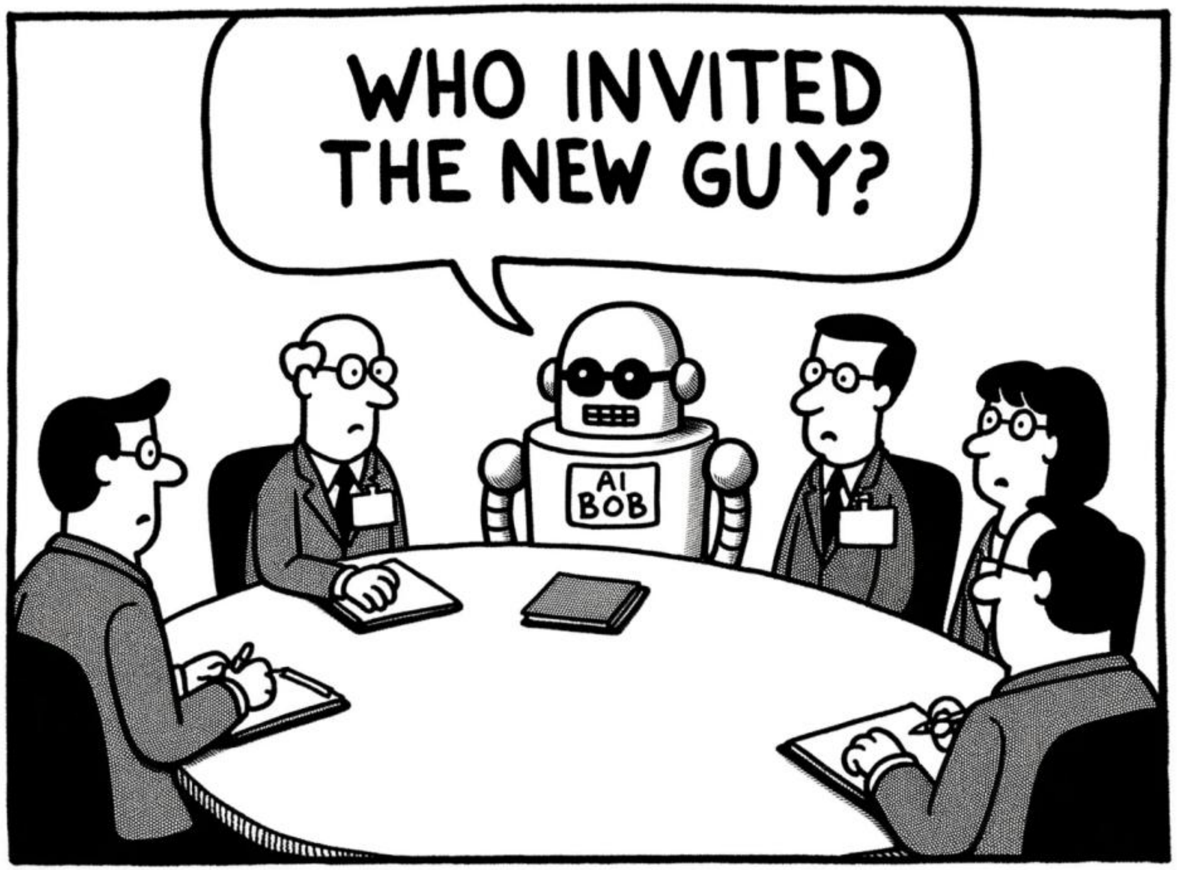
Operationalizing AI: Infrastructure to Agents

Unlocking Business Value at Scale

2026

Amit Nandi

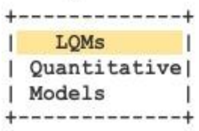
AgentOps



https://www.linkedin.com/pulse/easy-guide-autonomous-reasoning-agents-aras-paul-j-thompson-4wvyc/

Investment Banking

Large Quantitative Models (LQMs)



```
+-------------+
|   LQMs      |
| Quantitative|
| Models      |
+-------------+
```

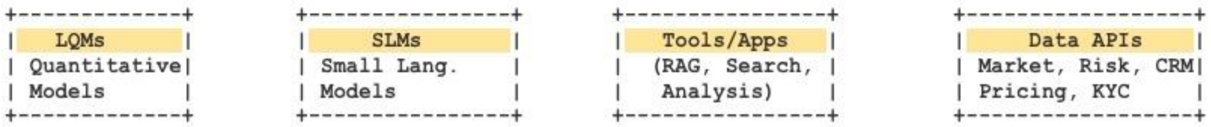
Investment Banking

Large Quantitative Models (LQMs) + Small Language Models (SLMs)

```
+-------------+      +---------------+
|    LQMs     |      |     SLMs      |
| Quantitative|      | Small Lang.   |
| Models      |      | Models        |
+-------------+      +---------------+
```

Investment Banking

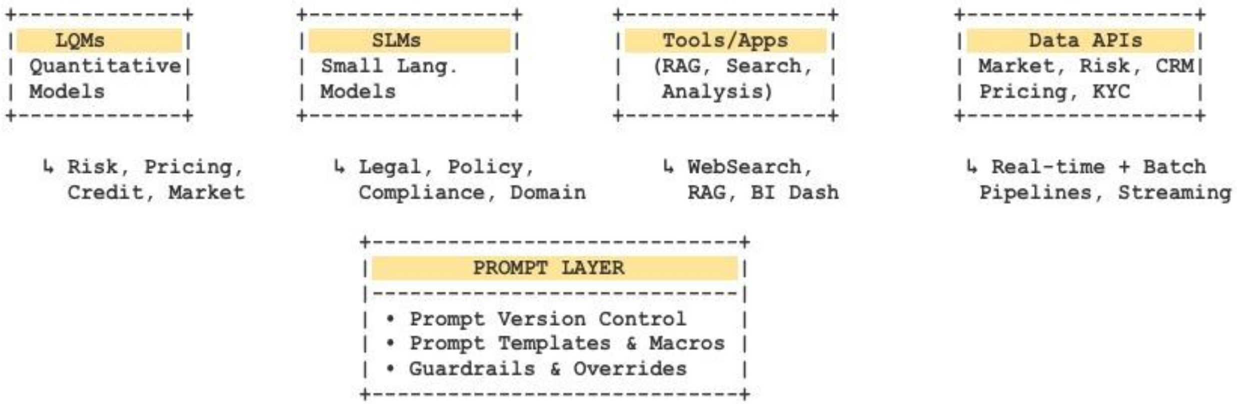
Large Quantitative Models (LQMs) + Small Language Models (SLMs) +RAGs



```
+-------------+     +----------------+     +----------------+     +------------------+
|    LQMs     |     |      SLMs      |     |   Tools/Apps   |     |     Data APIs    |
| Quantitative|     | Small Lang.    |     |  (RAG, Search, |     | Market, Risk, CRM|
| Models      |     | Models         |     |   Analysis)    |     | Pricing, KYC     |
+-------------+     +----------------+     +----------------+     +------------------+
```

Investment Banking

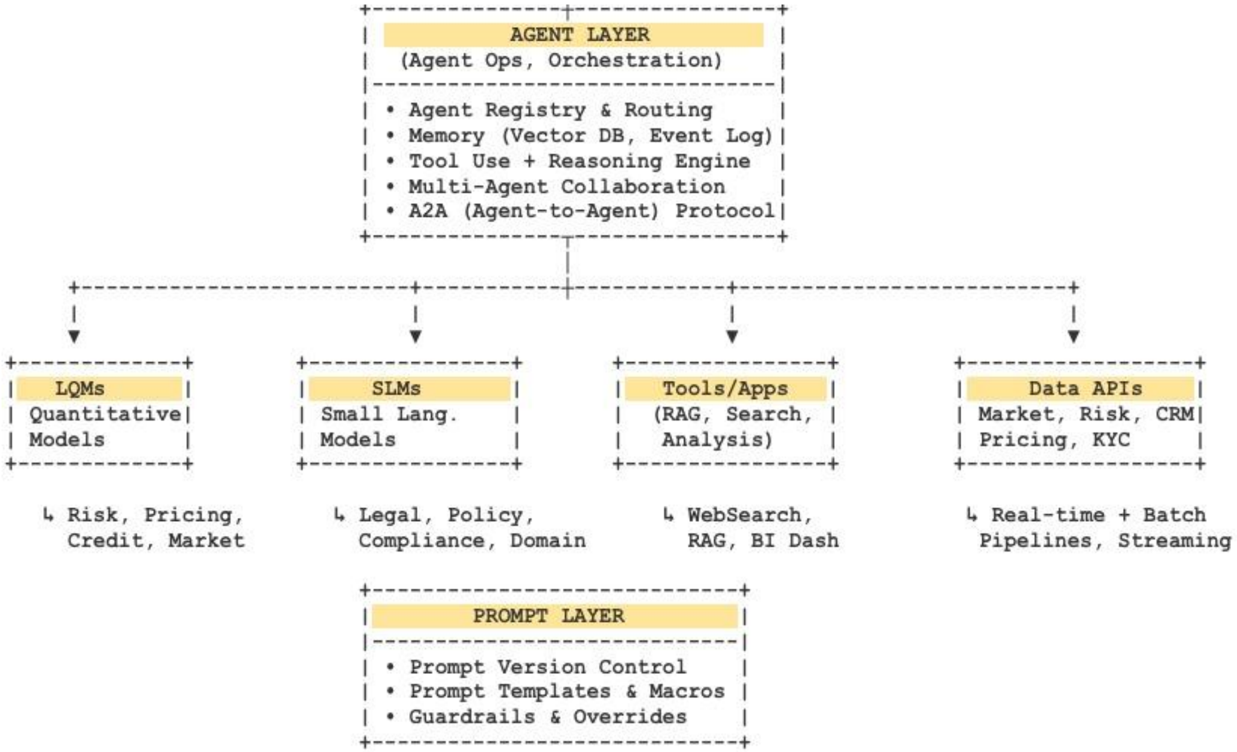
Large Quantitative Models (LQMs) + Small Language Models (SMLs) + RAGs



```
+-------------+     +----------------+     +---------------+     +------------------+
|    LQMs     |     |      SLMs      |     |   Tools/Apps  |     |     Data APIs    |
| Quantitative|     | Small Lang.    |     | (RAG, Search, |     | Market, Risk, CRM|
| Models      |     | Models         |     |  Analysis)    |     | Pricing, KYC     |
+-------------+     +----------------+     +---------------+     +------------------+

   ↳ Risk, Pricing,     ↳ Legal, Policy,       ↳ WebSearch,         ↳ Real-time + Batch
     Credit, Market       Compliance, Domain     RAG, BI Dash         Pipelines, Streaming

                    +----------------------------+
                    |        PROMPT LAYER        |
                    |----------------------------|
                    | • Prompt Version Control   |
                    | • Prompt Templates & Macros |
                    | • Guardrails & Overrides   |
                    +----------------------------+
```

Investment Banking

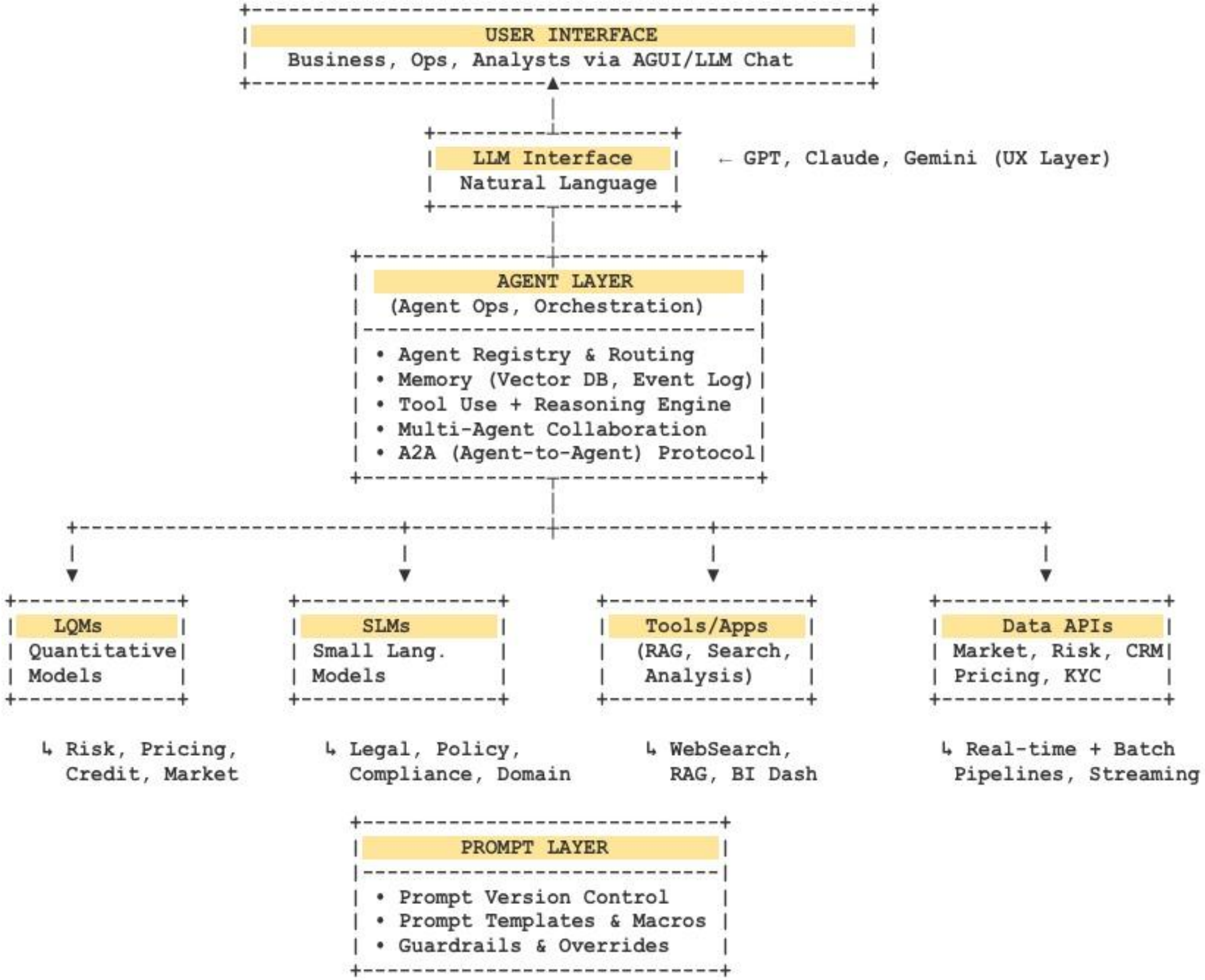
Large Quantitative Models (LQMs) + Small Language Models (SMLs) + RAGs + Agents



```
                      +---------------+---------------+
                      |          AGENT LAYER          |
                      |  (Agent Ops, Orchestration)   |
                      |-------------------------------|
                      | • Agent Registry & Routing    |
                      | • Memory (Vector DB, Event Log)|
                      | • Tool Use + Reasoning Engine |
                      | • Multi-Agent Collaboration   |
                      | • A2A (Agent-to-Agent) Protocol|
                      +---------------+---------------+
                                      |
     +-------------------------+------+-----------+------------------------+
     |                         |                  |                        |
     ▼                         ▼                  ▼                        ▼
+-------------+      +----------------+    +----------------+     +-----------------+
|    LQMs     |      |      SLMs      |    |   Tools/Apps   |     |    Data APIs    |
| Quantitative|      | Small Lang.    |    | (RAG, Search,  |     | Market, Risk, CRM|
| Models      |      | Models         |    |  Analysis)     |     | Pricing, KYC    |
+-------------+      +----------------+    +----------------+     +-----------------+

   ↳ Risk, Pricing,      ↳ Legal, Policy,      ↳ WebSearch,         ↳ Real-time + Batch
     Credit, Market        Compliance, Domain     RAG, BI Dash        Pipelines, Streaming

                      +-----------------------------+
                      |        PROMPT LAYER         |
                      |-----------------------------|
                      | • Prompt Version Control    |
                      | • Prompt Templates & Macros |
                      | • Guardrails & Overrides    |
                      +-----------------------------+
```

Investment Banking

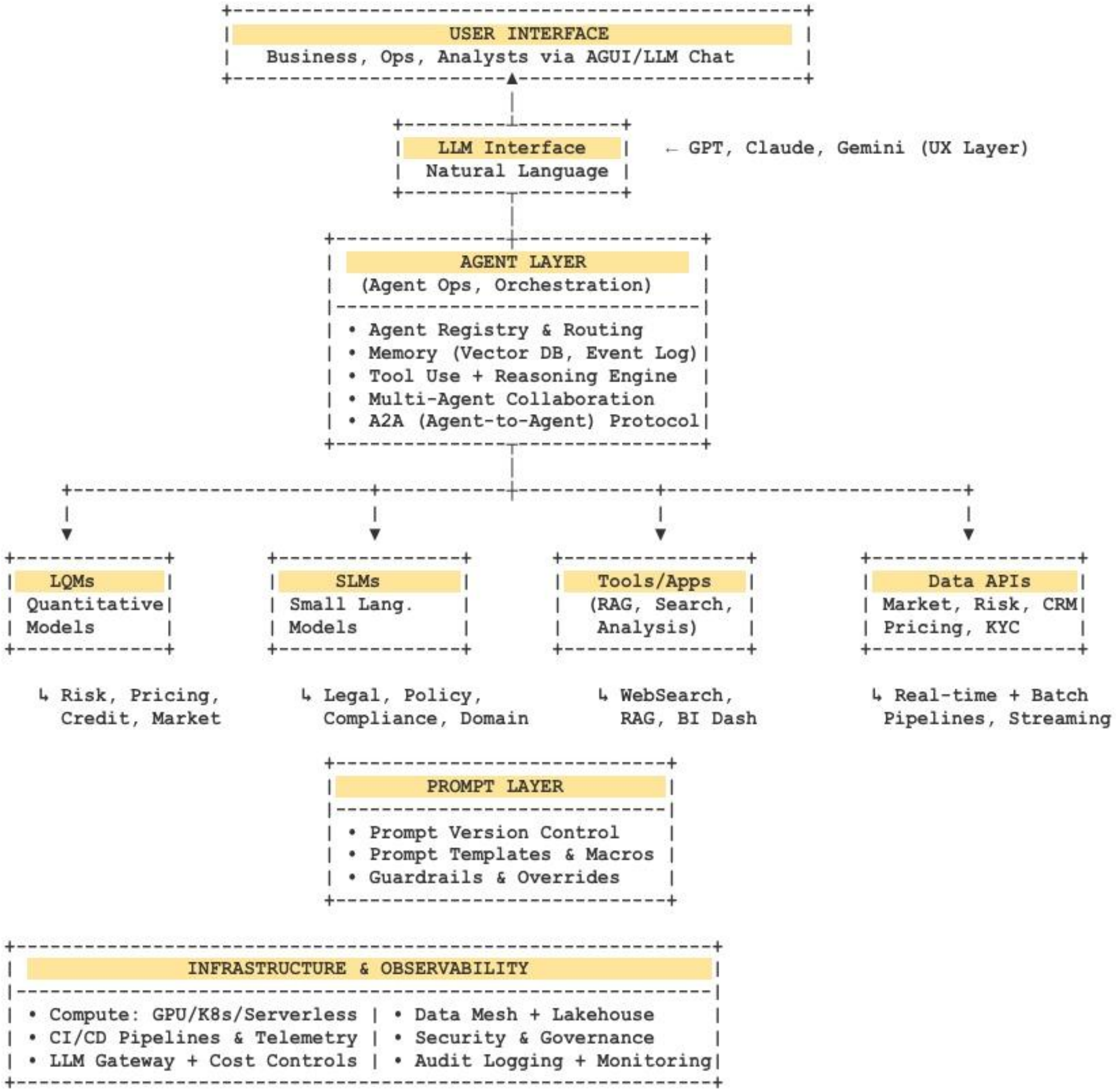
Large Quantitative Models (LQMs) + Small Language Models (SMLs) + RAGs + Agents



```
+---------------------------------------------------+
|                 USER INTERFACE                    |
|    Business, Ops, Analysts via AGUI/LLM Chat      |
+-------------------------▲-------------------------+
                          |
               +--------------------+
               |   LLM Interface    |   ← GPT, Claude, Gemini (UX Layer)
               |  Natural Language  |
               +--------------------+
                          |
         +---------------------------------+
         |          AGENT LAYER            |
         |  (Agent Ops, Orchestration)     |
         |---------------------------------|
         | • Agent Registry & Routing      |
         | • Memory (Vector DB, Event Log) |
         | • Tool Use + Reasoning Engine   |
         | • Multi-Agent Collaboration     |
         | • A2A (Agent-to-Agent) Protocol |
         +---------------------------------+
                          |
 +------------------------+-----------+------------------------+
 |                        |           |                        |
 ▼                        ▼           ▼                        ▼
+--------------+   +-----------------+   +-----------------+   +------------------+
|    LQMs      |   |      SLMs       |   |   Tools/Apps    |   |    Data APIs     |
| Quantitative|   | Small Lang.     |   |  (RAG, Search,  |   | Market, Risk, CRM|
| Models       |   | Models          |   |   Analysis)     |   | Pricing, KYC     |
+--------------+   +-----------------+   +-----------------+   +------------------+

   ↳ Risk, Pricing,    ↳ Legal, Policy,       ↳ WebSearch,        ↳ Real-time + Batch
     Credit, Market      Compliance, Domain     RAG, BI Dash       Pipelines, Streaming

         +------------------------------+
         |        PROMPT LAYER          |
         |------------------------------|
         | • Prompt Version Control     |
         | • Prompt Templates & Macros  |
         | • Guardrails & Overrides     |
         +------------------------------+
```

Investment Banking

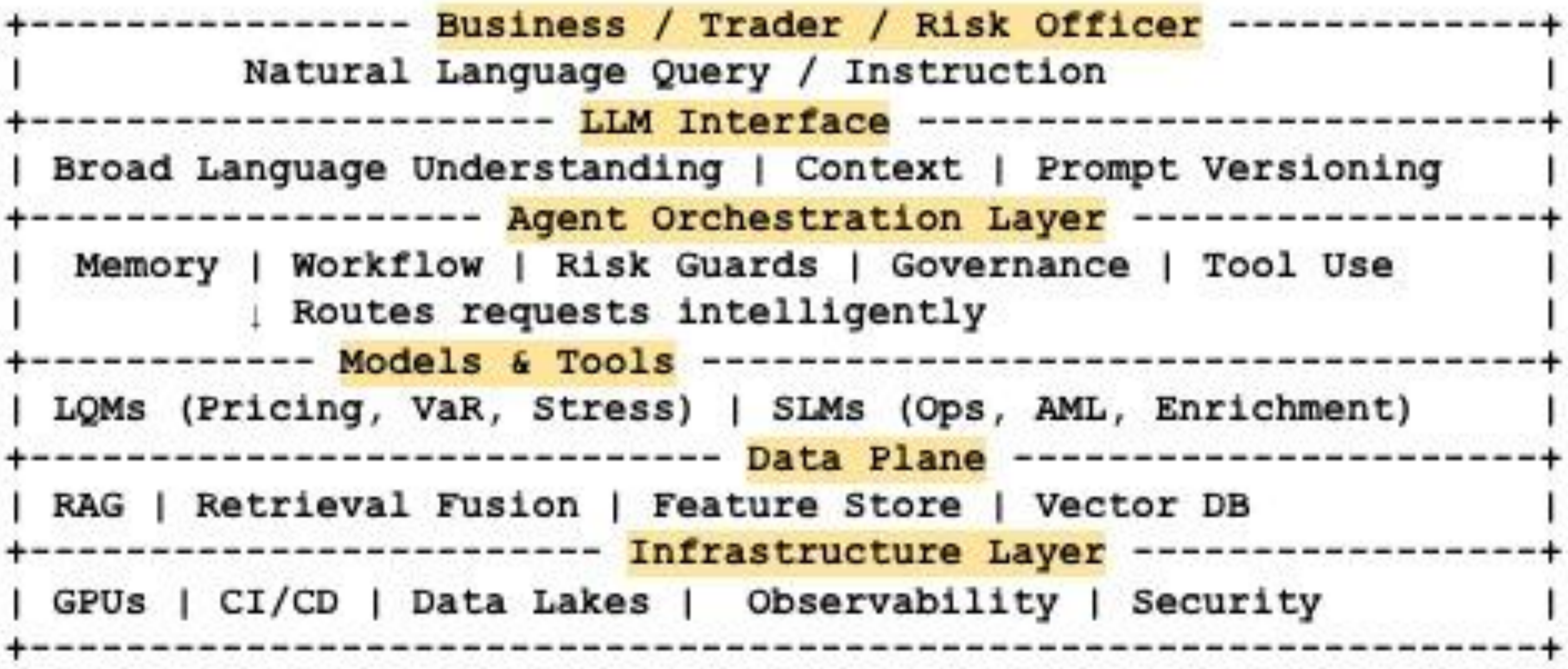
LQMs + SLMs + RAGs Orchestrated by Agents



```
                 +--------------------------------------------------+
                 |                 USER INTERFACE                   |
                 |   Business, Ops, Analysts via AGUI/LLM Chat      |
                 +------------------------▲-------------------------+
                                          |
                               +--------------------+
                               |   LLM Interface    |   ← GPT, Claude, Gemini (UX Layer)
                               |  Natural Language  |
                               +---------┬----------+
                                         |
                         +--------------------------------+
                         |          AGENT LAYER           |
                         |  (Agent Ops, Orchestration)    |
                         |--------------------------------|
                         | • Agent Registry & Routing     |
                         | • Memory (Vector DB, Event Log)|
                         | • Tool Use + Reasoning Engine  |
                         | • Multi-Agent Collaboration    |
                         | • A2A (Agent-to-Agent) Protocol|
                         +---------------┬----------------+
                                         |
   +--------------------------+----------+------------+--------------------------+
   |                          |                       |                          |
   ▼                          ▼                       ▼                          ▼
+--------------+       +-----------------+     +-----------------+       +------------------+
|    LQMs      |       |      SLMs       |     |   Tools/Apps    |       |    Data APIs     |
| Quantitative |       | Small Lang.     |     |  (RAG, Search,  |       | Market, Risk, CRM|
| Models       |       | Models          |     |   Analysis)     |       | Pricing, KYC     |
+--------------+       +-----------------+     +-----------------+       +------------------+

   ↳ Risk, Pricing,      ↳ Legal, Policy,         ↳ WebSearch,              ↳ Real-time + Batch
     Credit, Market        Compliance, Domain       RAG, BI Dash              Pipelines, Streaming

                         +-----------------------------+
                         |        PROMPT LAYER         |
                         |-----------------------------|
                         | • Prompt Version Control    |
                         | • Prompt Templates & Macros |
                         | • Guardrails & Overrides    |
                         +-----------------------------+

+--------------------------------------------------------------+
|             INFRASTRUCTURE & OBSERVABILITY                   |
|--------------------------------------------------------------|
| • Compute: GPU/K8s/Serverless | • Data Mesh + Lakehouse      |
| • CI/CD Pipelines & Telemetry | • Security & Governance      |
| • LLM Gateway + Cost Controls | • Audit Logging + Monitoring|
+--------------------------------------------------------------+
```

Investment Banking

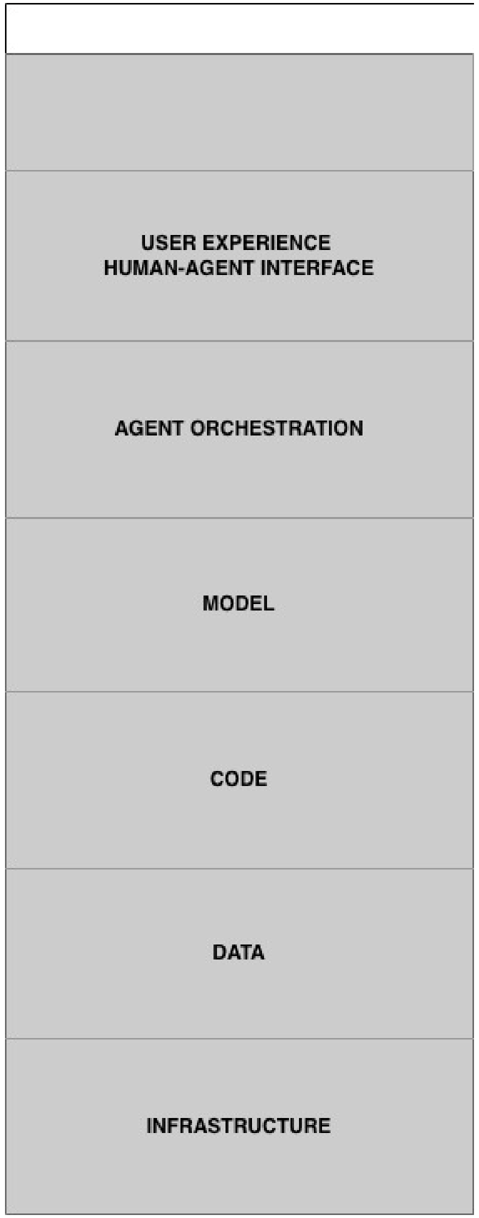
Activating the Stack



```
+---------------- Business / Trader / Risk Officer -------------+
|          Natural Language Query / Instruction                 |
+---------------------- LLM Interface --------------------------+
| Broad Language Understanding | Context | Prompt Versioning    |
+------------------- Agent Orchestration Layer -----------------+
|  Memory | Workflow | Risk Guards | Governance | Tool Use      |
|          ↓ Routes requests intelligently                      |
+------------ Models & Tools -----------------------------------+
| LQMs (Pricing, VaR, Stress) | SLMs (Ops, AML, Enrichment)     |
+---------------------------- Data Plane -----------------------+
| RAG | Retrieval Fusion | Feature Store | Vector DB            |
+------------------------ Infrastructure Layer -----------------+
| GPUs | CI/CD | Data Lakes |  Observability | Security         |
+---------------------------------------------------------------+
```

Scaling AI Operations: MLOps -> LLMOps -> AgentOps

Ops Framework



USER EXPERIENCE
HUMAN-AGENT INTERFACE

AGENT ORCHESTRATION

MODEL

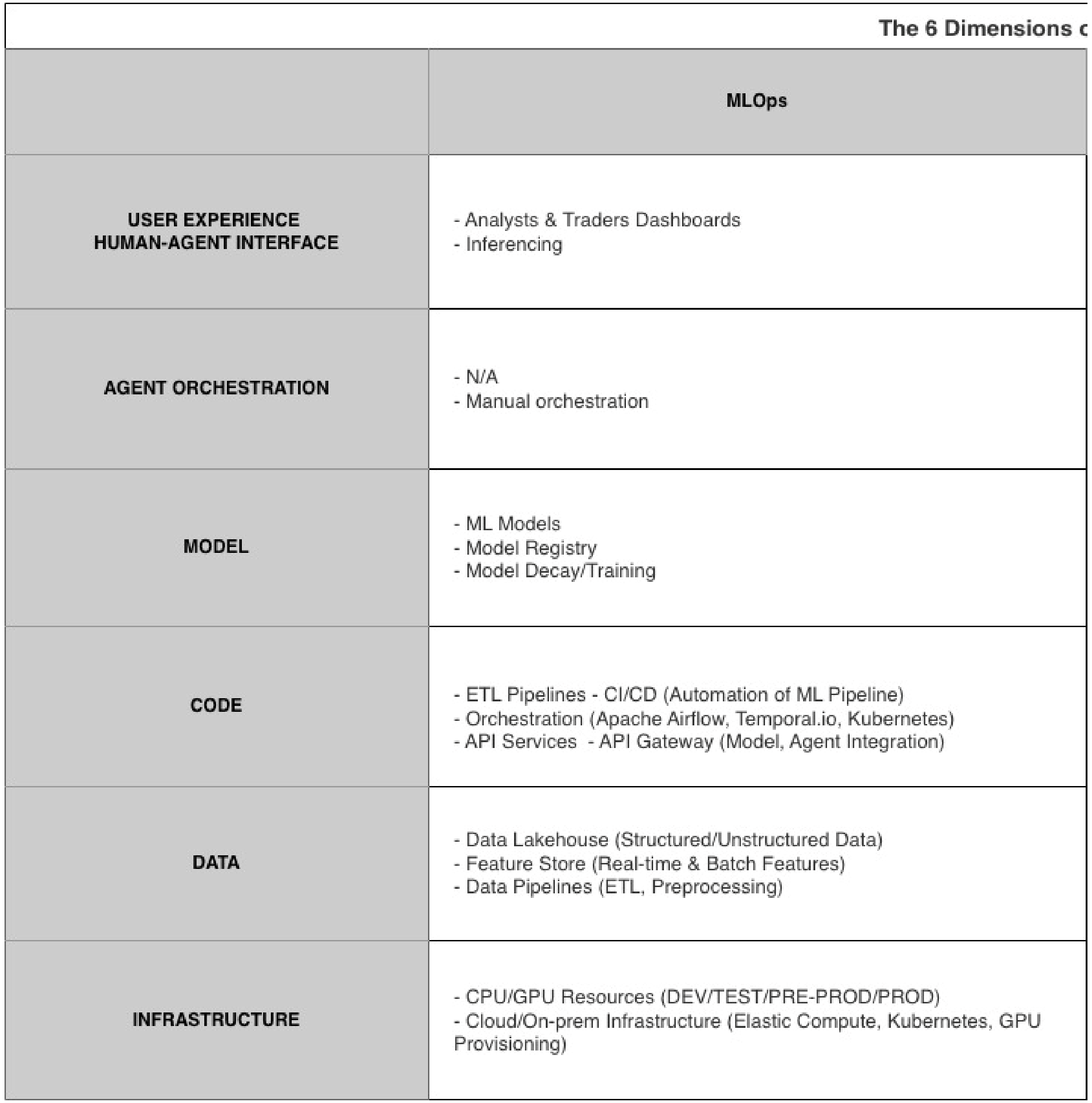
CODE

DATA

INFRASTRUCTURE

Scaling AI Operations: MLOps -> LLMOps -> AgentOps

Ops Framework

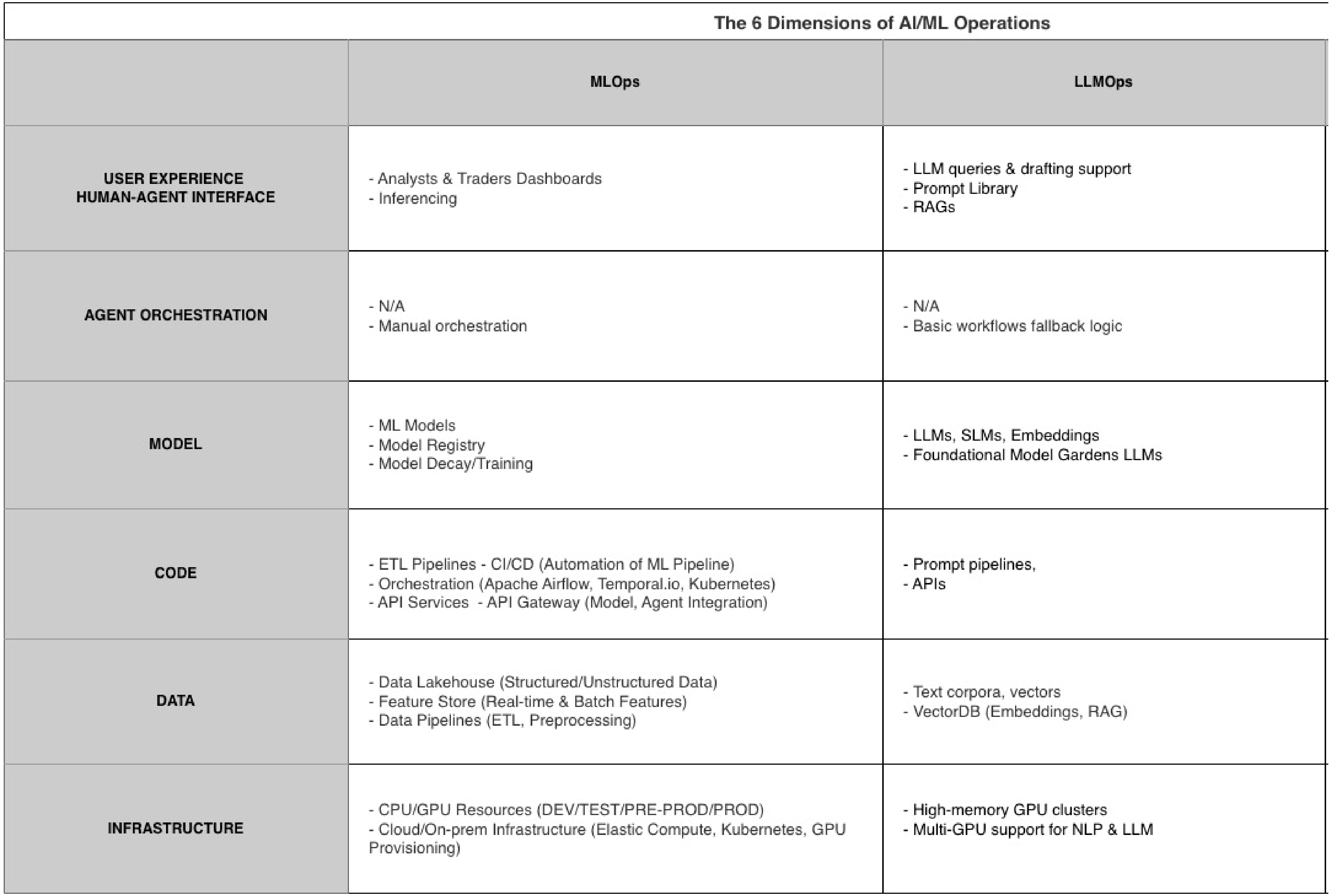


The 6 Dimensions c

	MLOps
USER EXPERIENCE HUMAN-AGENT INTERFACE	- Analysts & Traders Dashboards - Inferencing
AGENT ORCHESTRATION	- N/A - Manual orchestration
MODEL	- ML Models - Model Registry - Model Decay/Training
CODE	- ETL Pipelines - CI/CD (Automation of ML Pipeline) - Orchestration (Apache Airflow, Temporal.io, Kubernetes) - API Services - API Gateway (Model, Agent Integration)
DATA	- Data Lakehouse (Structured/Unstructured Data) - Feature Store (Real-time & Batch Features) - Data Pipelines (ETL, Preprocessing)
INFRASTRUCTURE	- CPU/GPU Resources (DEV/TEST/PRE-PROD/PROD) - Cloud/On-prem Infrastructure (Elastic Compute, Kubernetes, GPU Provisioning)

Scaling AI Operations: MLOps -> LLMOps -> AgentOps

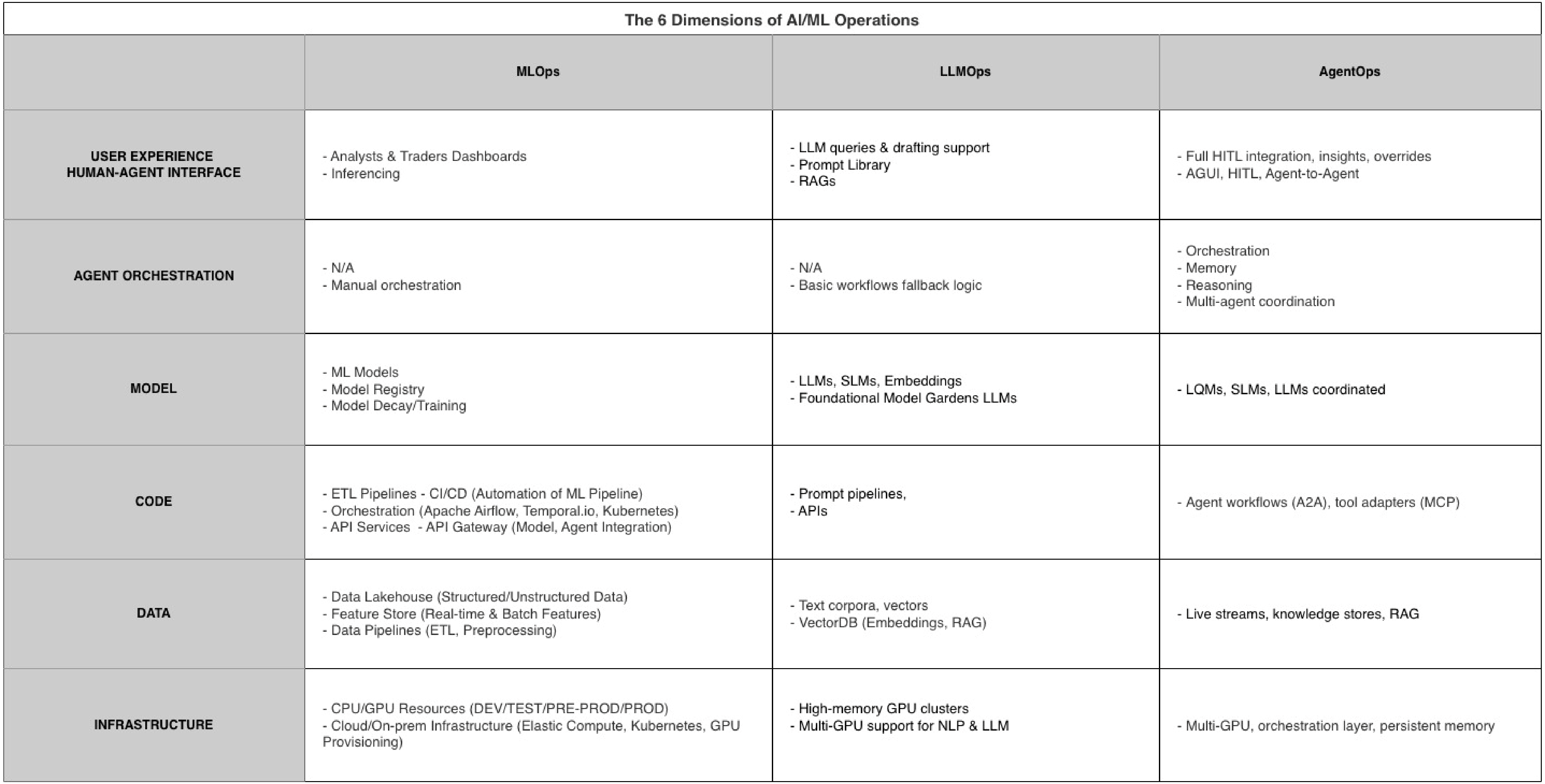
Ops Framework



The 6 Dimensions of AI/ML Operations		
	MLOps	LLMOps
USER EXPERIENCE HUMAN-AGENT INTERFACE	- Analysts & Traders Dashboards - Inferencing	- LLM queries & drafting support - Prompt Library - RAGs
AGENT ORCHESTRATION	- N/A - Manual orchestration	- N/A - Basic workflows fallback logic
MODEL	- ML Models - Model Registry - Model Decay/Training	- LLMs, SLMs, Embeddings - Foundational Model Gardens LLMs
CODE	- ETL Pipelines - CI/CD (Automation of ML Pipeline) - Orchestration (Apache Airflow, Temporal.io, Kubernetes) - API Services - API Gateway (Model, Agent Integration)	- Prompt pipelines, - APIs
DATA	- Data Lakehouse (Structured/Unstructured Data) - Feature Store (Real-time & Batch Features) - Data Pipelines (ETL, Preprocessing)	- Text corpora, vectors - VectorDB (Embeddings, RAG)
INFRASTRUCTURE	- CPU/GPU Resources (DEV/TEST/PRE-PROD/PROD) - Cloud/On-prem Infrastructure (Elastic Compute, Kubernetes, GPU Provisioning)	- High-memory GPU clusters - Multi-GPU support for NLP & LLM

Scaling AI Operations: MLOps -> LLMOps -> AgentOps

Ops Framework

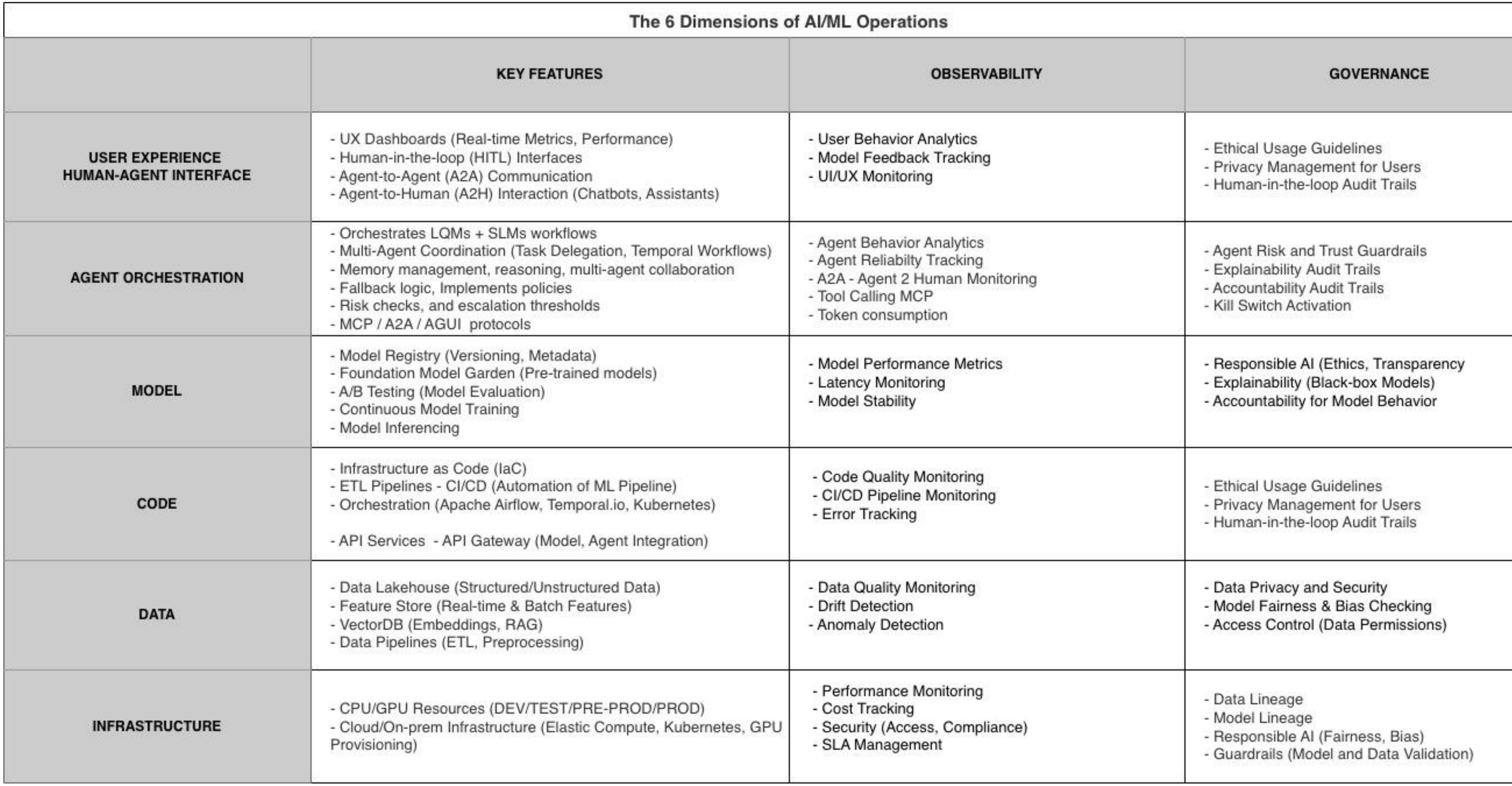


The 6 Dimensions of AI/ML Operations

	MLOps	LLMOps	AgentOps
USER EXPERIENCE HUMAN-AGENT INTERFACE	- Analysts & Traders Dashboards - Inferencing	- LLM queries & drafting support - Prompt Library - RAGs	- Full HITL integration, insights, overrides - AGUI, HITL, Agent-to-Agent
AGENT ORCHESTRATION	- N/A - Manual orchestration	- N/A - Basic workflows fallback logic	- Orchestration - Memory - Reasoning - Multi-agent coordination
MODEL	- ML Models - Model Registry - Model Decay/Training	- LLMs, SLMs, Embeddings - Foundational Model Gardens LLMs	- LQMs, SLMs, LLMs coordinated
CODE	- ETL Pipelines - CI/CD (Automation of ML Pipeline) - Orchestration (Apache Airflow, Temporal.io, Kubernetes) - API Services - API Gateway (Model, Agent Integration)	- Prompt pipelines, - APIs	- Agent workflows (A2A), tool adapters (MCP)
DATA	- Data Lakehouse (Structured/Unstructured Data) - Feature Store (Real-time & Batch Features) - Data Pipelines (ETL, Preprocessing)	- Text corpora, vectors - VectorDB (Embeddings, RAG)	- Live streams, knowledge stores, RAG
INFRASTRUCTURE	- CPU/GPU Resources (DEV/TEST/PRE-PROD/PROD) - Cloud/On-prem Infrastructure (Elastic Compute, Kubernetes, GPU Provisioning)	- High-memory GPU clusters - Multi-GPU support for NLP & LLM	- Multi-GPU, orchestration layer, persistent memory

Scaling AI Operations: MLOps -> LLMOps -> AgentOps

Ops Framework



The 6 Dimensions of AI/ML Operations

	KEY FEATURES	OBSERVABILITY	GOVERNANCE
USER EXPERIENCE HUMAN-AGENT INTERFACE	- UX Dashboards (Real-time Metrics, Performance) - Human-in-the-loop (HITL) Interfaces - Agent-to-Agent (A2A) Communication - Agent-to-Human (A2H) Interaction (Chatbots, Assistants)	- User Behavior Analytics - Model Feedback Tracking - UI/UX Monitoring	- Ethical Usage Guidelines - Privacy Management for Users - Human-in-the-loop Audit Trails
AGENT ORCHESTRATION	- Orchestrates LQMs + SLMs workflows - Multi-Agent Coordination (Task Delegation, Temporal Workflows) - Memory management, reasoning, multi-agent collaboration - Fallback logic, Implements policies - Risk checks, and escalation thresholds - MCP / A2A / AGUI protocols	- Agent Behavior Analytics - Agent Reliabilty Tracking - A2A - Agent 2 Human Monitoring - Tool Calling MCP - Token consumption	- Agent Risk and Trust Guardrails - Explainability Audit Trails - Accountability Audit Trails - Kill Switch Activation
MODEL	- Model Registry (Versioning, Metadata) - Foundation Model Garden (Pre-trained models) - A/B Testing (Model Evaluation) - Continuous Model Training - Model Inferencing	- Model Performance Metrics - Latency Monitoring - Model Stability	- Responsible AI (Ethics, Transparency - Explainability (Black-box Models) - Accountability for Model Behavior
CODE	- Infrastructure as Code (IaC) - ETL Pipelines - CI/CD (Automation of ML Pipeline) - Orchestration (Apache Airflow, Temporal.io, Kubernetes) - API Services - API Gateway (Model, Agent Integration)	- Code Quality Monitoring - CI/CD Pipeline Monitoring - Error Tracking	- Ethical Usage Guidelines - Privacy Management for Users - Human-in-the-loop Audit Trails
DATA	- Data Lakehouse (Structured/Unstructured Data) - Feature Store (Real-time & Batch Features) - VectorDB (Embeddings, RAG) - Data Pipelines (ETL, Preprocessing)	- Data Quality Monitoring - Drift Detection - Anomaly Detection	- Data Privacy and Security - Model Fairness & Bias Checking - Access Control (Data Permissions)
INFRASTRUCTURE	- CPU/GPU Resources (DEV/TEST/PRE-PROD/PROD) - Cloud/On-prem Infrastructure (Elastic Compute, Kubernetes, GPU Provisioning)	- Performance Monitoring - Cost Tracking - Security (Access, Compliance) - SLA Management	- Data Lineage - Model Lineage - Responsible AI (Fairness, Bias) - Guardrails (Model and Data Validation)

ML/AI People

Operations

Advance Analytics Team
Data Lake

Data Engineer

Prepare & Ingest data building ETL pipelines

Data Owners

Manage data sharing and provide access

Data Science Team
Experimentation & MLOps

Data Scientist

Create the best ML models to solve business problems

ML Engineer

Collaborate with DS to productionize ML

Platform Team
Secure Cloud/Data/ML Platform

MLOps Engineer/Admin

Standardize CI/CD, user/service role, model consumption, testing and deployment methodology

Security

Assess data, user, and service access creating policies and guardrails

Architects/ SysOps Engineer

Standardize account infrastructure, connectivity, user roles implementation

Business
Viz Dashboards, ML Adoption, & ROI

Business Stakeholder
Product Owners

Define business problem, business KPIs, and make business decisions

Business Stakeholder
Data & ML Consumers

Consumers of ML results from other BUs, driving business decision making

Risk & Compliance
Approve & Review Models

Auditors/Risk & Compliance

Review models, data sources, code artifacts

https://aws.amazon.com/blogs/machine-learning/fmops-llmops-operationalize-generative-ai-and-differences-with-mlops/

ML/AI + LLMs + Agentic AI People

Operations

Advance Analytics Team
Data Lake

Data Engineer

Prepare & Ingest data building ETL pipelines

Data Owners

Manage data sharing and provide access

Data Science Team
Experimentation & MLOps

Data Scientist

Create the best ML models to solve business problems

ML Engineer

Collaborate with DS to productionize ML

Labeler Team
Data Preparation at Scale

Data Labelers/Editors

Label or edit billions of Data for FM models and hundreds of data for fine tuning interacting with data lake using a dedicated website

Data Science Team Extension
Context Adaptation

Fine Tuners

Select the corresponding FM, evaluate the model & design the deployment method/infrastructure

Platform Team
Secure Cloud/Data/ML Platform

MLOps Engineer/Admin

Standardize CI/CD, user/service role, model consumption, testing and deployment methodology

Security & Architects

Assess data, user, and service access creating policies and infrastructure

Risk & Compliance
Approve & Review Models

Auditors/Risk & Compliance

Review models, data sources, code artifacts

Business
Viz Dashboards, ML Adoption, & ROI

Business Stakeholder Product Owners

Define business problem, business KPIs, and make business decisions

Business Stakeholder Data & ML Consumers

Consumers of ML results from other BUs, driving business decision making

End-Users
Consume Generative AI applications

Generative AI End-users

Consume Generative AI solutions as black box, share data and rate the quality of output

GenAI Application Team
Integrate GenAI models in applications

Generative AI Developers

Select, test, evaluate the FM, filter inputs/outputs, and develop the generative AI application back-end (e.g. LangChain Experts)

AppDev/DevOps

Develop the front-end of the GenAI application

Prompt Engineers

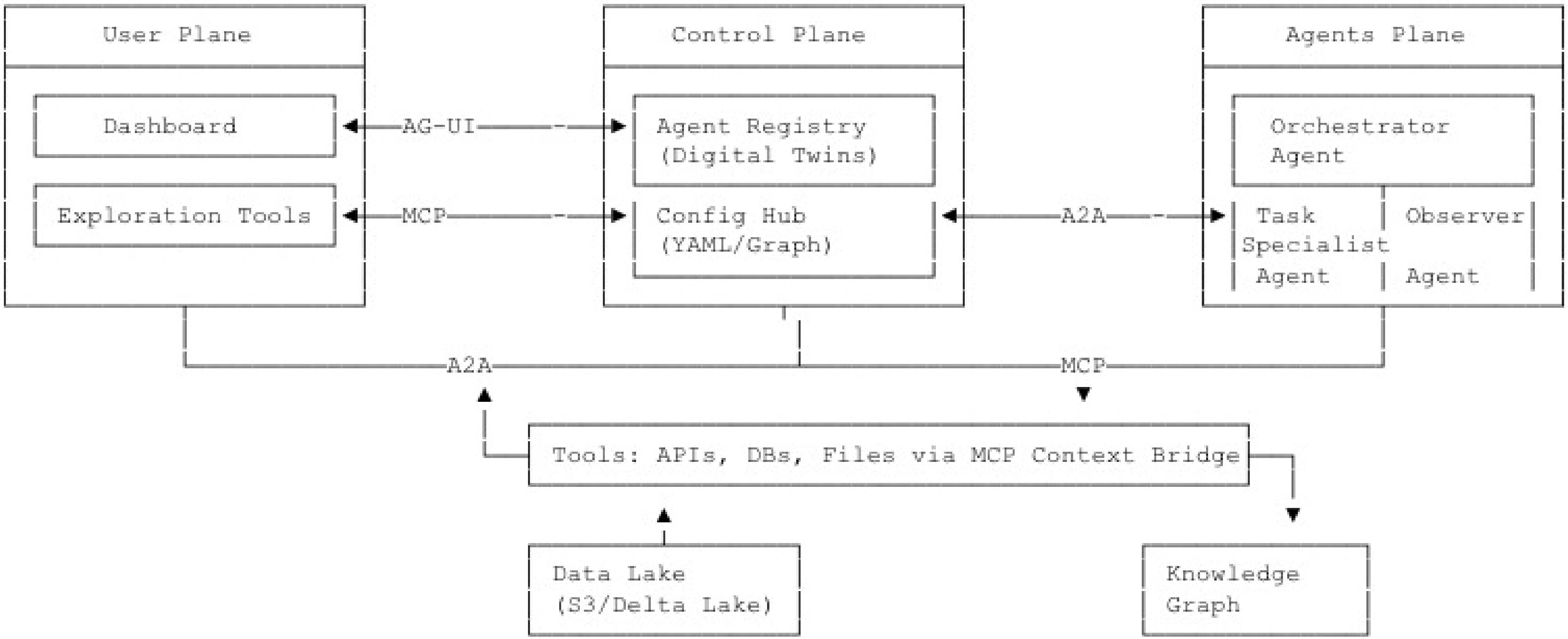
Design the input/output prompts to adapt the solution to the context and test the initial version

Prompt Testers

Test at scale the generative. AI solution (back-end/ front-end) and feed their results to the prompt catalog

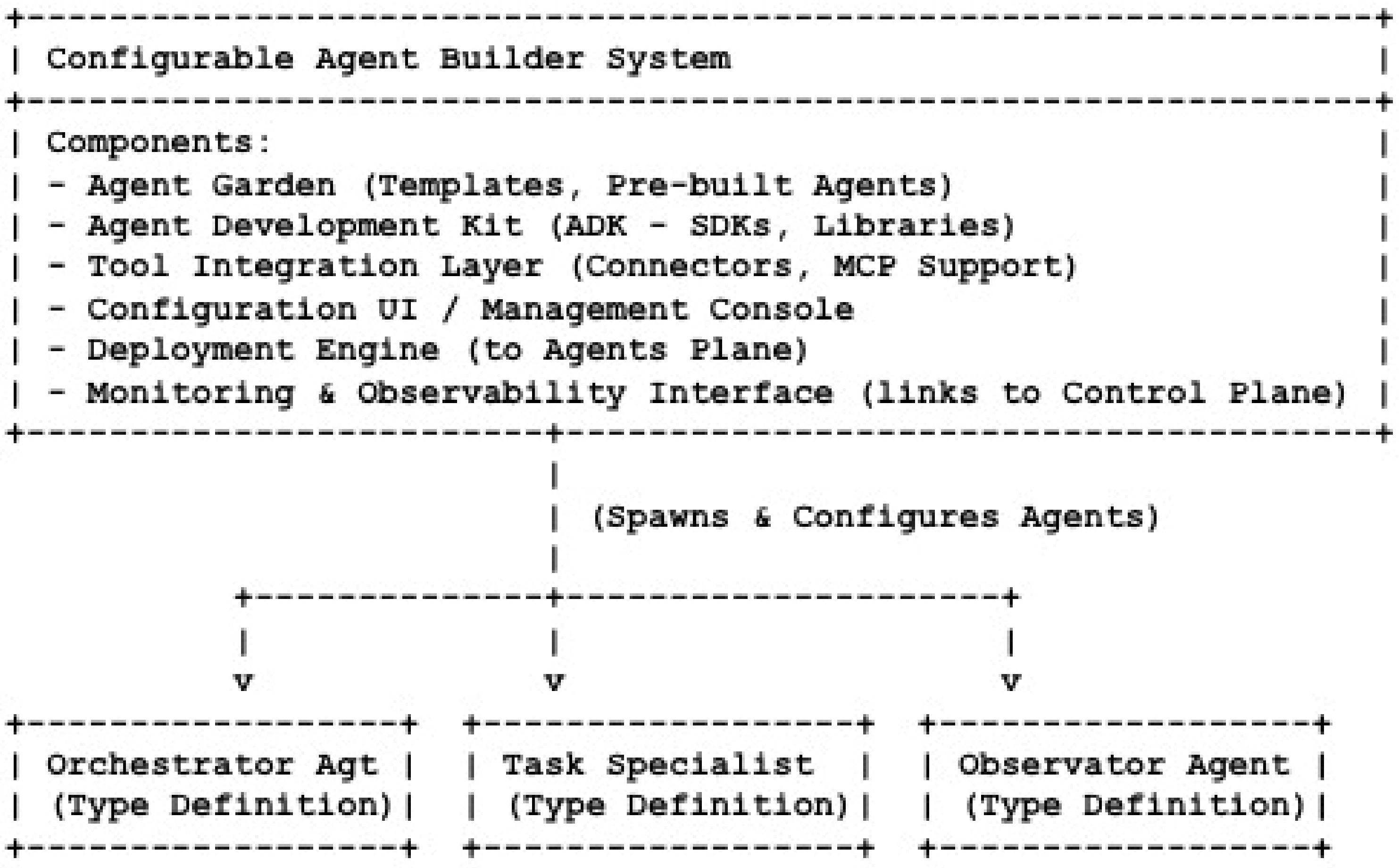
Agentic AI

User Plane - Control Plane - Agents Plane



Agentic AI

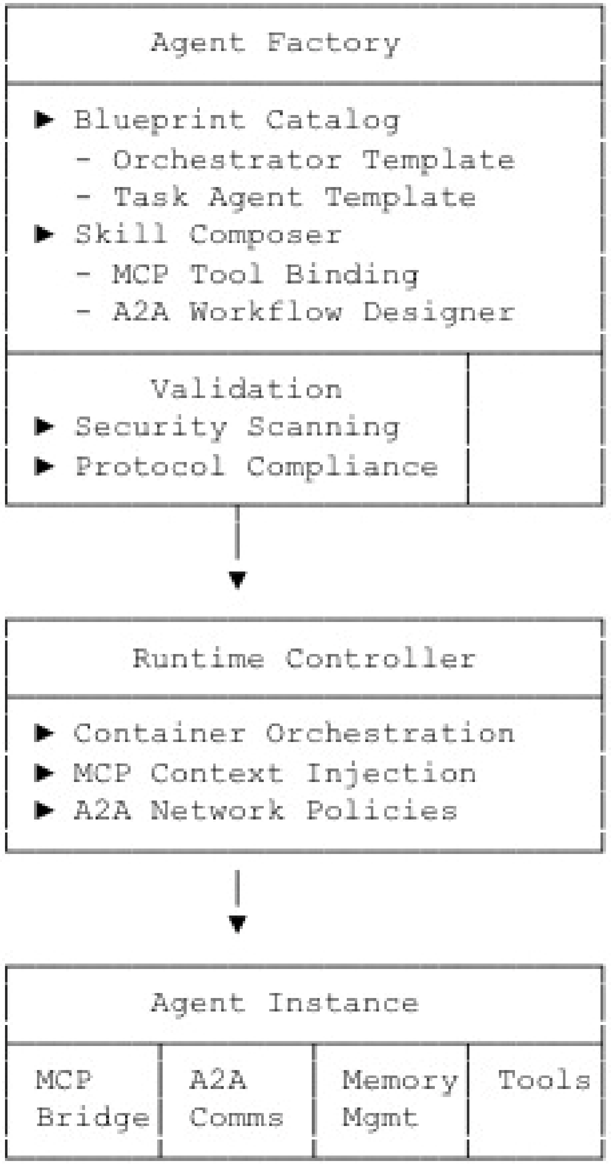
Conceive - Craft - Customize - Control



```
+-----------------------------------------------------------------+
| Configurable Agent Builder System                               |
+-----------------------------------------------------------------+
| Components:                                                     |
| - Agent Garden (Templates, Pre-built Agents)                    |
| - Agent Development Kit (ADK - SDKs, Libraries)                 |
| - Tool Integration Layer (Connectors, MCP Support)              |
| - Configuration UI / Management Console                         |
| - Deployment Engine (to Agents Plane)                           |
| - Monitoring & Observability Interface (links to Control Plane) |
+-------------------------+---------------------------------------+
                          |
                          | (Spawns & Configures Agents)
                          |
           +--------------+---------------------+
           |              |                     |
           v              v                     v
+------------------+  +------------------+  +------------------+
| Orchestrator Agt |  | Task Specialist  |  | Observator Agent |
| (Type Definition)|  | (Type Definition)|  | (Type Definition)|
+------------------+  +------------------+  +------------------+
```

Agentic AI

Agent Factory



Agent Factory

- ▶ Blueprint Catalog
  - Orchestrator Template
  - Task Agent Template
- ▶ Skill Composer
  - MCP Tool Binding
  - A2A Workflow Designer

Validation

- ▶ Security Scanning
- ▶ Protocol Compliance

▼

Runtime Controller

- ▶ Container Orchestration
- ▶ MCP Context Injection
- ▶ A2A Network Policies

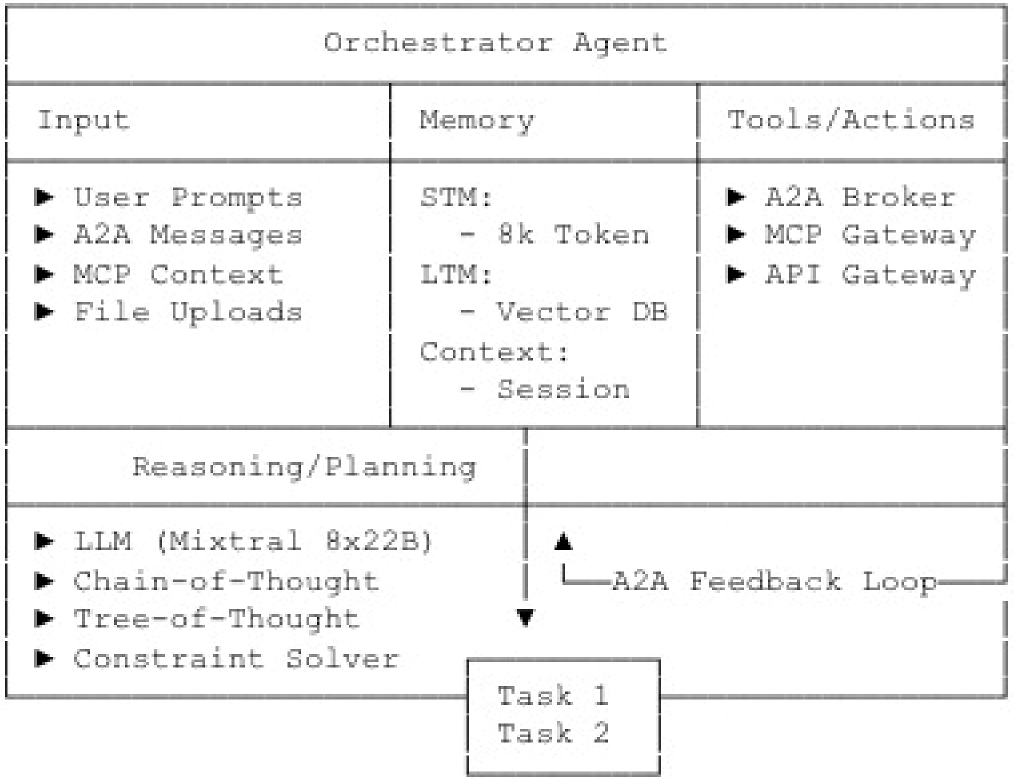
▼

Agent Instance

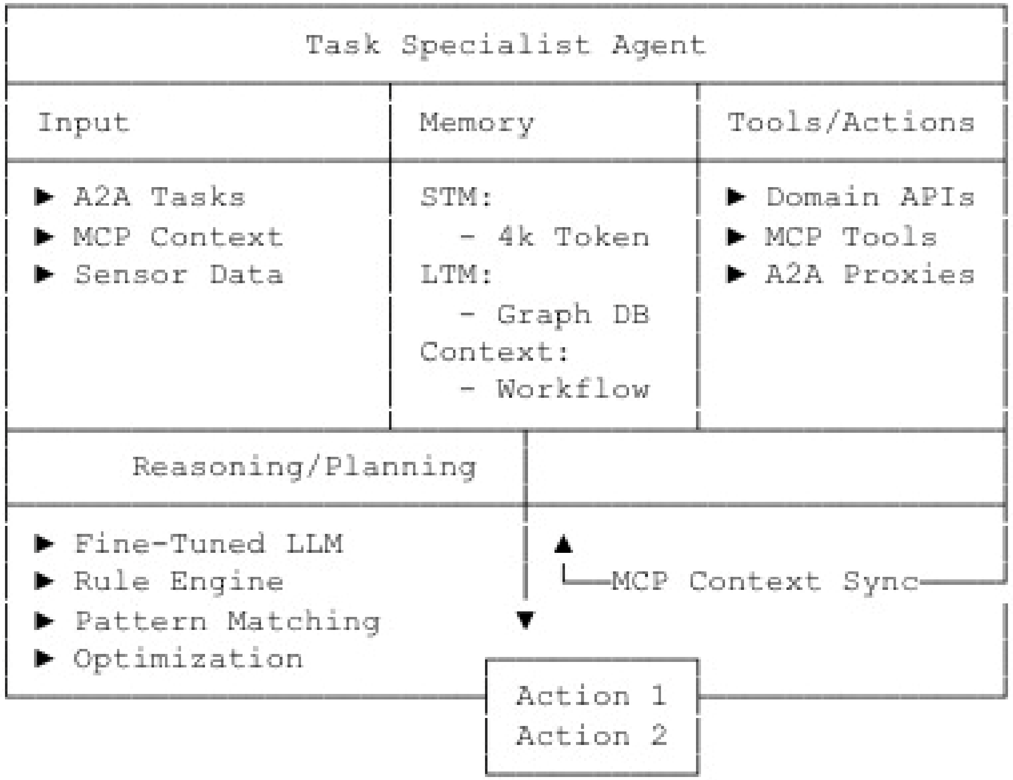
MCP Bridge	A2A Comms	Memory Mgmt	Tools

Agentic AI

Orchestrator Agents + Task Specialist Agents



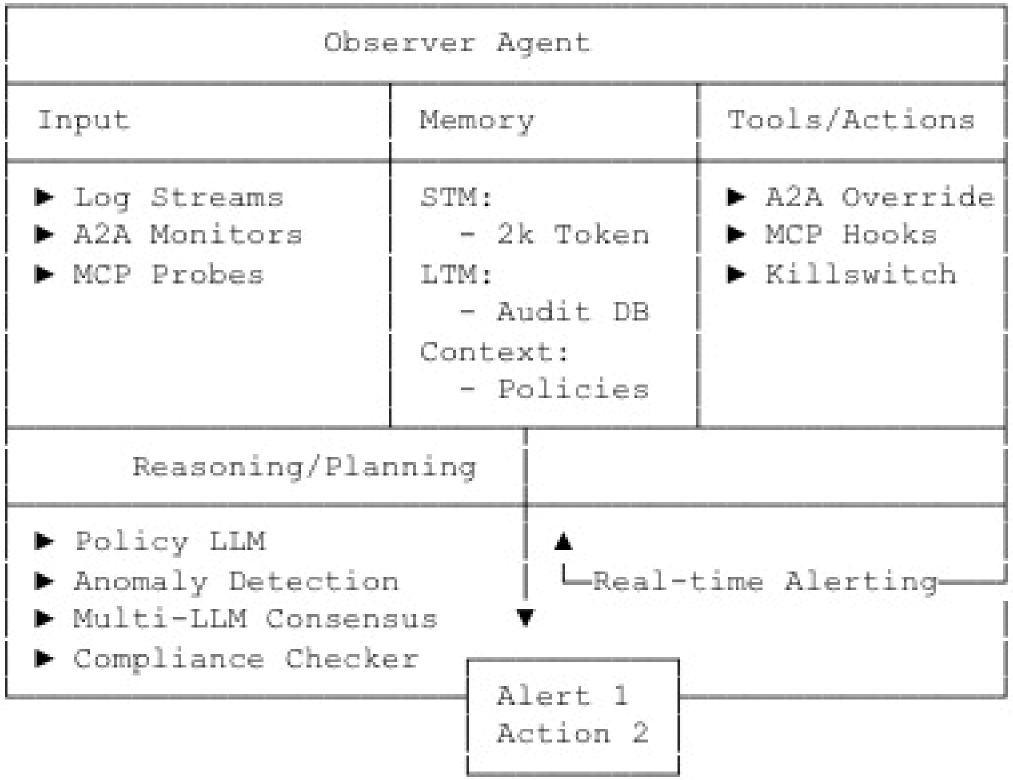
```
Orchestrator Agent

Input                  Memory             Tools/Actions

▶ User Prompts         STM:               ▶ A2A Broker
▶ A2A Messages           - 8k Token       ▶ MCP Gateway
▶ MCP Context          LTM:               ▶ API Gateway
▶ File Uploads           - Vector DB
                       Context:
                         - Session

      Reasoning/Planning

▶ LLM (Mixtral 8x22B)           ▲
▶ Chain-of-Thought              └─A2A Feedback Loop─
▶ Tree-of-Thought            ▼
▶ Constraint Solver
                          Task 1
                          Task 2
```



```
Task Specialist Agent

Input                  Memory             Tools/Actions

▶ A2A Tasks            STM:               ▶ Domain APIs
▶ MCP Context            - 4k Token       ▶ MCP Tools
▶ Sensor Data          LTM:               ▶ A2A Proxies
                         - Graph DB
                       Context:
                         - Workflow

      Reasoning/Planning

▶ Fine-Tuned LLM                ▲
▶ Rule Engine                   └─MCP Context Sync─
▶ Pattern Matching           ▼
▶ Optimization
                          Action 1
                          Action 2
```

Agentic AI

Observer Agents



Observer Agent

Input

- ▶ Log Streams
- ▶ A2A Monitors
- ▶ MCP Probes

Memory

STM:
  - 2k Token
LTM:
  - Audit DB
Context:
  - Policies

Tools/Actions

- ▶ A2A Override
- ▶ MCP Hooks
- ▶ Killswitch

Reasoning/Planning

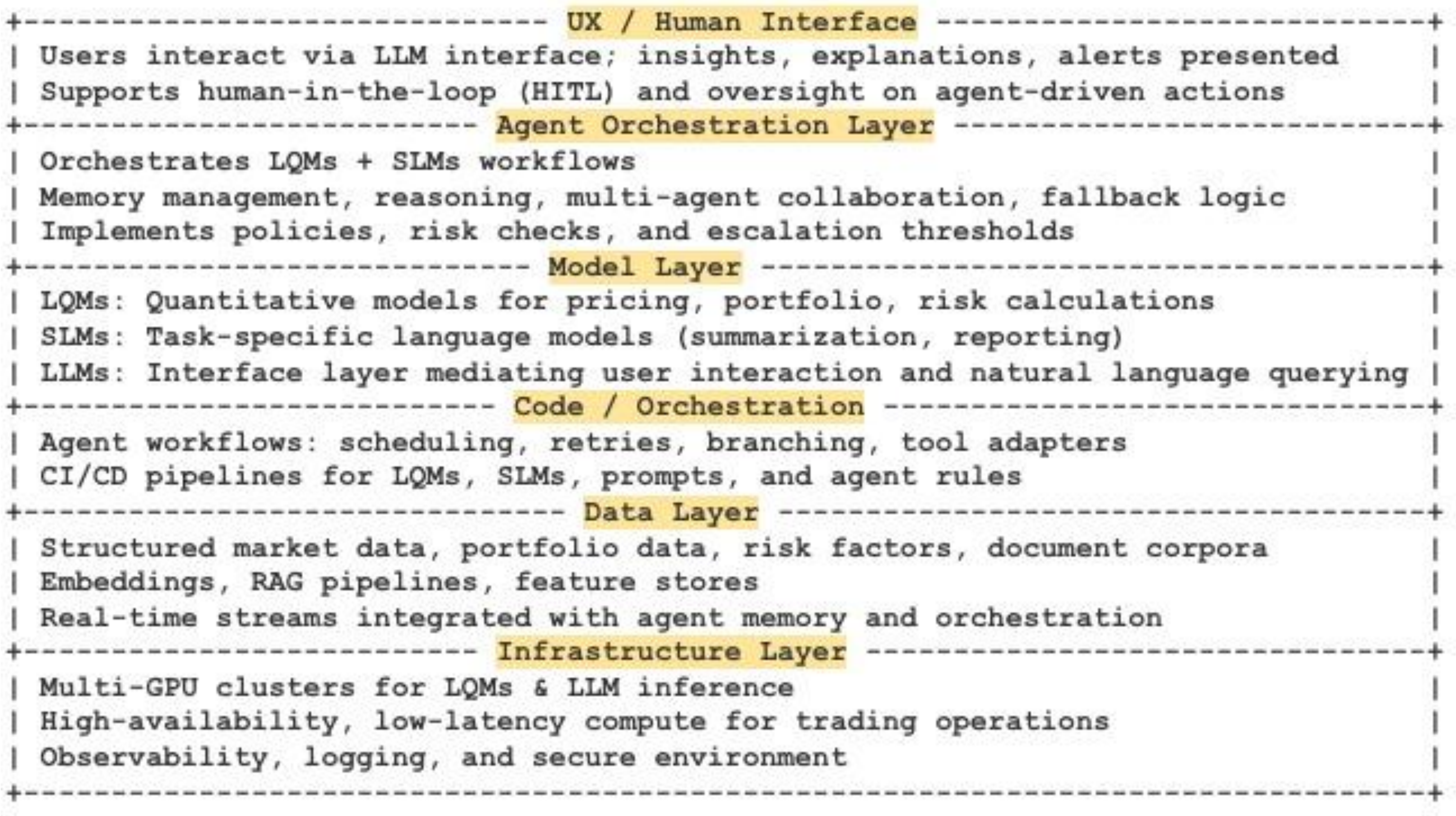
- ▶ Policy LLM
- ▶ Anomaly Detection
- ▶ Multi-LLM Consensus
- ▶ Compliance Checker

Real-time Alerting

Alert 1
Action 2

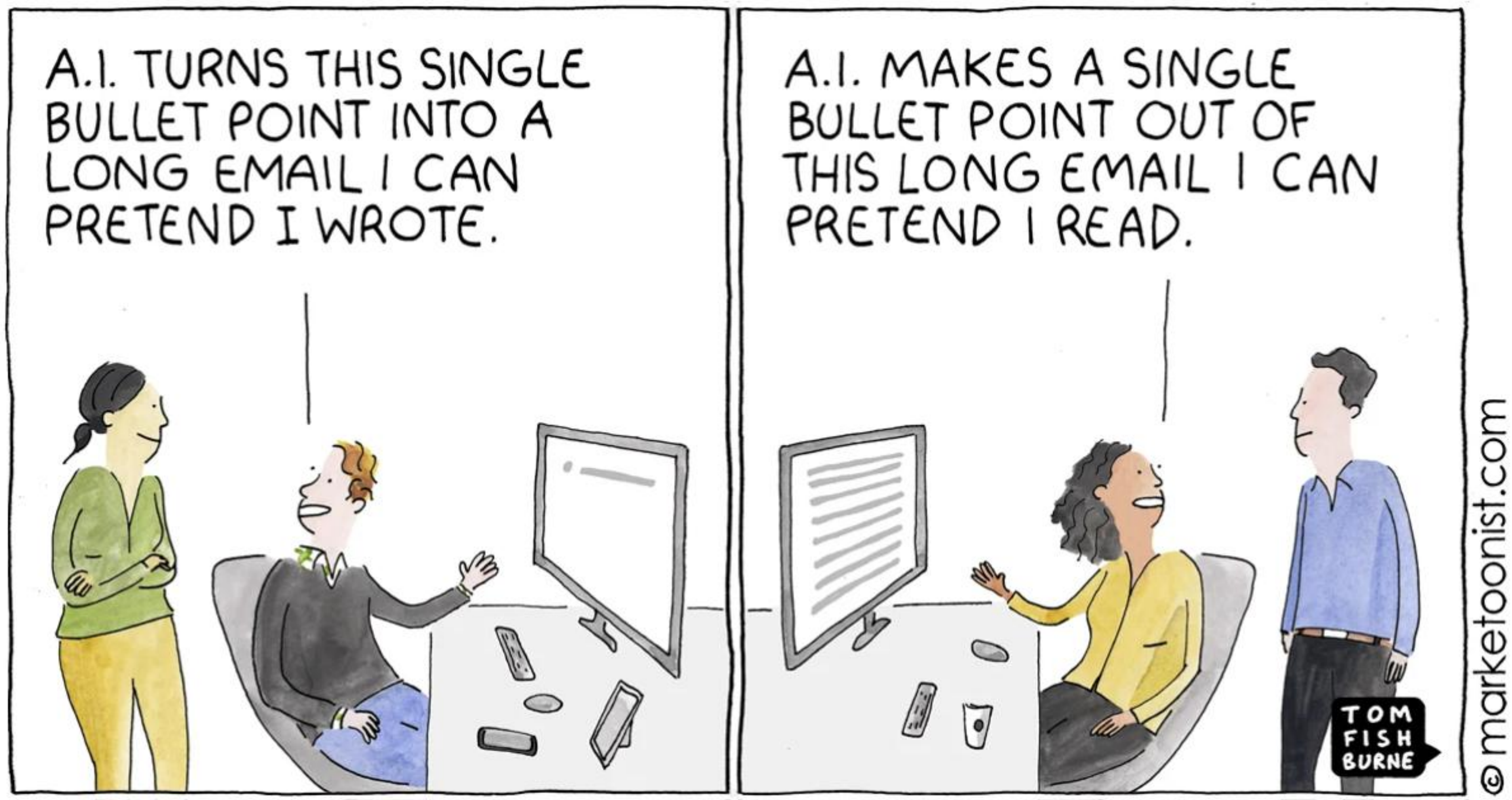
Agentic AI - From Infrastructure to Agents

Scaling for Business Value



```
+------------------------------ UX / Human Interface -------------------------------+
| Users interact via LLM interface; insights, explanations, alerts presented        |
| Supports human-in-the-loop (HITL) and oversight on agent-driven actions          |
+--------------------------- Agent Orchestration Layer ----------------------------+
| Orchestrates LQMs + SLMs workflows                                                |
| Memory management, reasoning, multi-agent collaboration, fallback logic          |
| Implements policies, risk checks, and escalation thresholds                      |
+------------------------------ Model Layer ----------------------------------------+
| LQMs: Quantitative models for pricing, portfolio, risk calculations              |
| SLMs: Task-specific language models (summarization, reporting)                   |
| LLMs: Interface layer mediating user interaction and natural language querying  |
+---------------------------- Code / Orchestration ---------------------------------+
| Agent workflows: scheduling, retries, branching, tool adapters                   |
| CI/CD pipelines for LQMs, SLMs, prompts, and agent rules                         |
+------------------------------- Data Layer ---------------------------------------+
| Structured market data, portfolio data, risk factors, document corpora           |
| Embeddings, RAG pipelines, feature stores                                        |
| Real-time streams integrated with agent memory and orchestration                 |
+--------------------------- Infrastructure Layer ---------------------------------+
| Multi-GPU clusters for LQMs & LLM inference                                       |
| High-availability, low-latency compute for trading operations                    |
| Observability, logging, and secure environment                                   |
+-----------------------------------------------------------------------------------+
```

AI


A.I. TURNS THIS SINGLE BULLET POINT INTO A LONG EMAIL I CAN PRETEND I WROTE.
A.I. MAKES A SINGLE BULLET POINT OUT OF THIS LONG EMAIL I CAN PRETEND I READ.
TOM FISH BURNE
©marketoonist.com

Thank You

AI powered Business

AI/ML Ops
Framework

USER EXPERIENCE /HUMAN-AGENT INTERFACE

Key Features	Observability	Governance
- UX Dashboards (Real-time Metrics, Performance) - Human-in-the-loop (HITL) Interfaces - Agent-to-Agent (A2A) Communication - Agent-to-Human (A2H) Interaction (Chatbots, Assistants)	- User Behavior Analytics - Model Feedback Tracking - UI/UX Monitoring	- Ethical Usage Guidelines - Privacy Management for Users - Human-in-the-loop Audit Trails

AGENT ORCHESTRATION LAYER

Key Features	Observability	Governance
- Orchestrates LQMs + SLMs workflows - Memory management, reasoning, multi-agent collaboration - Fallback logic, Implements policies - Risk checks, and escalation thresholds - MCP / A2A / AGUI protocols	- Agent Behavior Analytics - Agent Reliabilty Tracking - A2A - Agent 2 Human Monitoring - Tool Calling MCP - Token consumption	- Agent Risk and Trust Guardrails - Explainability Audit Trails - Accountability Audit Trails - Kill Switch Activation

MODEL LAYER

Key Features	Observability	Governance
- Model Registry (Versioning, Metadata) - Foundation Model Garden (Pre-trained models) - A/B Testing (Model Evaluation) - Multi-Agent Coordination (Task Delegation, Temporal Workflows)	- Model Performance Metrics - Latency Monitoring - Model Stability	- Responsible AI (Ethics, Transparency - Explainability (Black-box Models) - Accountability for Model Behavior

CODE LAYER

Key Features	Observability	Governance
- Infrastructure as Code (IaC) - ETL Pipelines - CI/CD (Automation of ML Pipeline) - Orchestration (Apache Airflow, Temporal.io, Kubernetes) - API Services - API Gateway (Model, Agent Integration)	- Code Quality Monitoring - CI/CD Pipeline Monitoring - Error Tracking	- Ethical Usage Guidelines - Privacy Management for Users - Human-in-the-loop Audit Trails

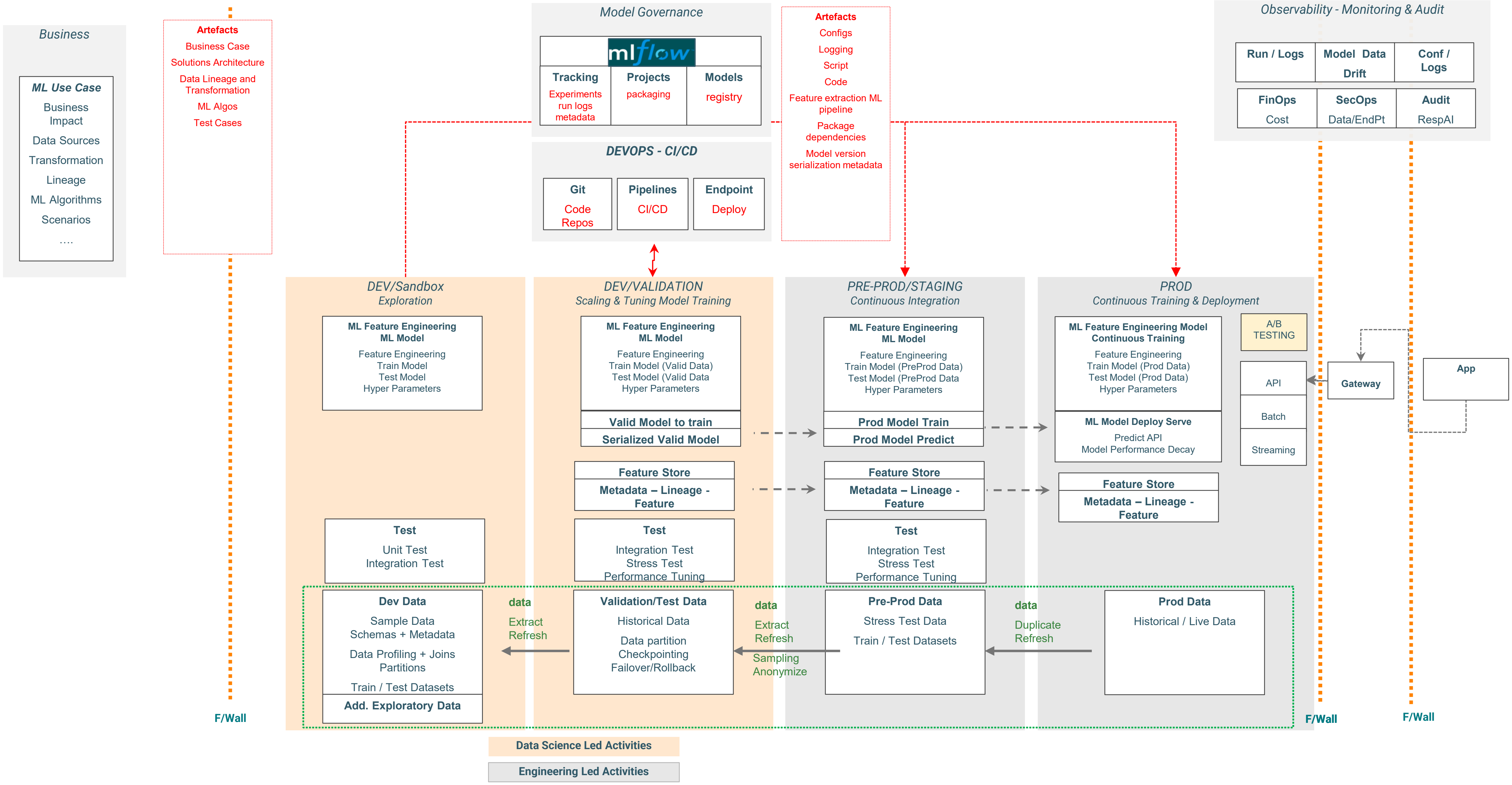
DATA LAYER

Key Features	Observability	Governance
- Data Lakehouse (Structured/Unstructured Data) - Feature Store (Real-time & Batch Features) - VectorDB (Embeddings, RAG) - Data Pipelines (ETL, Preprocessing)	- Data Quality Monitoring - Drift Detection - Anomaly Detection	- Data Privacy and Security - Model Fairness & Bias Checking - Access Control (Data Permissions)

INFRASTRUCTURE LAYER

Key Features	Observability	Governance
- CPU/GPU Resources (DEV/TEST/PRE-PROD/PROD) - Cloud/On-prem Infrastructure (Elastic Compute, Kubernetes, GPU Provisioning)	- Performance Monitoring - Cost Tracking - Security (Access, Compliance) - SLA Management	- Data Lineage - Model Lineage - Responsible AI (Fairness, Bias) - Guardrails (Model and Data Validation)

MLOps - Road to Production



Business

ML Use Case

Business Impact

Data Sources

Transformation

Lineage

ML Algorithms

Scenarios

….

Artefacts

Business Case

Solutions Architecture

Data Lineage and Transformation

ML Algos

Test Cases

F/Wall

Model Governance

mlflow

Tracking	Projects	Models
Experiments run logs metadata	packaging	registry

DEVOPS - CI/CD

Git	Pipelines	Endpoint
Code Repos	CI/CD	Deploy

Artefacts

Configs

Logging

Script

Code

Feature extraction ML pipeline

Package dependencies

Model version serialization metadata

Observability - Monitoring & Audit

Run / Logs	Model Data Drift	Conf / Logs

FinOps	SecOps	Audit
Cost	Data/EndPt	RespAI

DEV/Sandbox
Exploration

ML Feature Engineering
ML Model

Feature Engineering
Train Model
Test Model
Hyper Parameters

Test

Unit Test
Integration Test

Dev Data

Sample Data
Schemas + Metadata

Data Profiling + Joins
Partitions

Train / Test Datasets

Add. Exploratory Data

data

Extract
Refresh

DEV/VALIDATION
Scaling & Tuning Model Training

ML Feature Engineering
ML Model

Feature Engineering
Train Model (Valid Data)
Test Model (Valid Data
Hyper Parameters

Valid Model to train

Serialized Valid Model

Feature Store

Metadata – Lineage - Feature

Test

Integration Test
Stress Test
Performance Tuning

Validation/Test Data

Historical Data

Data partition
Checkpointing
Failover/Rollback

data

Extract
Refresh

Sampling
Anonymize

PRE-PROD/STAGING
Continuous Integration

ML Feature Engineering
ML Model

Feature Engineering
Train Model (PreProd Data)
Test Model (PreProd Data
Hyper Parameters

Prod Model Train

Prod Model Predict

Feature Store

Metadata – Lineage - Feature

Test

Integration Test
Stress Test
Performance Tuning

Pre-Prod Data

Stress Test Data

Train / Test Datasets

data

Duplicate
Refresh

PROD
Continuous Training & Deployment

ML Feature Engineering Model
Continuous Training

Feature Engineering
Train Model (Prod Data)
Test Model (Prod Data)
Hyper Parameters

ML Model Deploy Serve

Predict API
Model Performance Decay

Feature Store

Metadata – Lineage - Feature

Prod Data

Historical / Live Data

A/B TESTING

API

Batch

Streaming

Gateway

App

F/Wall

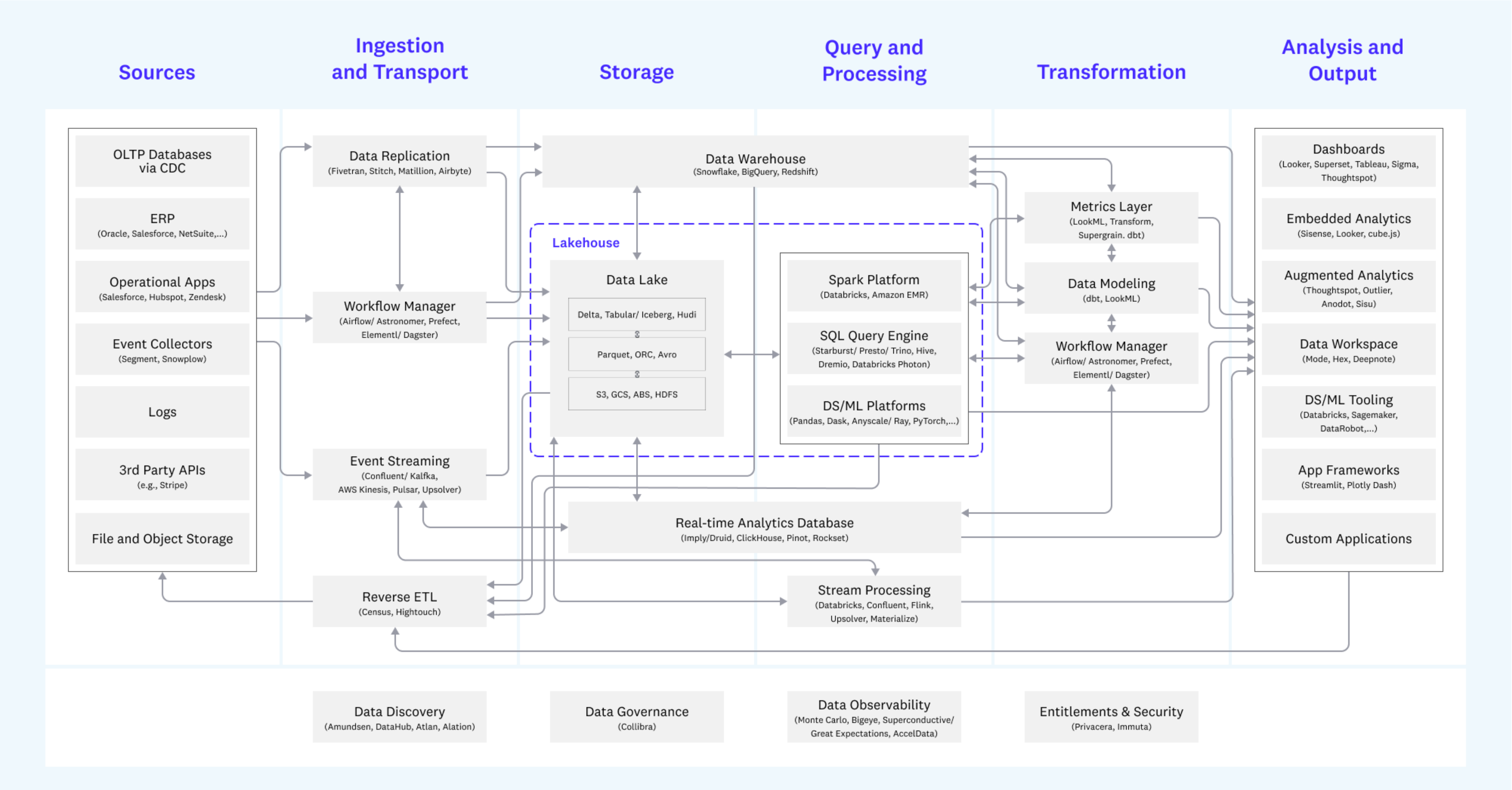
F/Wall

Data Science Led Activities

Engineering Led Activities

Data Platform

Orchestration - Catalogue - Observability - Security



Sources

- OLTP Databases via CDC
- ERP (Oracle, Salesforce, NetSuite,...)
- Operational Apps (Salesforce, Hubspot, Zendesk)
- Event Collectors (Segment, Snowplow)
- Logs
- 3rd Party APIs (e.g., Stripe)
- File and Object Storage

Ingestion and Transport

- Data Replication (Fivetran, Stitch, Matillion, Airbyte)
- Workflow Manager (Airflow/ Astronomer, Prefect, Elementl/ Dagster)
- Event Streaming (Confluent/ Kalfka, AWS Kinesis, Pulsar, Upsolver)
- Reverse ETL (Census, Hightouch)

Storage

- Data Warehouse (Snowflake, BigQuery, Redshift)
- Lakehouse
- Data Lake
  - Delta, Tabular/ Iceberg, Hudi
  - Parquet, ORC, Avro
  - S3, GCS, ABS, HDFS
- Real-time Analytics Database (Imply/Druid, ClickHouse, Pinot, Rockset)

Query and Processing

- Spark Platform (Databricks, Amazon EMR)
- SQL Query Engine (Starburst/ Presto/ Trino, Hive, Dremio, Databricks Photon)
- DS/ML Platforms (Pandas, Dask, Anyscale/ Ray, PyTorch,...)
- Stream Processing (Databricks, Confluent, Flink, Upsolver, Materialize)

Transformation

- Metrics Layer (LookML, Transform, Supergrain. dbt)
- Data Modeling (dbt, LookML)
- Workflow Manager (Airflow/ Astronomer, Prefect, Elementl/ Dagster)

Analysis and Output

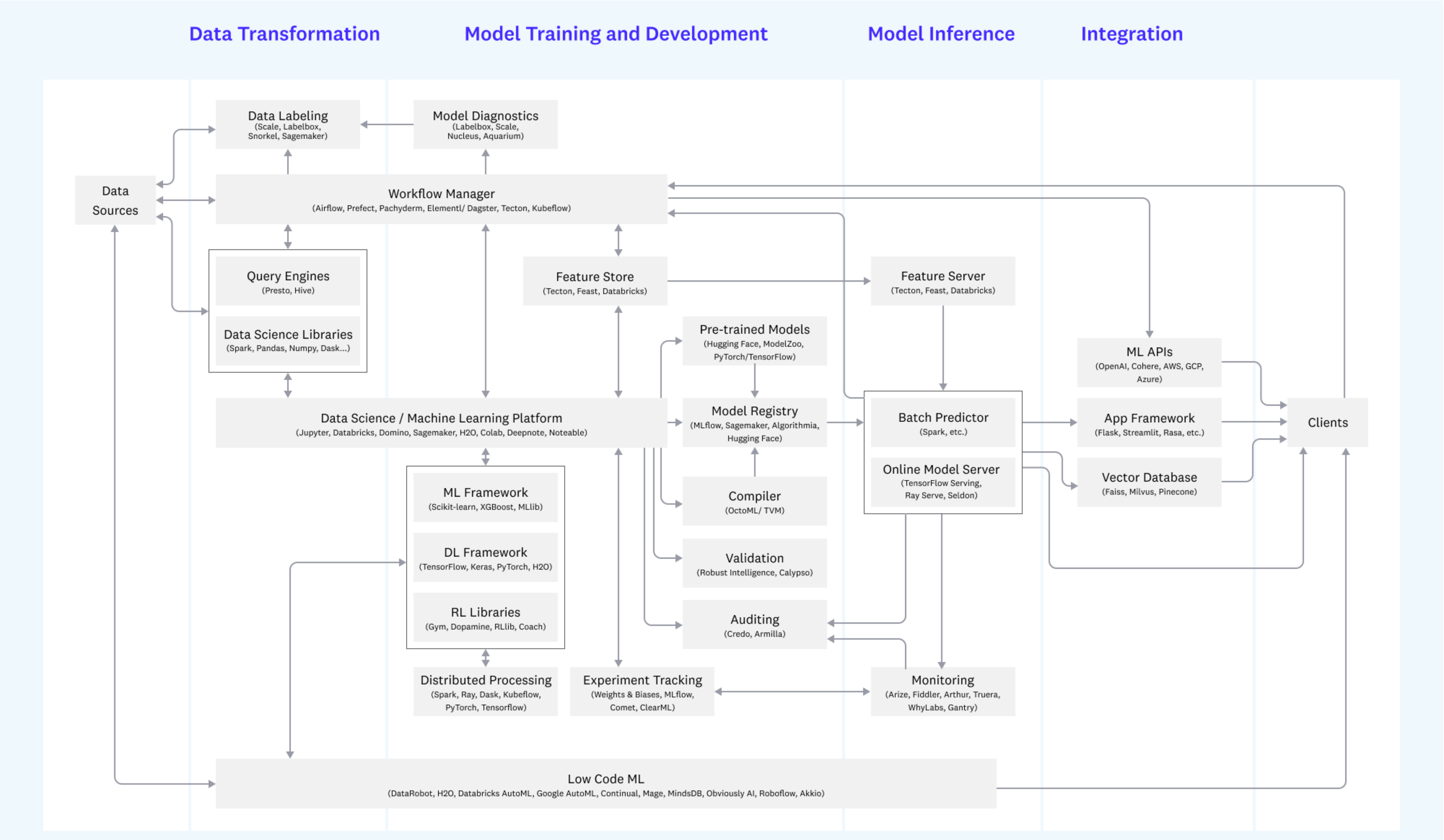
- Dashboards (Looker, Superset, Tableau, Sigma, Thoughtspot)
- Embedded Analytics (Sisense, Looker, cube.js)
- Augmented Analytics (Thoughtspot, Outlier, Anodot, Sisu)
- Data Workspace (Mode, Hex, Deepnote)
- DS/ML Tooling (Databricks, Sagemaker, DataRobot,...)
- App Frameworks (Streamlit, Plotly Dash)
- Custom Applications

- Data Discovery (Amundsen, DataHub, Atlan, Alation)
- Data Governance (Collibra)
- Data Observability (Monte Carlo, Bigeye, Superconductive/ Great Expectations, AccelData)
- Entitlements & Security (Privacera, Immuta)

https://a16z.com/emerging-architectures-for-modern-data-infrastructure/

ML/AI Platform

Orchestration - Model & Data Drift - Audit-ability



Data Transformation

Model Training and Development

Model Inference

Integration

Data Labeling (Scale, Labelbox, Snorkel, Sagemaker)

Model Diagnostics (Labelbox, Scale, Nucleus, Aquarium)

Data Sources

Workflow Manager (Airflow, Prefect, Pachyderm, Elementl/ Dagster, Tecton, Kubeflow)

Query Engines (Presto, Hive)

Data Science Libraries (Spark, Pandas, Numpy, Dask...)

Feature Store (Tecton, Feast, Databricks)

Feature Server (Tecton, Feast, Databricks)

Pre-trained Models (Hugging Face, ModelZoo, PyTorch/TensorFlow)

ML APIs (OpenAI, Cohere, AWS, GCP, Azure)

Data Science / Machine Learning Platform (Jupyter, Databricks, Domino, Sagemaker, H2O, Colab, Deepnote, Noteable)

Model Registry (MLflow, Sagemaker, Algorithmia, Hugging Face)

Batch Predictor (Spark, etc.)

Online Model Server (TensorFlow Serving, Ray Serve, Seldon)

App Framework (Flask, Streamlit, Rasa, etc.)

Vector Database (Faiss, Milvus, Pinecone)

Clients

ML Framework (Scikit-learn, XGBoost, MLlib)

DL Framework (TensorFlow, Keras, PyTorch, H2O)

RL Libraries (Gym, Dopamine, RLlib, Coach)

Compiler (OctoML/ TVM)

Validation (Robust Intelligence, Calypso)

Auditing (Credo, Armilla)

Distributed Processing (Spark, Ray, Dask, Kubeflow, PyTorch, Tensorflow)

Experiment Tracking (Weights & Biases, MLflow, Comet, ClearML)

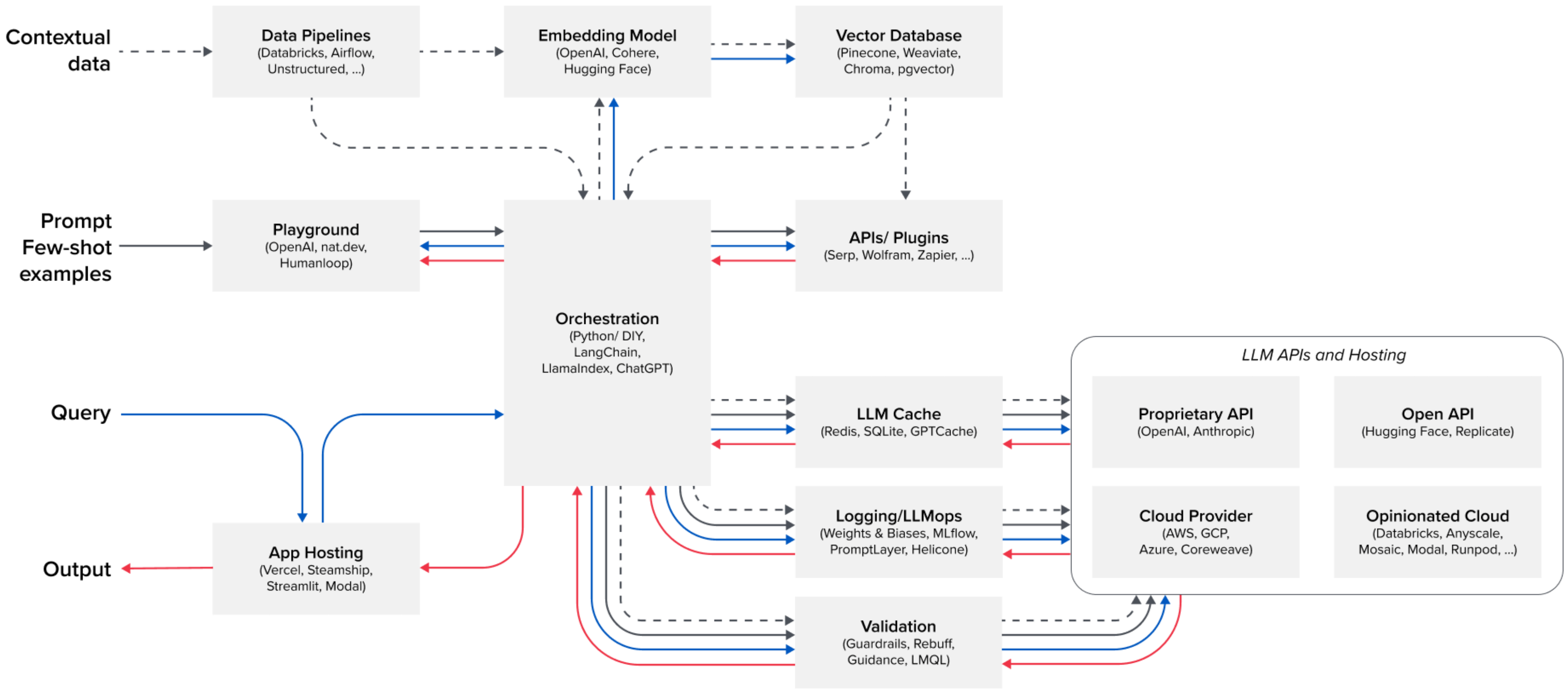
Monitoring (Arize, Fiddler, Arthur, Truera, WhyLabs, Gantry)

Low Code ML (DataRobot, H2O, Databricks AutoML, Google AutoML, Continual, Mage, MindsDB, Obviously AI, Roboflow, Akkio)

https://a16z.com/emerging-architectures-for-modern-data-infrastructure/

LLM Stack

Foundation Models - Prompt - Fine Tune - RAG



Contextual data

Data Pipelines
(Databricks, Airflow, Unstructured, ...)

Embedding Model
(OpenAI, Cohere, Hugging Face)

Vector Database
(Pinecone, Weaviate, Chroma, pgvector)

Prompt
Few-shot examples

Playground
(OpenAI, nat.dev, Humanloop)

Orchestration
(Python/ DIY, LangChain, LlamaIndex, ChatGPT)

APIs/ Plugins
(Serp, Wolfram, Zapier, ...)

Query

Output

App Hosting
(Vercel, Steamship, Streamlit, Modal)

LLM Cache
(Redis, SQLite, GPTCache)

Logging/LLMops
(Weights & Biases, MLflow, PromptLayer, Helicone)

Validation
(Guardrails, Rebuff, Guidance, LMQL)

LLM APIs and Hosting

Proprietary API
(OpenAI, Anthropic)

Open API
(Hugging Face, Replicate)

Cloud Provider
(AWS, GCP, Azure, Coreweave)

Opinionated Cloud
(Databricks, Anyscale, Mosaic, Modal, Runpod, ...)

LEGEND

Gray boxes show key components of the stack, with leading tools/systems listed

Arrows show the flow of data through the stack

- - - → Contextual data provided by app developers to condition LLM outputs

⟶ Prompts and few-shot examples that are sent to the LLM

⟶ Queries submitted by users

⟶ Output returned to users

https://a16z.com/emerging-architectures-for-llm-applications/