AI INSIDER
powered by RESONANCE

# 7-layer AI Stack framework & AI Market Maps

April 2025

# Executive summary

The AI ecosystem is becoming increasingly complex, making it more challenging for organizations to navigate how systems are built, scaled, and governed. To bring structure to this landscape, ***AI Insider*** **has developed a 7-layer AI Stack framework** that clarifies the interdependencies between infrastructure, models, workflows, and governance.

Across these seven layers — Hardware & Datacenter Enablers, Data Structure & Processing, Model Development & Deployment, Inference, Orchestration, Tooling, and Security & Governance — **we have mapped over 450 solutions to provide a detailed and comprehensive view of the AI market.**

The framework is intended as a baseline for evaluating AI strategies, priorities and understand the AI ecosystem from its foundational infrastructure to services and applications, building a structured reference for understanding how the market is evolving. **Organizations adopting a full-stack perspective may be better positioned to manage complexity, optimize resources, and align AI initiatives with broader enterprise objectives.**

This deck presents a curated selection of the full mapping for illustrative purposes. To access the complete maps or discuss the framework in more detail, please contact us at ***hello@resonance.holdings***.

# The 7 Layers of the AI Stack
## A strategic framework for market understanding

6 **Tooling**

The top layer empowers builders. It includes developer environments, agent builders, app creation tools, and MLOps — making it easier to prototype, ship, and scale AI-powered applications.

5 **Orchestration**

AI systems often use multiple models and tools in sequence. This layer decides how components interact — routing prompts, managing agent behavior, and evaluating outputs across workflows.

4 **Inference**

Once models are live, this layer handles real-time execution. Inference systems, APIs, and microservices ensure models can respond at scale, in milliseconds, to user inputs or system queries.

3 **Model development & deployment**

With compute and structured data ready, this is where models are trained, fine-tuned, and deployed. It includes foundation models, deep learning frameworks, and interfaces to production environments.

2 **Data Structure & Processing**

Raw data must be cleaned, normalized, and organized before AI can learn. This layer structures data into usable formats and enriches it with embeddings and vectorization for learning and retrieval.

1 **Hardware & AI Datacentre Enablers**

The stack begins with infrastructure. This layer provides the computational foundation — from GPUs and AI chips to power, cooling, and storage. Without it, no training or inference is possible.

7 **Security & Governance**

A horizontal layer that secures the entire stack. It spans risk mitigation (e.g., LLM red teaming, adversarial testing), data governance (e.g., masking, consent, lineage), privacy, and compliance. It ensures trust, auditability, and safety across the AI infrastructure.

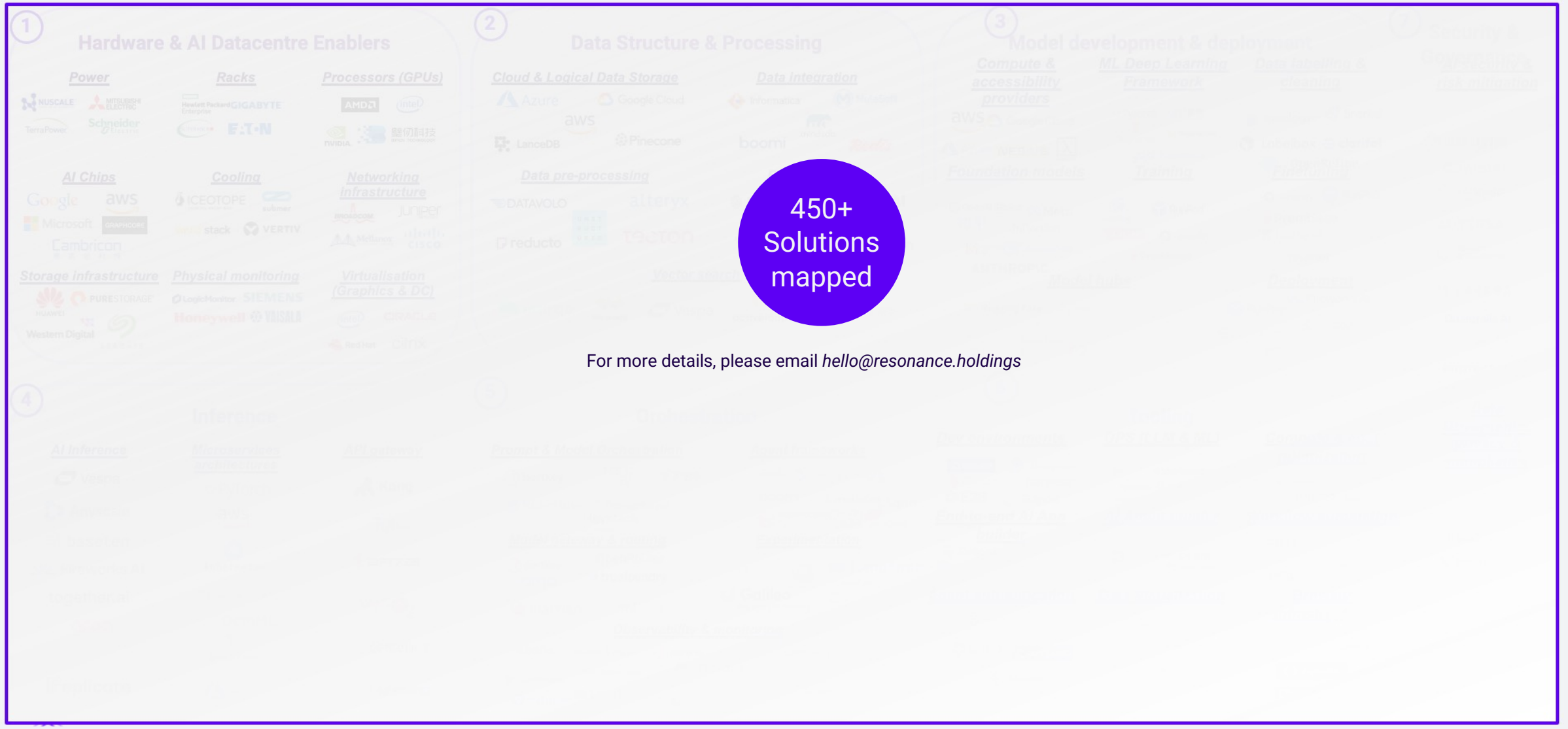
# The 7 Layers of the AI Stack

## 41 building blocks to map the AI Stack across the 7 layers

# Market Map: Core AI Stack

*April 2025 – A work in progress*

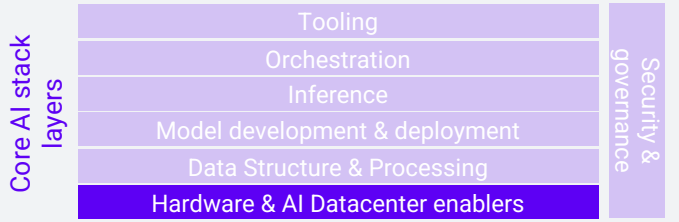
*Selected vendors, non exhaustive*




For more details, please email *hello@resonance.holdings*

# Market Map: Hardware & AI Datacenter enablers

*Workflow*




### Power

1. The flow begins with Power, which delivers a continuous and stable energy supply to the entire AI infrastructure. This includes redundant power distribution systems, battery backups, and (in cutting-edge cases) small modular nuclear reactors. Power is critical, because even brief interruptions can crash long training runs or interrupt 24/7 inference systems.

### Racks

2. With power in place, Racks physically house the hardware — such as GPUs, AI chips, and storage systems — in dense, modular configurations optimized for airflow and cable management. These racks form the scalable physical core of compute clusters.

### Processors (GPUs)

3. Inside the racks, GPUs and AI chips handle the heavy lifting. GPUs specialize in parallel matrix operations critical for deep learning, unlike general-purpose CPUs. Their massive parallelism accelerates AI tasks: training, fine-tuning, and inference.

### AI Chips

4. AI chips may be generalist (like TPUs) or specialist (e.g., vision-only or edge inference chips) and provide higher efficiency, speed, and energy savings compared to traditional GPUs in targeted AI scenarios. They are especially useful for companies that need scalable, repeatable AI performance without the overhead of general-purpose compute.

### Cooling

5. These processors generate a massive amount of heat, so Cooling systems are used to stabilize thermal conditions. Whether air-based, liquid-cooled, or immersed, cooling keeps the cluster operational and prevents performance throttling or failure.

### Networking infrastructure

6. While computation happens, Networking Infrastructure connects all nodes across the datacentre — enabling distributed training, data movement from storage, and real-time communication between inference servers. High-bandwidth, low-latency networking (e.g. InfiniBand, 400G Ethernet) is essential to scale.

### Storage infrastructure

7. Alongside compute, storage infrastructure keeps datasets, logs, checkpoints, and model outputs safely stored. This may include HDDs, SSDs, or flash arrays — all optimized for throughput and durability to support both training and inference workflows.

### Physical monitoring

8. Physical Monitoring tools continuously track temperature, power draw, uptime, and failures across the rack and server infrastructure. This ensures reliability and early problem detection, feeding telemetry into datacentre management systems.

### Virtualization (Graphics & DC)

9. Finally, Virtualisation (Graphics & Datacentre) abstracts the underlying compute into shareable, scalable resources. Through virtualization layers, GPU capacity is assigned to users and jobs, allowing teams to run workloads in isolated environments (e.g., vGPU, Docker, or K8s).

# Market Map: Hardware & AI Datacenter enablers

*April 2025 – A work in progress*

*Non exhaustive*

Core AI stack layers

- Tooling
- Orchestration
- Inference
- Model development & deployment
- Data Structure & Processing
- **Hardware & AI Datacenter enablers**

Security & governance

## Power

NUSCALE Power for all humankind · EnerSys Power/Full Solutions · NEXTera ENERGY · energy · ROLLS ROYCE · TESLA · OKLO · newcleo Futurable Energy · saft · DUKE ENERGY · TerraPower · EATON · Schneider Electric · ENERGY VAULT Enabling a Renewable World · ABB · ROSATOM · Kairos Power · MITSUBISHI ELECTRIC

## Racks

Hewlett Packard Enterprise · DELL · Schneider Electric · cisco · SUPERMICRO · Lenovo · GIGABYTE · IBM · NVIDIA · VERTIV · EATON · legrand · HUAWEI

## Processors (GPUs)

NVIDIA · AMD · intel · 摩尔线程 MOORE THREADS · Imagination · arm · Qualcomm · 景嘉微 JINGJIA MICRO · 壁仞科技 BIREN TECHNOLOGY · SAMSUNG ELECTRONICS

## AI Chips

Google · aws · Microsoft · Alibaba · NVIDIA · AMD · cerebras · GRAPHCORE · tenstorrent · Qualcomm · UNTETHER AI · neuchips · Cambricon 寒武纪科技 · FURIOSA · intel · Enflame 燧原科技 · KALRAY THE POWER OF MORE · AXELERA ARTIFICIAL INTELLIGENCE

## Cooling

ICEOTOPE COOLING REDEFINED · GRC GREEN REVOLUTION COOLING The Immersion Cooling Authority · submer · DAIKIN · VERTIV · GIGABYTE

130+ Solutions mapped

## Networking infrastructure

AMD · NVIDIA · cisco · ARISTA · JUNIPER NETWORKS · Mellanox TECHNOLOGIES · BROADCOM

## Storage infrastructure

Western Digital · PURESTORAGE · Micron · PURE STORAGE · NetApp · Hewlett Packard Enterprise · SOLIDIGM · DELL · HUAWEI · cisco · SUPERMICRO · IBM · SEAGATE

## Physical monitoring

LogicMonitor · APC Legendary Reliability · ZABBIX · RoomAlert We've got this. · Honeywell · RFCODE · Sunbird DCIM that's easy, fast, and complete · ABB · VAISALA · SIEMENS

## Virtualization (Graphics & DC)

intel · vmware · ORACLE · NVIDIA · citrix · Red Hat · HUAWEI · cisco · SCALE COMPUTING · Virtuozzo · Hewlett Packard Enterprise

# Market Map: Data Structure & Processing

*Workflow*

Core AI stack layers

- Tooling
- Orchestration
- Inference
- Model development & deployment
- **Data Structure & Processing**
- Hardware & AI Datacenter enablers

Security & governance

In many environments, this logical storage layer is also connected to Storage Infrastructure (Layer 1) to enable caching or fast local access during runtime (e.g., training or inference). However, the core focus here is to provide a platform-agnostic, structured foundation for all downstream AI tasks — including pre-processing, embedding, retrieval, and labeling.

**Cloud & Logical Data Storage**

1. After the infrastructure is provisioned and compute is enabled at the hardware level (Layer 1), AI workflows begin by organizing and storing raw data inside Logical Data Storage systems. At this stage, files such as text corpora, image datasets, or structured exports are ingested into cloud object stores, structured data lakes, or dataset catalogs. These systems are abstracted from physical hardware and are accessed via APIs and pipelines — enabling reproducible, versioned, and collaborative data workflows.

**Data integration**

2. Because datasets often originate from fragmented systems or formats, the next step is Data Infra
Pipelines ingest content from databases
warehouses, or SaaS platforms
consistent schemas and
to ensure that all
coherent

For more details, please email *hello@resonance.holdings*

# Market Map: Data Structure & Processing

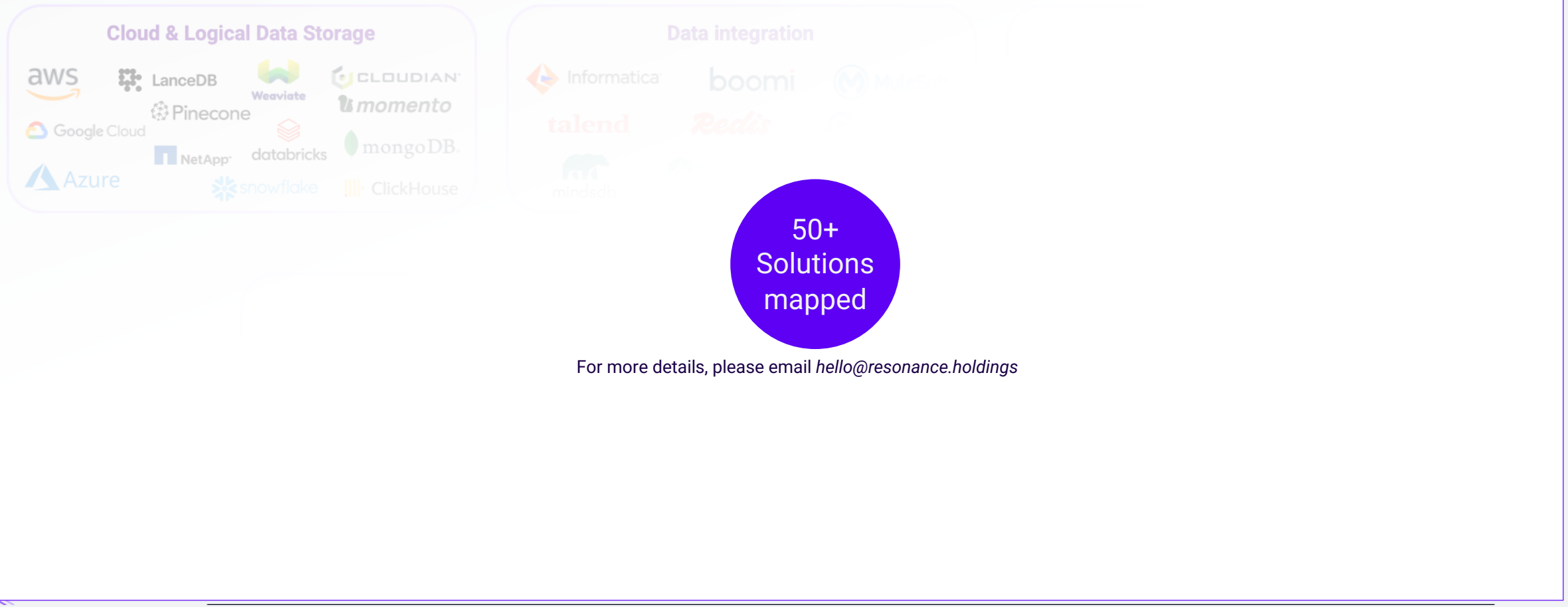
*April 2025 – A work in progress*

*Non exhaustive*

Core AI stack layers

- Tooling
- Orchestration
- Inference
- Model development & deployment
- **Data Structure & Processing**
- Hardware & AI Datacenter enablers

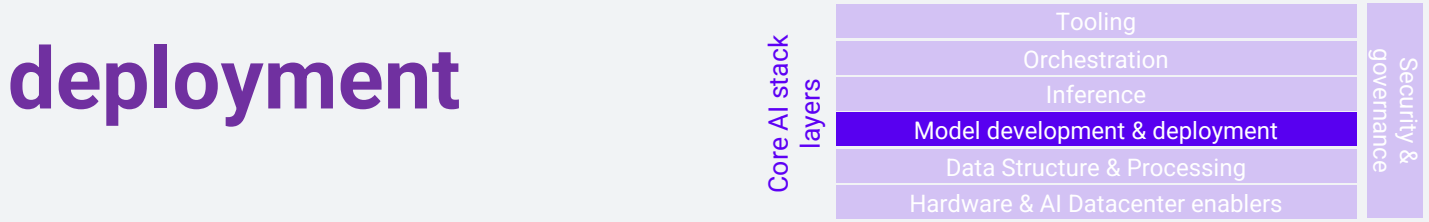
Security & governance




For more details, please email *hello@resonance.holdings*

# Market Map: Model dev. & deployment

*Workflow*




**Compute & accessibility providers**

1. With data structured and embeddings generated, the system calls on Compute & Accessibility Providers to provision infrastructure — GPUs, TPUs, or AI-optimized chips, either via the cloud or on-prem clusters. This is the raw horsepower required to run large-scale training and inference jobs.

**ML deep learning framework**

2. Once compute is available, developers use ML deep learning frameworks to define model architectures, loss functions, optimization routines, and training workflows. These frameworks are where the training logic is programmed and run.

**Data labelling**

3. Simultaneously, the Data [illegible]

**Foundation models**

4. With data and frameworks ready, teams may choose to start from scratch — or use a Foundation Model. These are large, pre-trained models like GPT, Claude, or LLaMA that already capture general knowledge. They're retrieved from [illegible] and serve as a strong base for

5. [illegible]

For more details, please email *hello@resonance.holdings*

# Market Map: Model dev. & deployment

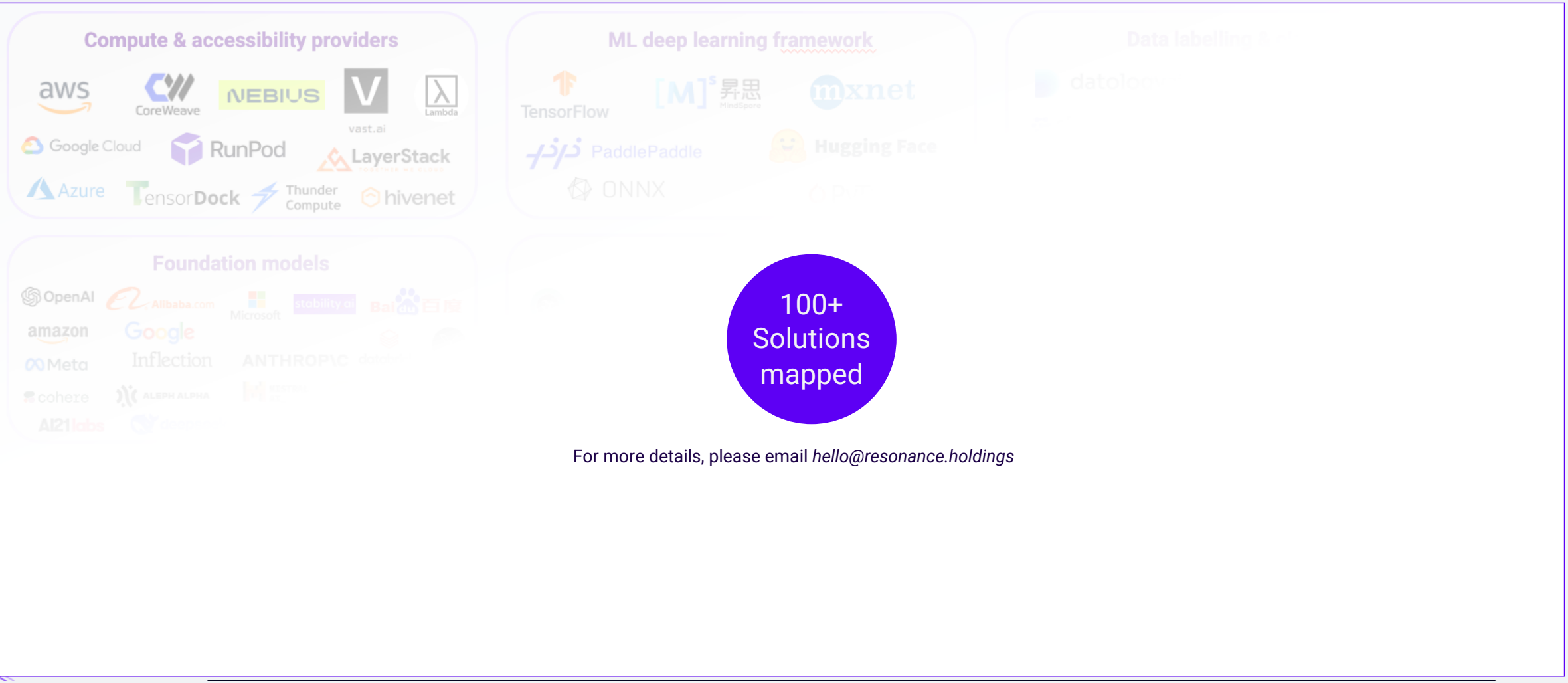
*April 2025 – A work in progress*

Core AI stack layers

- Tooling
- Orchestration
- Inference
- **Model development & deployment**
- Data Structure & Processing
- Hardware & AI Datacenter enablers

Security & governance

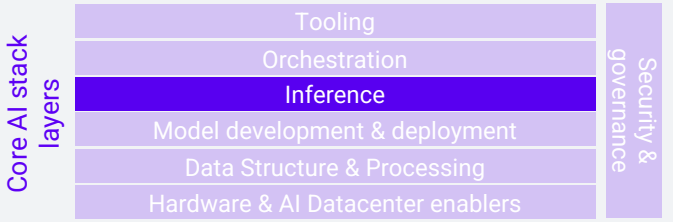
*Non exhaustive*




For more details, please email *hello@resonance.holdings*

# Market Map: Inference

*Workflow*




**AI inference**

1. With models deployed, the real-time AI Inference Systems kick in. These systems load the model into memory, receive input data (like a prompt or image), run the model forward pass, and return the prediction or generated output.

**Microservices architectures**

2. For modularity and scale, inferenc
often wrapped in Microserv
Architectures. These
services exp
functi

For more details, please email *hello@resonance.holdings*

# Market Map: Inference

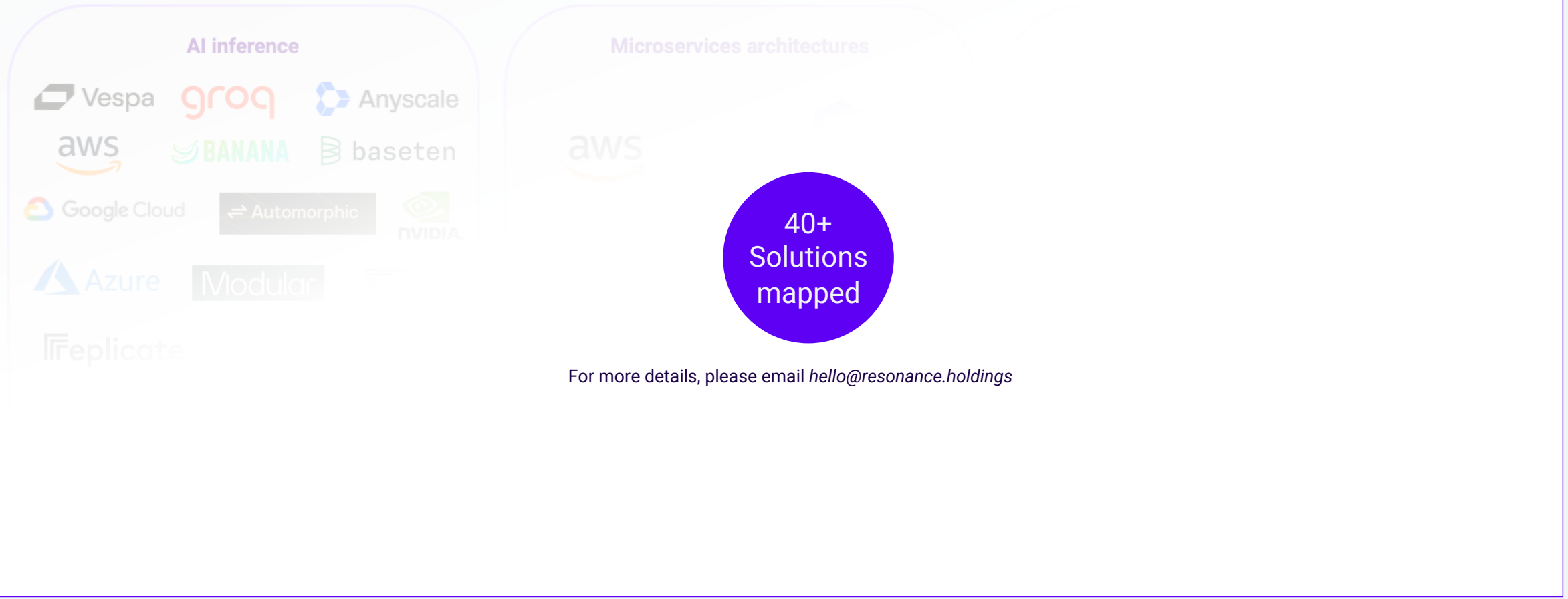
*April 2025 – A work in progress*

*Non exhaustive*

Core AI stack layers

- Tooling
- Orchestration
- **Inference**
- Model development & deployment
- Data Structure & Processing
- Hardware & AI Datacenter enablers

Security & governance




For more details, please email *hello@resonance.holdings*

# Market Map: Orchestration

*Workflow*

Core AI stack layers

- Tooling
- **Orchestration**
- Inference
- Model development & deployment
- Data Structure & Processing
- Hardware & AI Datacenter enablers

Security & governance

### Prompt & Model Orchestration

1. Orchestration begins with Prompt & Model Orchestration, where developers define the logic that sequences prompts, tools, and model outputs. This is the “brain wiring” of an LLM system — it decides how to query models, chain tasks, and retrieve external data.

### Agent frameworks

2. Once logic is defined, Agent Frameworks
autonomy to the system. These fr
models into stateful agents
maintain memory a
respond to
and

### Model gateway & routing

3. As multiple models or tools are used, Model Gate
& Routing systems dynamically decide whic
endpoint to use. They enable A/B te
load-based routing — choosi
Mistral for speed, for

For more details, please email *hello@resonance.holdings*

# Market Map: Orchestration

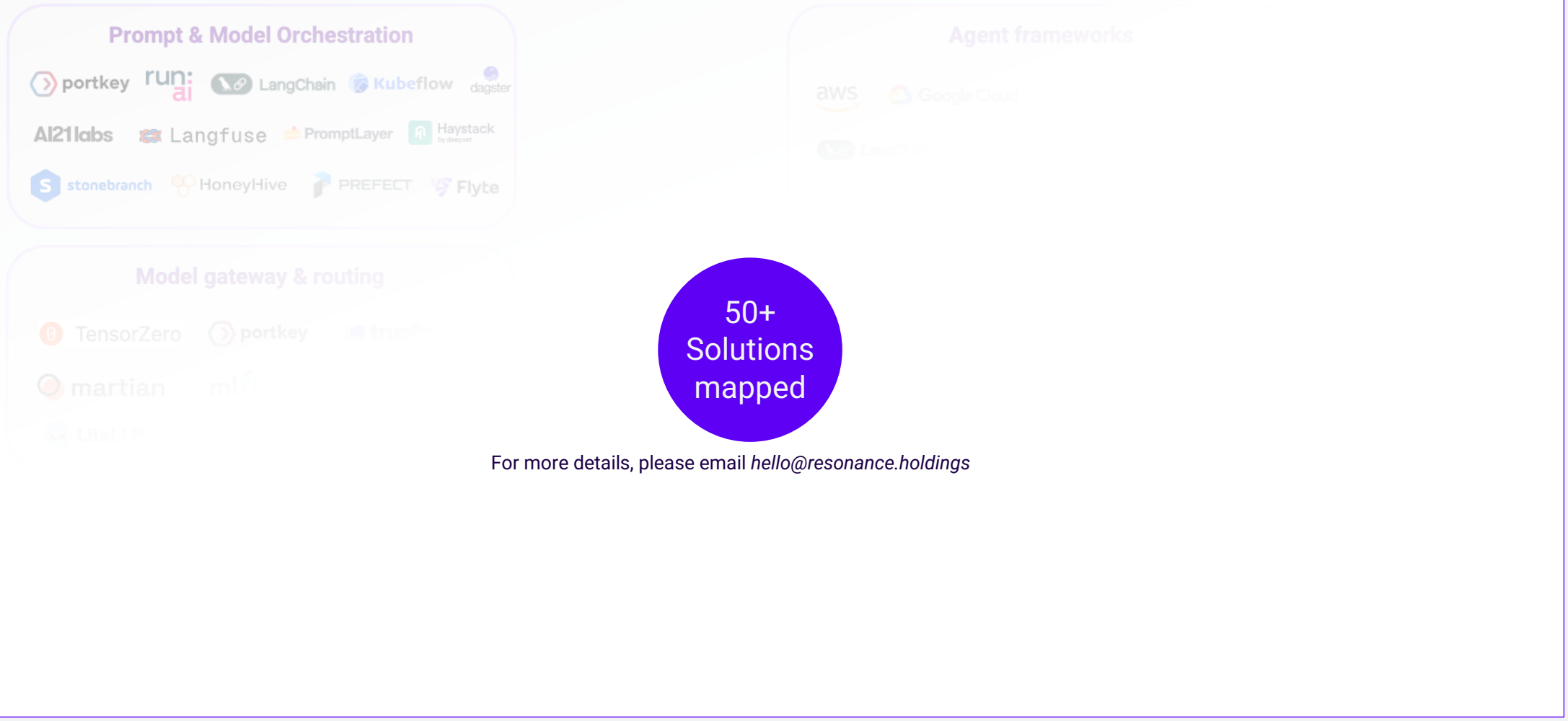
*April 2025 – A work in progress*

*Non exhaustive*

Core AI stack layers

- Tooling
- **Orchestration**
- Inference
- Model development & deployment
- Data Structure & Processing
- Hardware & AI Datacenter enablers

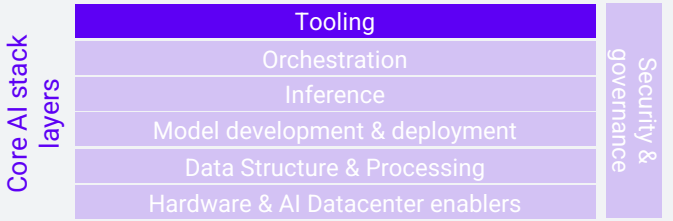
Security & governance




For more details, please email *hello@resonance.holdings*

# Market Map: Tooling

*Workflow*




### Dev environments

Cloud-native or local environments that enable developers to build, test, and debug AI code with preconfigured dependencies and GPU access. These platforms streamline reproducibility, CI/CD workflows, and agent/model prototyping.

### AI Ops (LLM & ML)

Operational tooling that monitors, evaluates, and improves AI systems in production. Includes LLMOps for prompt tracking, hallucination detection, and token usage, and MLOps for retraining, model versioning, and pipeline observability.

### Compute & cost

Platforms that improve infrastructure [illegible] across

### End-to-end AI app builder

Platforms to build, integrate, and deploy full AI-powered applications — combining logic, models, APIs, UI, and workflows. Targets builders looking to ship LLM apps, dashboards, and copilots [illegible] managing infrastructure.

[illegible]

For more details, please email *hello@resonance.holdings*

# Market Map: Tooling

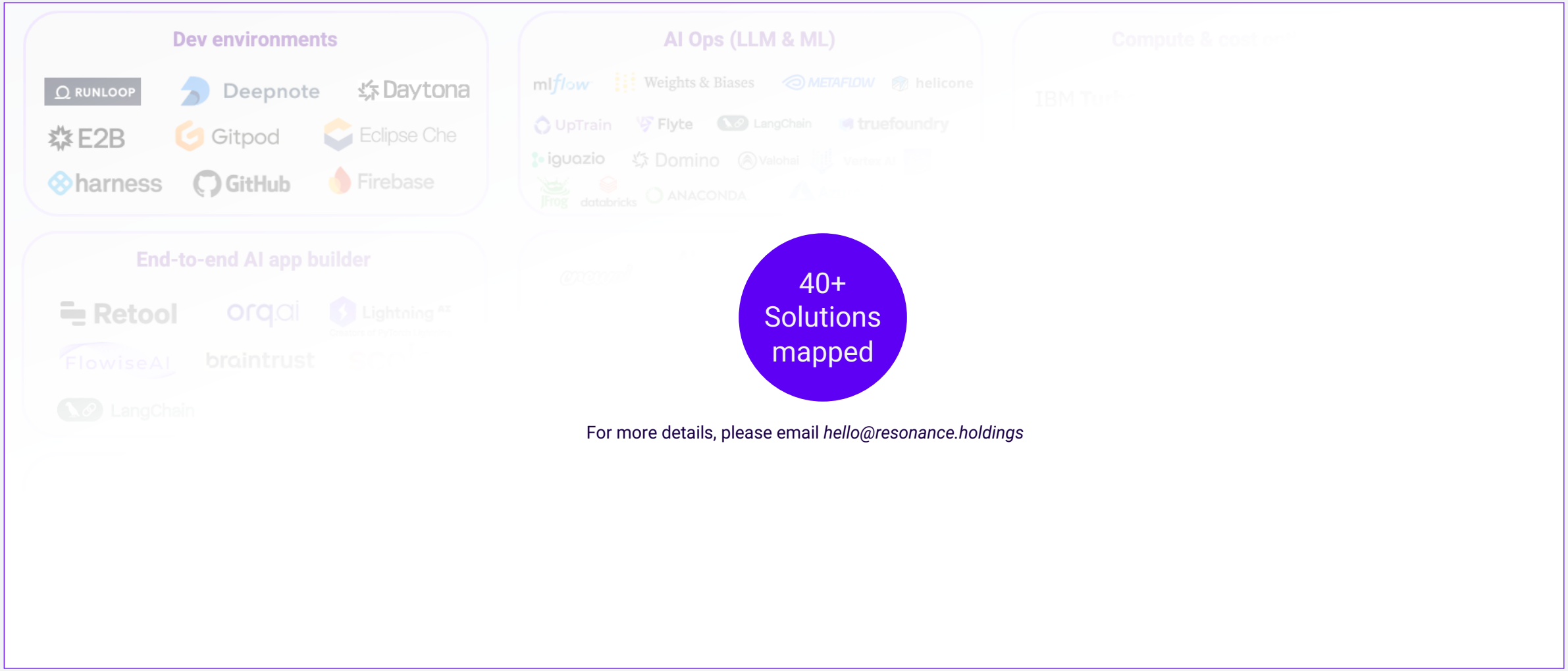
*April 2025 – A work in progress*

*Non exhaustive*

Core AI stack layers

- Tooling
- Orchestration
- Inference
- Model development & deployment
- Data Structure & Processing
- Hardware & AI Datacenter enablers

Security & governance

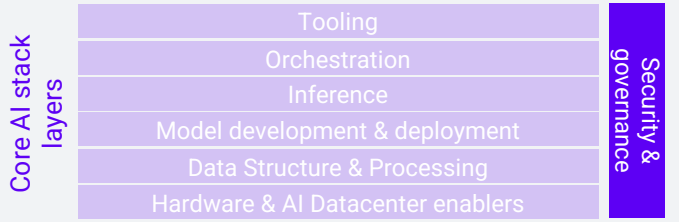



40+ Solutions mapped

For more details, please email *hello@resonance.holdings*

# Market Map: Security & governance

*Workflow*




**AI security & risk mitigation**

This category covers tools and practices that protect AI systems from emerging threats, vulnerabilities, and misuse — hacking, prompt injections, model theft, hallucinations, or unsafe outputs.
The main focus is to ensure the reliability, robustness, and ethical integrity of models throughout their life… training to inference…

**Data Governance, privacy & compliance**

Data Governance, priva… platforms mana… control o…

For more details, please email *hello@resonance.holdings*

# Market Map: Security & governance

*April 2025 – A work in progress*

*Non exhaustive*

Core AI stack layers

- Tooling
- Orchestration
- Inference
- Model development & deployment
- Data Structure & Processing
- Hardware & AI Datacenter enablers

Security & governance


20+ Solutions mapped

For more details, please email *hello@resonance.holdings*

# Thank you

AI INSIDER
powered by RESONANCE