AI SYSTEMS INTRO

From LLM Productivity to Tailored **Enterprise Growth**

Stratos Idreos

FinLab @McKinsey is hiring

If interested reach out to:
Gabriel_Morgan_Asaftei@mckinsey.com

Goals

1) Current state-of-the-art in AI systems
2) Definitions
3) Design space
4) What is next
5) Skills you may need/want

A small set of “simple” tricks/concepts

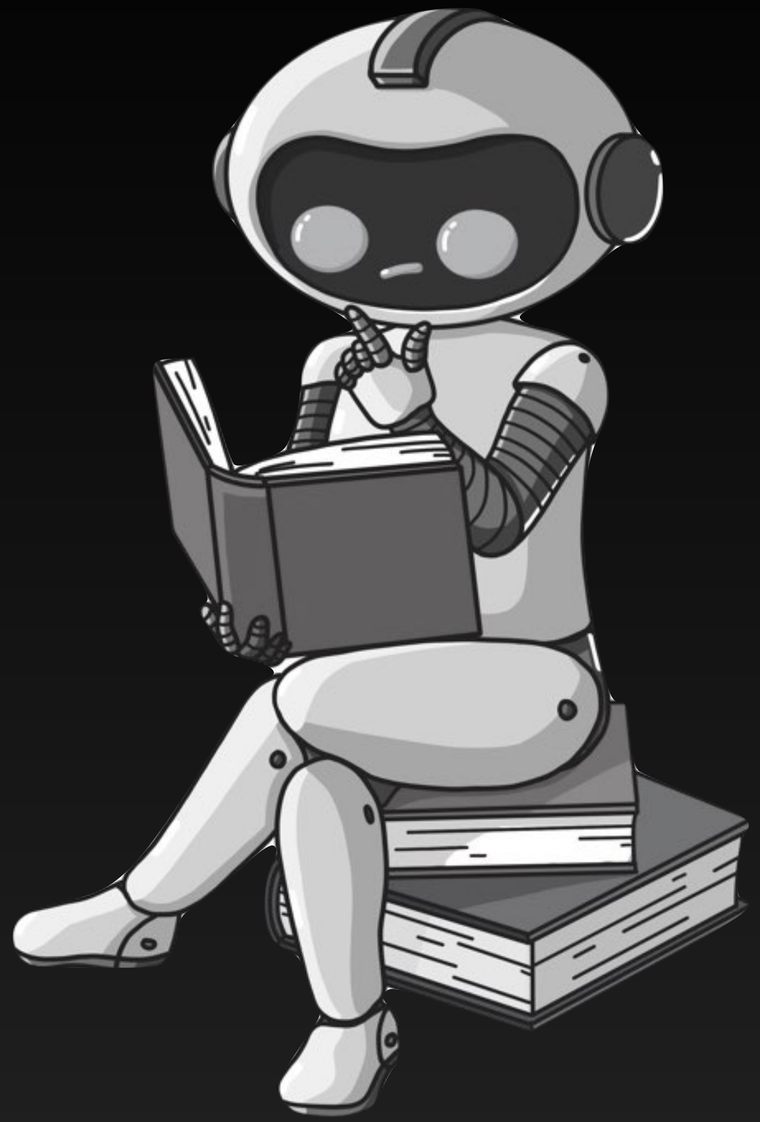
DATA LLMs AGENTS ML


Industry Evolution Path
VALUE & COMPLEXITY
1
Data Foundation
STORE & MANAGE
Data-first approach
DATA
2
Analytics
INSIGHT & REPORTING
Business intelligence
LLMs
3
Productivity Automation
BASIC AGENTS
Workflow efficiency
AGENTS
4
Core Automation
BUSINESS USE CASES
Mission-critical functions
ML
5
Custom ML Models
DOMAIN-SPECIFIC AI
Tailored intelligence
?
6
Systems of Intelligence
FULL INTEGRATION
Complex orchestration
MATURITY JOURNEY →

AI Systems

The growing ecosystem of specialized tools, libraries, and platforms, each addressing a distinct piece of the AI lifecycle:

1) training & fine-tuning models,
2) managing & querying data,
3) building & orchestrating agents,
4) …



OpenAI Anthropic Google Cohere

HuggingFace Ollama Replicate

LlamaIndex LangChain Haystack

TRL PEFT DSPy Axolotl

MODEL TRAINING & EXPERIMENTATION

Training Frameworks

Build, train, and iterate on neural networks and ML models at any scale

PyTorch TensorFlow JAX Keras

Classical ML & Data Science

Statistical models, tabular ML, preprocessing, and exploratory analysis

scikit-learn XGBoost Pandas NumPy

Experiment Tracking

Log runs, compare metrics, manage hyperparameters, reproduce results

MLflow W&B Neptune CometML

AutoML & HPO

Automated model selection, hyperparameter search, and neural architecture search

Optuna Ray Tune AutoGluon

DATA & STORAGE

Relational Databases

Structured data storage, transactions, SQL queries for tabular business data

PostgreSQL MySQL Snowflake

Vector Databases

Similarity search over embeddings for retrieval & RAG pipelines

Pinecone Weaviate Milvus

Data Lakes & Warehouses

Scalable storage for massive datasets across structured & unstructured formats

Databricks BigQuery Redshift

Data Pipelines & ETL

Ingestion, transformation, scheduling, and orchestration of data workflows

Airflow dbt Spark Kafka

Feature Stores

Manage, serve, and reuse engineered features across models and teams

Feast Tecton Hopsworks

COMPUTE & HARDWARE

GPU & Accelerators

Specialized hardware powering all modern training and inference workloads

NVIDIA H100 AMD MI300 Google TPU Trainium

Cloud AI Platforms

Managed infrastructure for training, serving, and scaling AI workloads

AWS GCP Azure Lambda Labs

Cluster & Job Management

Schedule, scale, and manage distributed compute jobs across clusters

Kubernetes Slurm Ray SkyPilot DeepSpeed

OpenAI Anthropic Google Cohere

HuggingFace Ollama Replicate

LlamaIndex LangChain Haystack

TRL PEFT DSPy Axolotl

MODEL TRAINING & EXPERIMENTATION

Training Frameworks
Build, train, and iterate on neural networks and ML models at any scale

PyTorch TensorFlow JAX Keras

Classical ML & Data Science
Statistical models, tabular ML, preprocessing, and exploratory analysis

scikit-learn XGBoost Pandas NumPy

Experiment Tracking
Log runs, compare metrics, manage hyperparameters, reproduce results

MLflow W&B Neptune CometML

AutoML & HPO
Automated model selection, hyperparameter search, and neural architecture search

Optuna Ray Tune AutoGluon

DATA & STORAGE

Relational Databases
Structured data storage, transactions, SQL queries for tabular business data

PostgreSQL MySQL Snowflake

Vector Databases
Similarity search over embeddings for retrieval & RAG pipelines

Pinecone Weaviate Milvus

Data Lakes & Warehouses
Scalable storage for massive datasets across structured & unstructured formats

Databricks BigQuery Redshift

Data Pipelines & ETL
Ingestion, transformation, scheduling, and orchestration of data workflows

Airflow dbt Spark Kafka

Feature Stores
Manage, serve, and reuse engineered features across models and teams

Feast Tecton Hopsworks

A LOT OF INNOVATION POSSIBLE ON FOUNDATIONS

COMPUTE & HARDWARE

GPU & Accelerators
Specialized hardware powering all modern training and inference workloads

NVIDIA H100 AMD MI300 Google TPU Trainium

Cloud AI Platforms
Managed infrastructure for training, serving, and scaling AI workloads

AWS GCP Azure Lambda Labs

Cluster & Job Management
Schedule, scale, and manage distributed compute jobs across clusters

Kubernetes Slurm Ray SkyPilot DeepSpeed

OpenAI Anthropic Google Cohere

HuggingFace Ollama Replicate

LlamaIndex LangChain Haystack

TRL PEFT DSPy Axolotl

MODEL TRAINING & EXPERIMENTATION

Training Frameworks
Build, train, and iterate on neural networks and ML models at any scale

PyTorch TensorFlow JAX Keras

Classical ML & Data Science
Statistical models, tabular ML, preprocessing, and exploratory analysis

scikit-learn XGBoost Pandas NumPy

Experiment Tracking
Log runs, compare metrics, manage hyperparameters, reproduce results

MLflow W&B Neptune CometML

AutoML & HPO
Automated model selection, hyperparameter search, and neural architecture search

Optuna Ray Tune AutoGluon

THE FOUNDATIONS FOR DATA

DATA & STORAGE

Relational Databases
Structured data storage, transactions, SQL queries for tabular business data

PostgreSQL MySQL Snowflake

Vector Databases
Similarity search over embeddings for retrieval & RAG pipelines

Pinecone Weaviate Milvus

Data Lakes & Warehouses
Scalable storage for massive datasets across structured & unstructured formats

Databricks BigQuery Redshift

Data Pipelines & ETL
Ingestion, transformation, scheduling, and orchestration of data workflows

Airflow dbt Spark Kafka

Feature Stores
Manage, serve, and reuse engineered features across models and teams

Feast Tecton Hopsworks

A LOT OF INNOVATION POSSIBLE ON FOUNDATIONS

COMPUTE & HARDWARE

GPU & Accelerators
Specialized hardware powering all modern training and inference workloads

NVIDIA H100 AMD MI300 Google TPU Trainium

Cloud AI Platforms
Managed infrastructure for training, serving, and scaling AI workloads

AWS GCP Azure Lambda Labs

Cluster & Job Management
Schedule, scale, and manage distributed compute jobs across clusters

Kubernetes Slurm Ray SkyPilot DeepSpeed

OpenAI Anthropic Google Cohere

HuggingFace Ollama Replicate

LlamaIndex LangChain Haystack

TRL PEFT DSPy Axolotl

MODEL TRAINING & EXPERIMENTATION

Training Frameworks
Build, train, and iterate on neural networks and ML models at any scale

PyTorch TensorFlow JAX Keras

Classical ML & Data Science
Statistical models, tabular ML, preprocessing, and exploratory analysis

scikit-learn XGBoost Pandas NumPy

Experiment Tracking
Log runs, compare metrics, manage hyperparameters, reproduce results

MLflow W&B Neptune CometML

AutoML & HPO
Automated model selection, hyperparameter search, and neural architecture search

Optuna Ray Tune AutoGluon

THE FOUNDATIONS FOR DATA

DATA & STORAGE

Relational Databases
Structured data storage, transactions, SQL queries for tabular business data

PostgreSQL MySQL Snowflake

Vector Databases
Similarity search over embeddings for retrieval & RAG pipelines

Pinecone Weaviate Milvus

Data Lakes & Warehouses
Scalable storage for massive datasets across structured & unstructured formats

Databricks BigQuery Redshift

Data Pipelines & ETL
Ingestion, transformation, scheduling, and orchestration of data workflows

Airflow dbt Spark Kafka

Feature Stores
Manage, serve, and reuse engineered features across models and teams

Feast Tecton Hopsworks

A LOT OF INNOVATION POSSIBLE ON FOUNDATIONS

COMPUTE & HARDWARE

GPU & Accelerators
Specialized hardware powering all modern training and inference workloads

NVIDIA H100 AMD MI300 Google TPU Trainium

Cloud AI Platforms
Managed infrastructure for training, serving, and scaling AI workloads

AWS GCP Azure Lambda Labs

Cluster & Job Management
Schedule, scale, and manage distributed compute jobs across clusters

Kubernetes Slurm Ray SkyPilot DeepSpeed

OpenAI Anthropic Google Cohere

HuggingFace Ollama Replicate

LlamaIndex LangChain Haystack

TRL PEFT DSPy Axolotl

MODEL TRAINING & EXPERIMENTATION

Training Frameworks
Build, train, and iterate on neural networks and ML models at any scale

PyTorch TensorFlow JAX Keras

Classical ML & Data Science
Statistical models, tabular ML, preprocessing, and exploratory analysis

scikit-learn XGBoost Pandas NumPy

Experiment Tracking
Log runs, compare metrics, manage hyperparameters, reproduce results

MLflow W&B Neptune CometML

AutoML & HPO
Automated model selection, hyperparameter search, and neural architecture search

Optuna Ray Tune AutoGluon

THE FOUNDATIONS FOR DATA

DATA & STORAGE

Relational Databases
Structured data storage, transactions, SQL queries for tabular business data

PostgreSQL MySQL Snowflake

Vector Databases
Similarity search over embeddings for retrieval & RAG pipelines

Pinecone Weaviate Milvus

Data Lakes & Warehouses
Scalable storage for massive datasets across structured & unstructured formats

Databricks BigQuery Redshift

Data Pipelines & ETL
Ingestion, transformation, scheduling, and orchestration of data workflows

Airflow dbt Spark Kafka

Feature Stores
Manage, serve, and reuse engineered features across models and teams

Feast Tecton Hopsworks

A LOT OF INNOVATION POSSIBLE ON FOUNDATIONS

COMPUTE & HARDWARE

GPU & Accelerators
Specialized hardware powering all modern training and inference workloads

NVIDIA H100 AMD MI300 Google TPU Trainium

Cloud AI Platforms
Managed infrastructure for training, serving, and scaling AI workloads

AWS GCP Azure Lambda Labs

Cluster & Job Management
Schedule, scale, and manage distributed compute jobs across clusters

Kubernetes Slurm Ray SkyPilot DeepSpeed

OpenAI Anthropic Google Cohere

HuggingFace Ollama Replicate

LlamaIndex LangChain Haystack

TRL PEFT DSPy Axolotl

MODEL TRAINING & EXPERIMENTATION

Training Frameworks
Build, train, and iterate on neural networks and ML models at any scale

PyTorch TensorFlow JAX Keras

Classical ML & Data Science
Statistical models, tabular ML, preprocessing, and exploratory analysis

scikit-learn XGBoost Pandas NumPy

Experiment Tracking
Log runs, compare metrics, manage hyperparameters, reproduce results

MLflow W&B Neptune CometML

AutoML & HPO
Automated model selection, hyperparameter search, and neural architecture search

Optuna Ray Tune AutoGluon

THE FOUNDATIONS FOR DATA

DATA & STORAGE

Relational Databases
Structured data storage, transactions, SQL queries for tabular business data

PostgreSQL MySQL Snowflake

Vector Databases
Similarity search over embeddings for retrieval & RAG pipelines

Pinecone Weaviate Milvus

Data Lakes & Warehouses
Scalable storage for massive datasets across structured & unstructured formats

Databricks BigQuery Redshift

Data Pipelines & ETL
Ingestion, transformation, scheduling, and orchestration of data workflows

Airflow dbt Spark Kafka

Feature Stores
Manage, serve, and reuse engineered features across models and teams

Feast Tecton Hopsworks

A LOT OF INNOVATION POSSIBLE ON FOUNDATIONS

COMPUTE & HARDWARE

GPU & Accelerators
Specialized hardware powering all modern training and inference workloads

NVIDIA H100 AMD MI300 Google TPU Trainium

Cloud AI Platforms
Managed infrastructure for training, serving, and scaling AI workloads

AWS GCP Azure Lambda Labs

Cluster & Job Management
Schedule, scale, and manage distributed compute jobs across clusters

Kubernetes Slurm Ray SkyPilot DeepSpeed

MONITORING, EVALUATION & SAFETY

LLM Evaluation
Benchmark, score, and compare LLM outputs across tasks and safety metrics

`LM Eval Harness` `HELM` `Ragas`

Drift & Monitoring
Detect data drift, concept drift, and model degradation in production

`Evidently` `WhyLabs` `Fiddler`

Observability & Tracing
Trace LLM calls, agent steps, latency, and costs across the stack

`LangSmith` `Arize` `Helicone`

Guardrails & Safety
Content filters, output validation, toxicity detection, policy enforcement

`Guardrails AI` `NeMo Guardr.`

SERVING, DEPLOYMENT & INFERENCE

LLM Serving
High-throughput, low-latency serving engines for large language models

`vLLM` `TGI` `TensorRT-LLM` `SGLang`

Model Serving
Deploy, version, and serve ML models behind production APIs

`TorchServe` `Triton` `BentoML` `Seldon`

MLOps & Registries
Model versioning, CI/CD, packaging, and reproducible deployment pipelines

`MLflow` `SageMaker` `Vertex AI` `Kubeflow`

Compression & Edge
Quantization, pruning, distillation for on-device and low-latency deployment

`GPTQ` `AWQ` `llama.cpp` `ONNX`

LLM & GENERATIVE AI INFRASTRUCTURE

Foundation Model APIs
Access to frontier LLMs via API for reasoning, generation, and dialog

`OpenAI` `Anthropic` `Google` `Cohere`

Open Model Hubs
Download, share, and deploy open-weight models and datasets

`HuggingFace` `Ollama` `Replicate`

RAG & Retrieval
Augment LLMs with external knowledge via retrieval pipelines and chunking

`LlamaIndex` `LangChain` `Haystack`

Fine-Tuning & Alignment
Adapt foundation models via RLHF, LoRA, prompt tuning, and distillation

`TRL` `PEFT` `DSPy` `Axolotl`

In practice, building an AI solution means selecting and manually stitching together >>1 AI systems from a massive set of tools into a working whole

AGENT FRAMEWORKS & ORCHESTRATION

Agent Frameworks
Build autonomous agents that reason, plan, use tools, and take actions

`LangGraph` `CrewAI` `AutoGen`

Multi-Agent Systems
Coordinate teams of agents with roles, delegation, and consensus protocols

`CrewAI` `AutoGen` `MetaGPT`

Tool & API Integration
Connect agents to external services, functions, databases, and code execution

`MCP` `Function Calling`

Workflow Orchestration
DAG-based pipelines, conditional routing, and scheduling of AI workflows

`Prefect` `Dagster` `Temporal`

MONITORING, EVALUATION & SAFETY

LLM Evaluation
Benchmark, score, and compare LLM outputs across tasks and safety metrics

`LM Eval Harness` `HELM` `Ragas`

Drift & Monitoring
Detect data drift, concept drift, and model degradation in production

`Evidently` `WhyLabs` `Fiddler`

Observability & Tracing
Trace LLM calls, agent steps, latency, and costs across the stack

`LangSmith` `Arize` `Helicone`

Guardrails & Safety
Content filters, output validation, toxicity detection, policy enforcement

`Guardrails AI` `NeMo Guardr.`

THERE ARE THREE CRITICAL FEATURES IN AI DEVELOPMENT

SPEED, SPEED, & SPEED

Time to market, Better models, More models

~90%

OF EFFORT IN AI GOES INTO “GLUE ENGINEERING”

OF LATENCY AND $$$ IN AI IS BECAUSE OF STORAGE