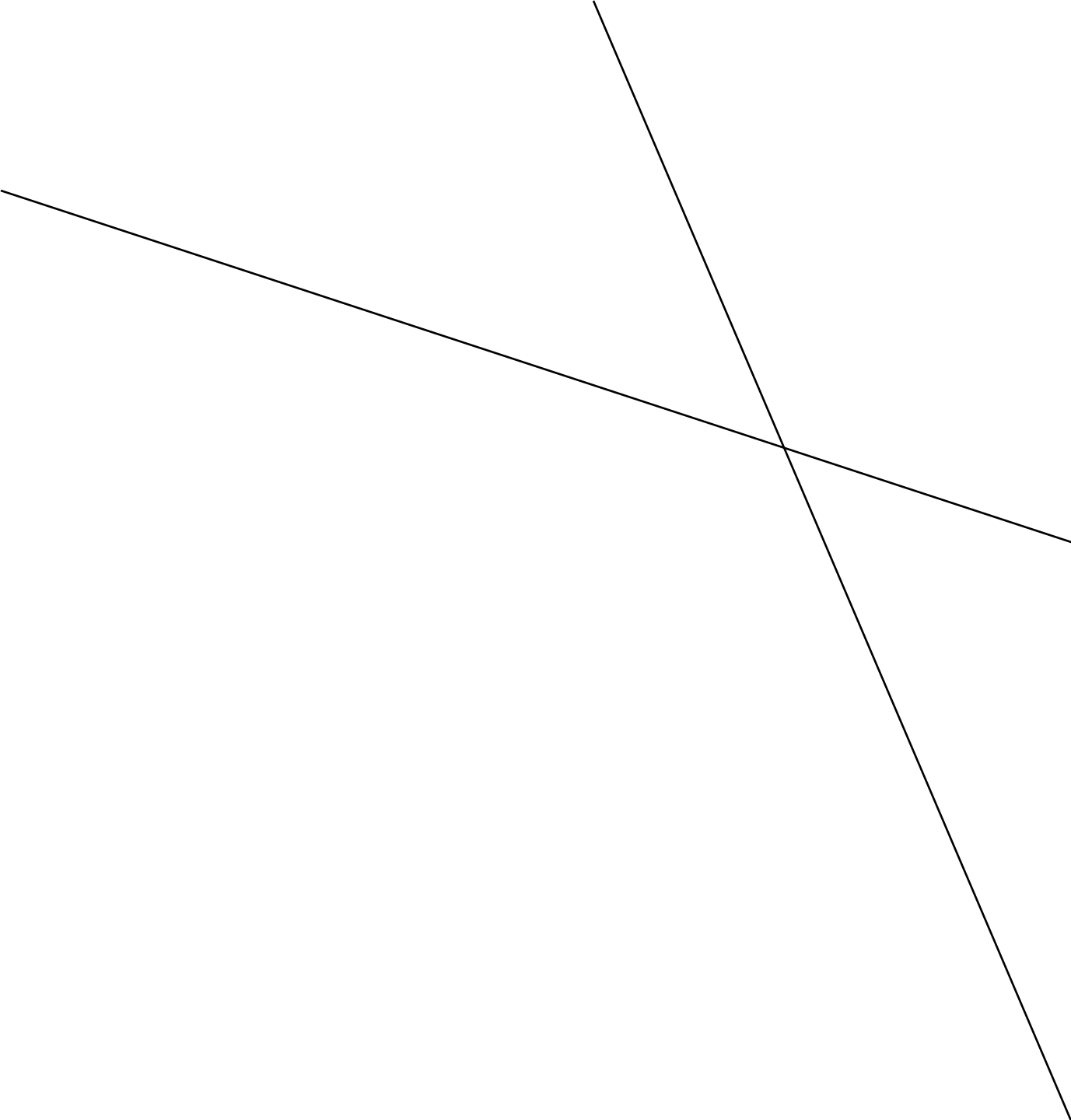
- Es isch vu MS (die droppeds wahrschindli ned grad und vill manpower dehinder)
- Ned supper supper neu
- 42k GithubStars
- Hed es UI (für mich ned supper wichtig aber nice zum werbig mache ^^)
- Guet Dokumentiert
- Guete level of abstraction
- Und de isch wüerkli wichtig für mich: Es isch s framework wo kagent brucht ^^

Ahh und sie unterstützed MCP



Praxisteil Agents

<https://github.com/enki-farm/MLCourse-I>



enki



Geschichte

- Ganz anderer Start
- Unser erstes Modell
- Warum wir nun hier sind

Enki gmbh

Dienstleistungen

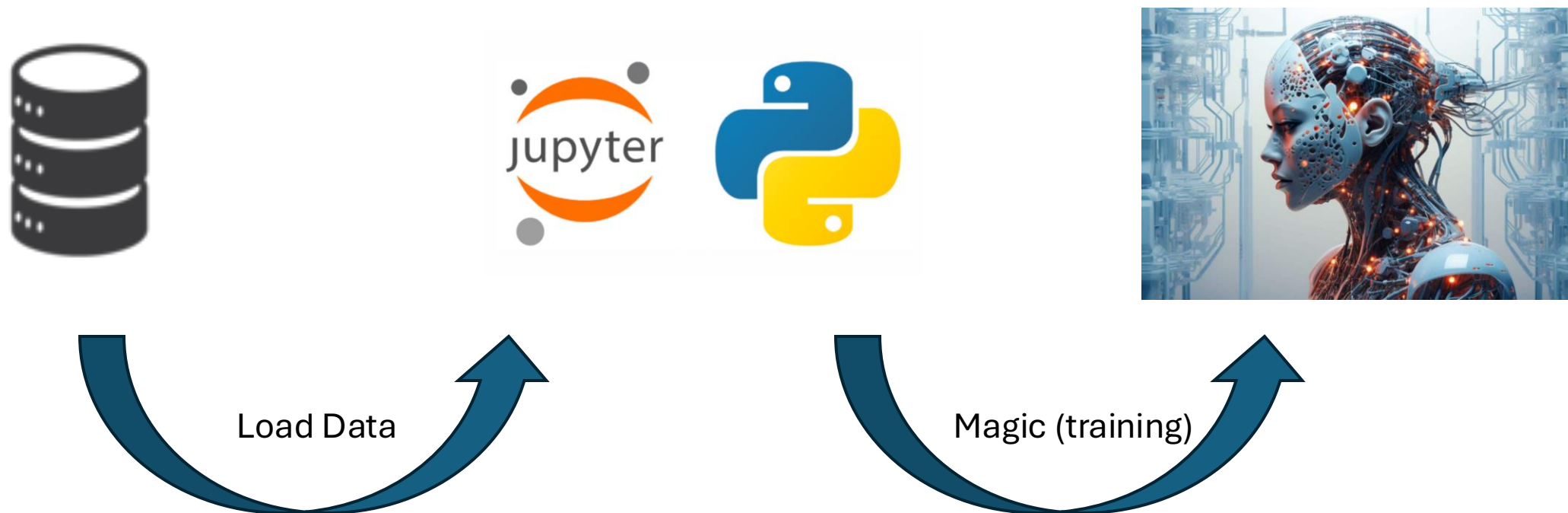
1. Schulung
2. Custom ML
3. Consulting

Idee

ML (Machine Learning) ist ein mächtiges Tool welches in die Hände aller gehört. Wir möchten KMU's helfen dieses auch für ihr Geschäft zu nutzen.

- Nah und persönlich
- Ihre Daten gehören Ihnen
- Am Puls der Zeit
- Kosteneffizient

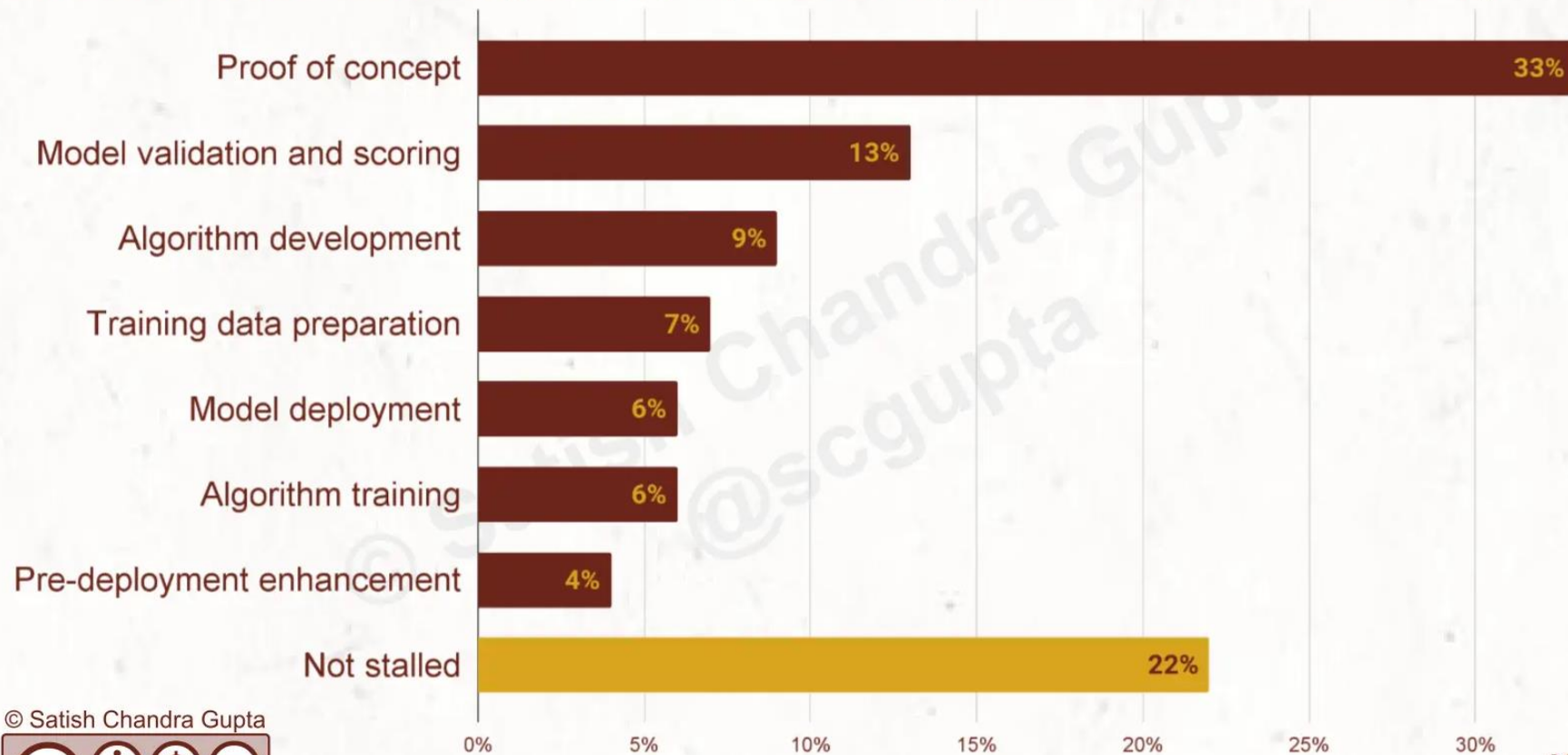
POC



78% of AI or ML Projects Stall at Some Stage Before Deployment



Source: Dimensional Research - Alegion Survey. <https://content.alegion.com/dimensional-researchs-survey>



© Satish Chandra Gupta

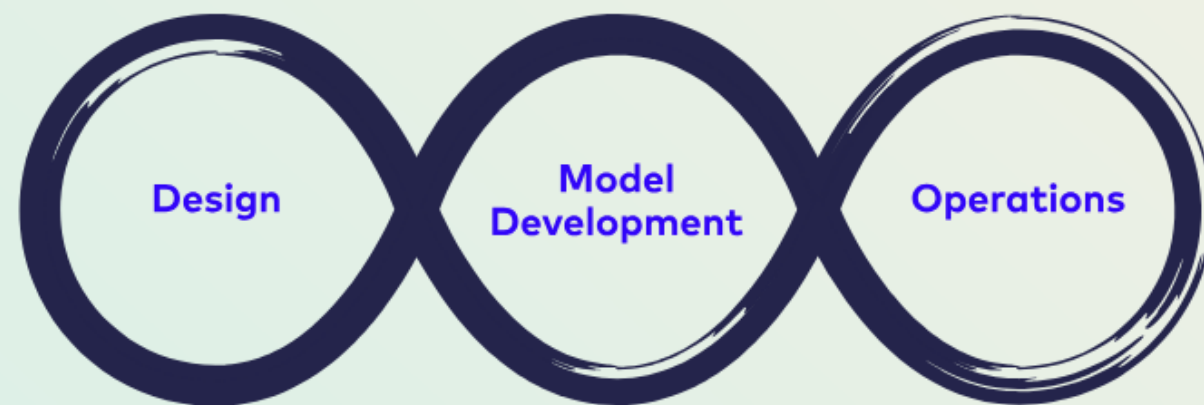


scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta

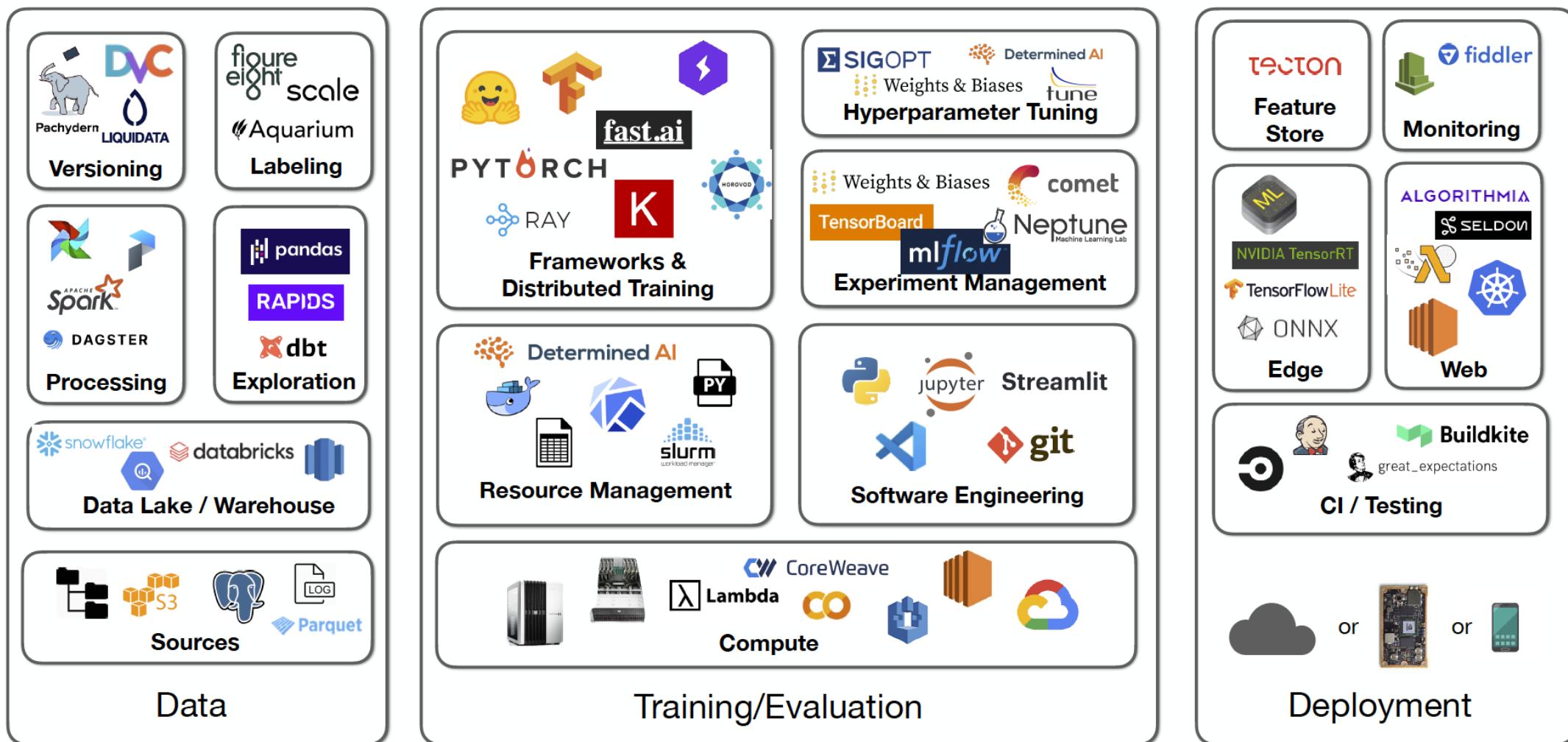
MLOps

Machine Learning Operations

With Machine Learning Model Operationalization Management (MLOps), we want to provide an end-to-end machine learning development process to design, build and manage reproducible, testable, and evolvable ML-powered software.



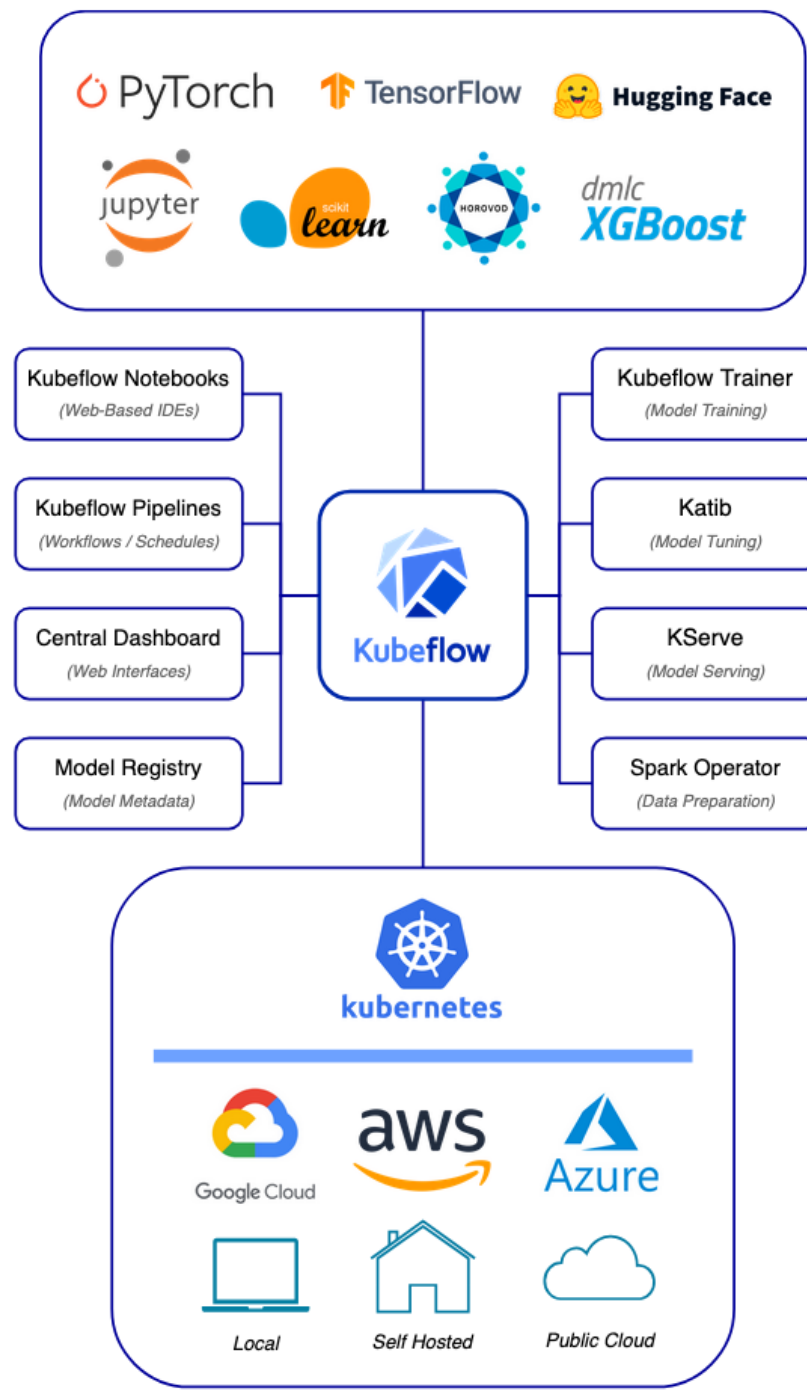
Production



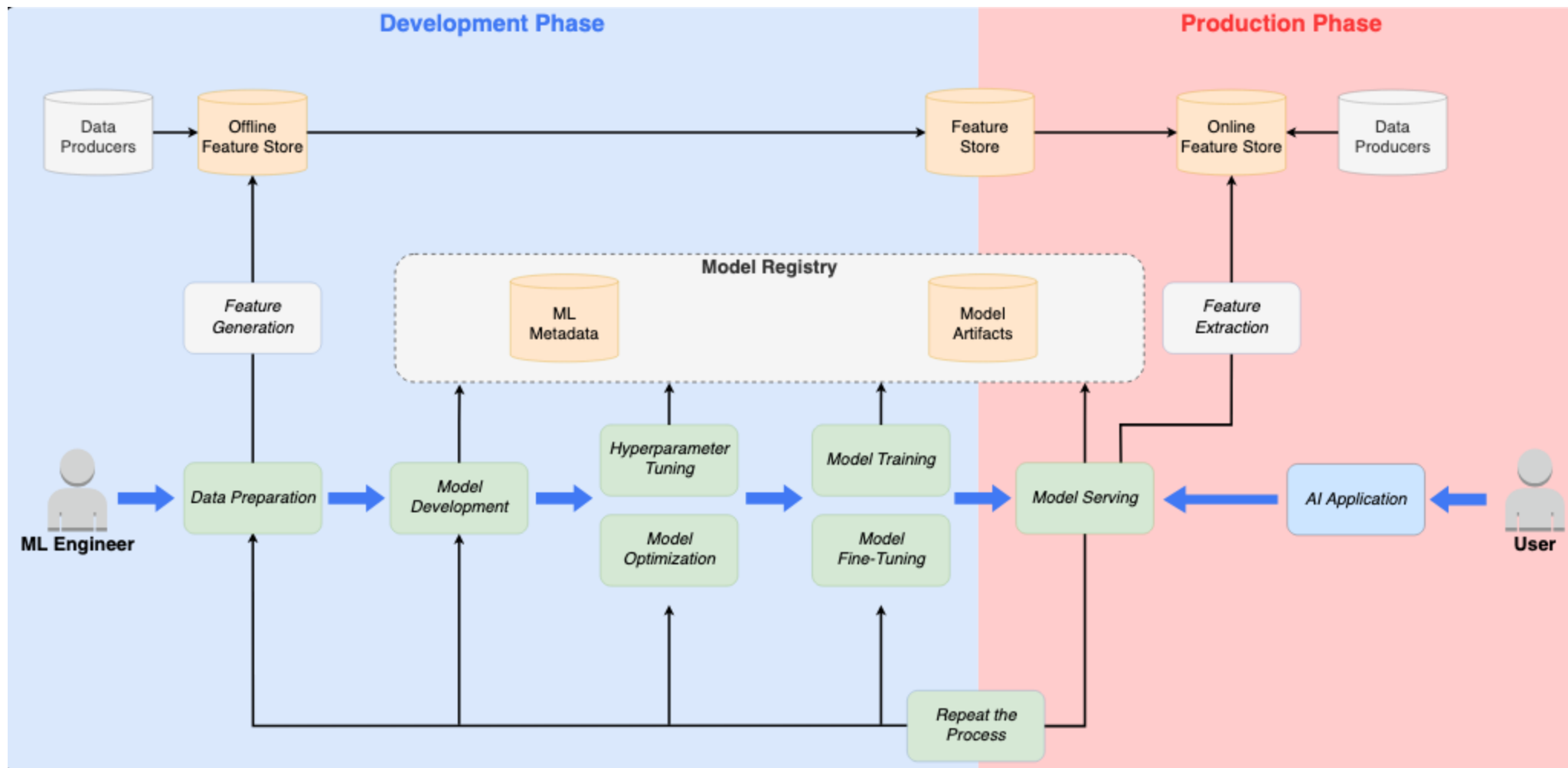
Kubeflow

Kubeflow is a community and ecosystem of open-source projects to address each stage in the machine learning (ML) lifecycle with support for best-in-class open-source tools and frameworks.

Kubeflow makes AI/ML on Kubernetes simple, portable, and scalable.

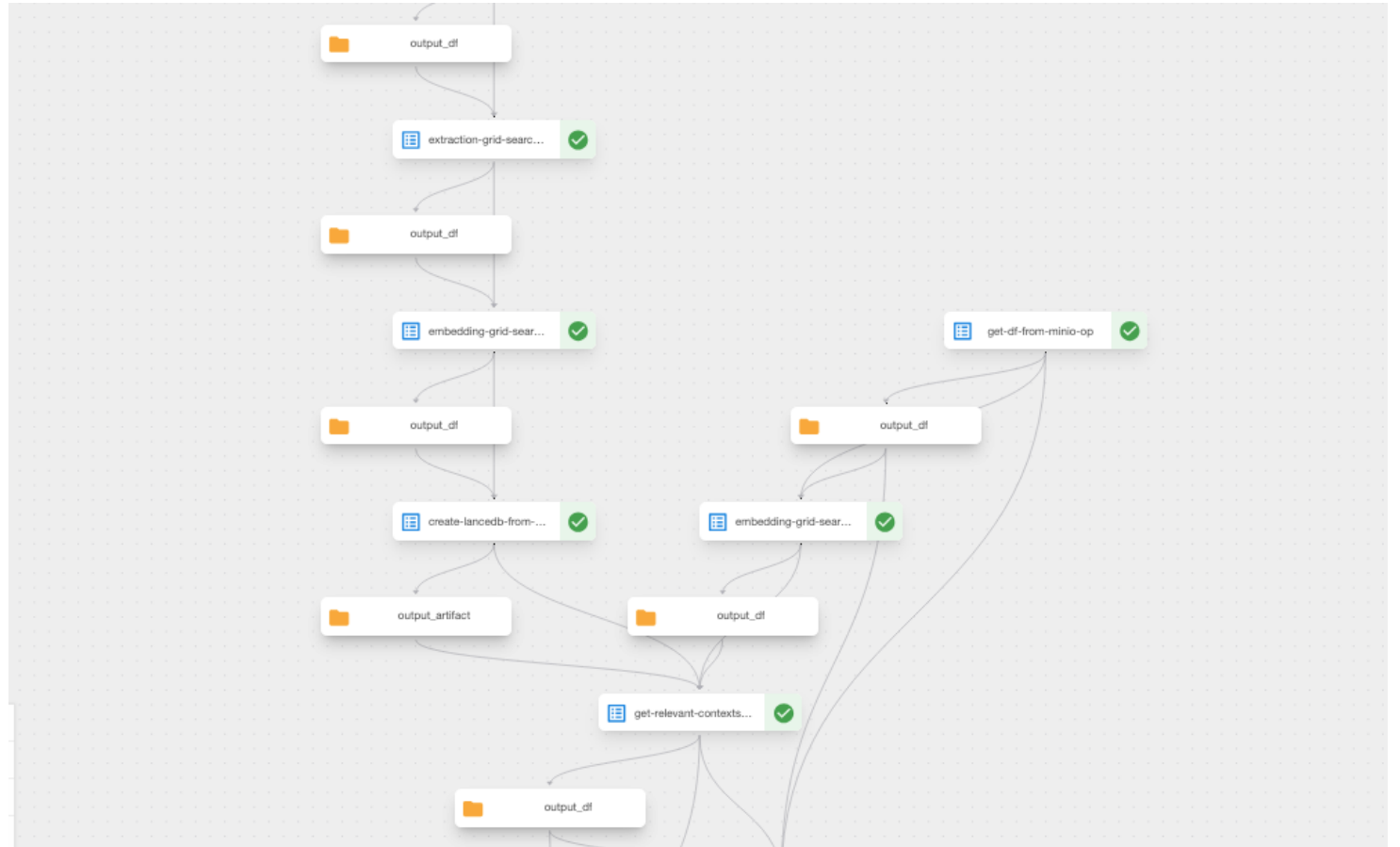


ML Lifecycle



Pipelines

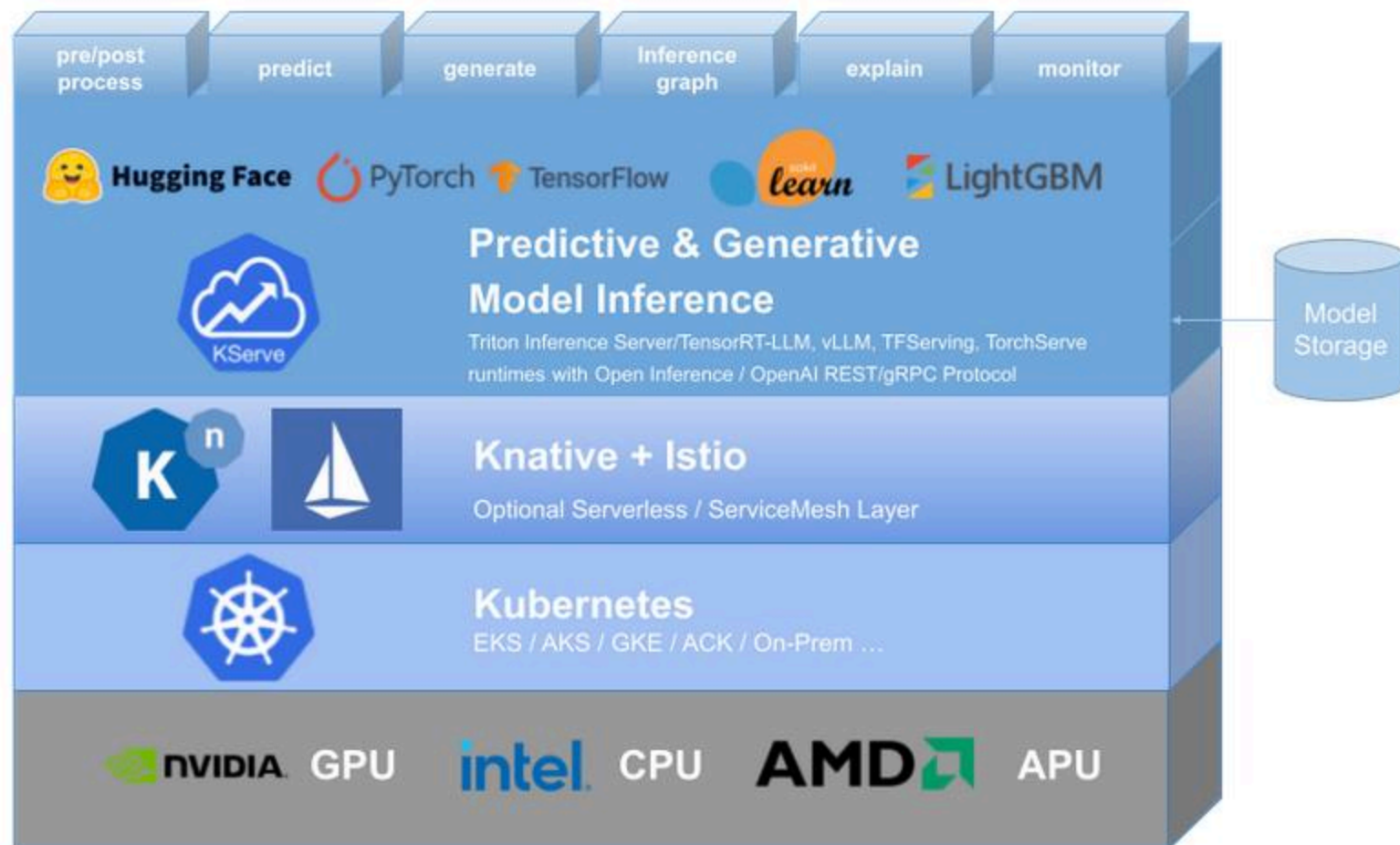
- Reproduzierbare Daten
- Wiederholbare Trainings
- Vergleichbarkeit schaffen
- Austauschbare Komponenten



KServe

KServe is an open-source project that enables serverless inferencing on Kubernetes.

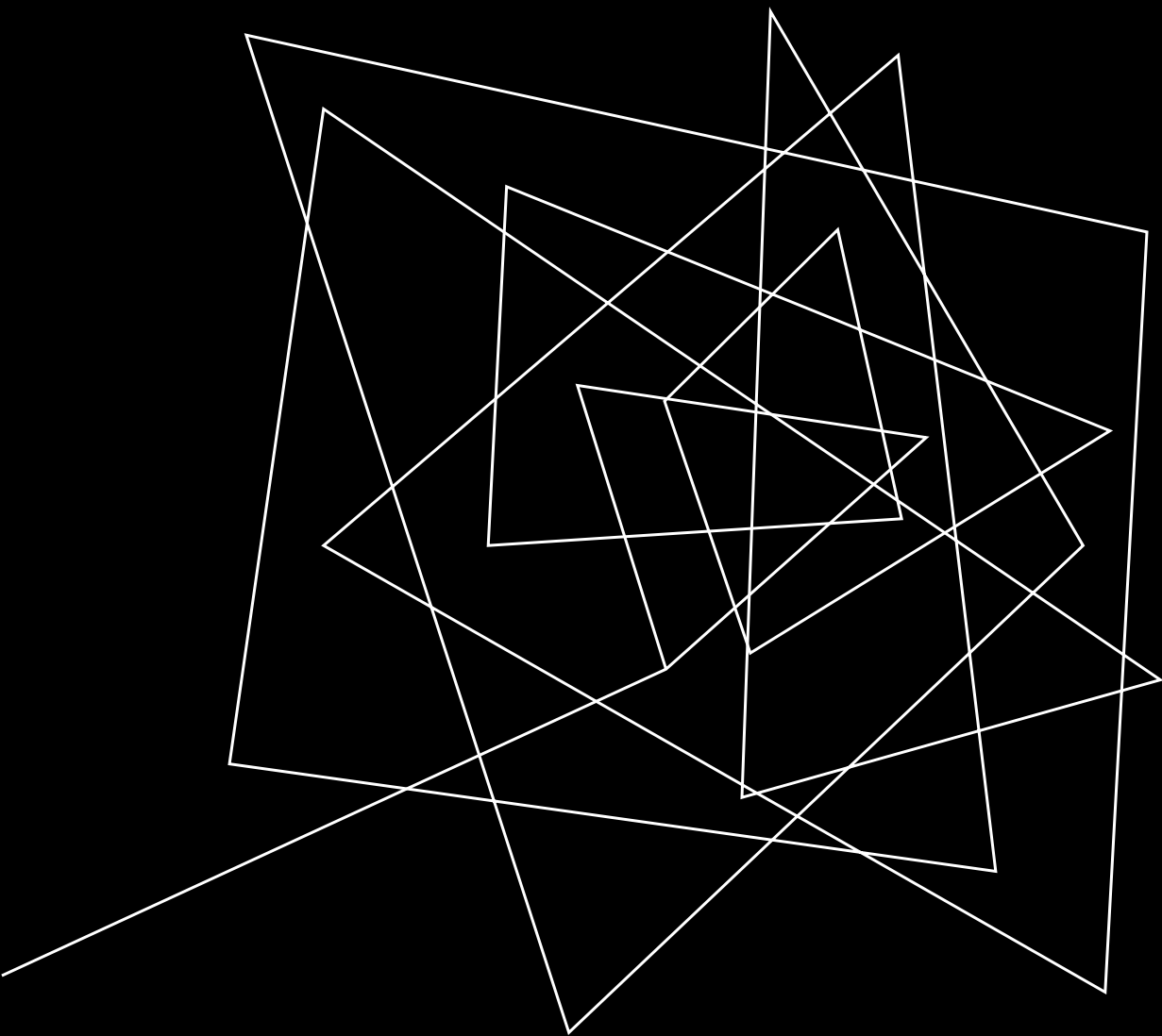
KServe provides performant, high abstraction interfaces for common machine learning (ML) frameworks like TensorFlow, XGBoost, scikit-learn, PyTorch, and ONNX to solve production model serving use cases.



Monitoring

- Das A und O von MLOps
- Grafana Stack
- Daten getriebene Entscheidungen





Diskussion



VIELEN DANK

Marco Crisafulli (marco@enki.swiss)

Daniel Dutli (dani@enki.swiss)

Enki GmbH

<https://enki.swiss>

Scan Me

