SCI for AI
Discussion points

Navveen Balani , Co-Chair, Standards working group - GSF

SCI Use Cases in AI

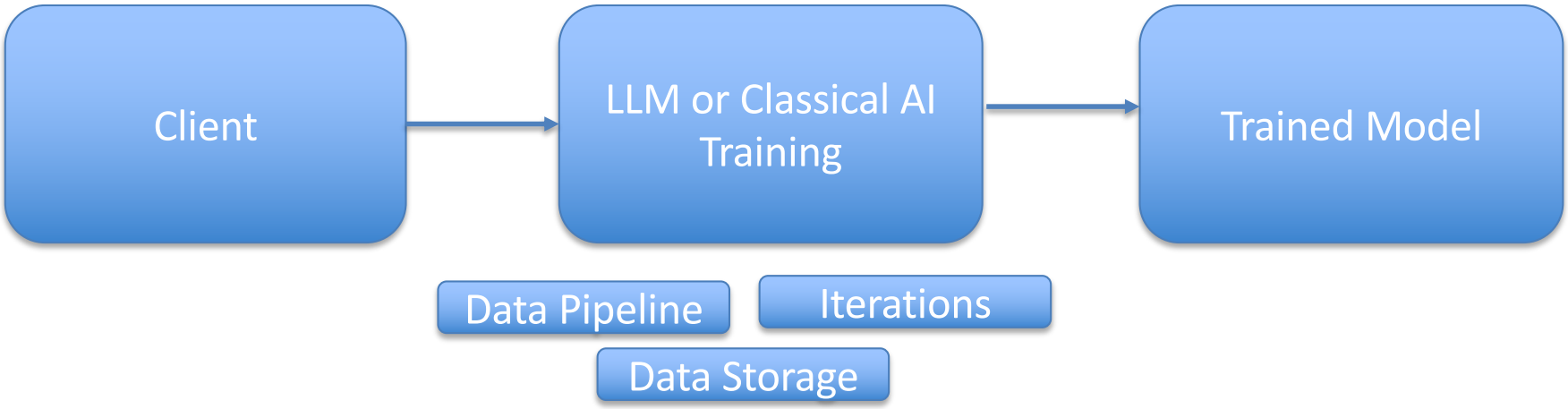
Use Case 1: LLM API Consumers (Prompt-based Usage)

•Key Questions:

- Should training emissions (of trained model) be **amortized** into per-prompt SCI?
- Is inference cost already **included by provider**?
- How to account for **dynamic model routing**?
- Some analogy - If you are using an managed database or API gateway, do you factor development cost for database or API gateway or only runtime ?

SCI Use Cases in AI

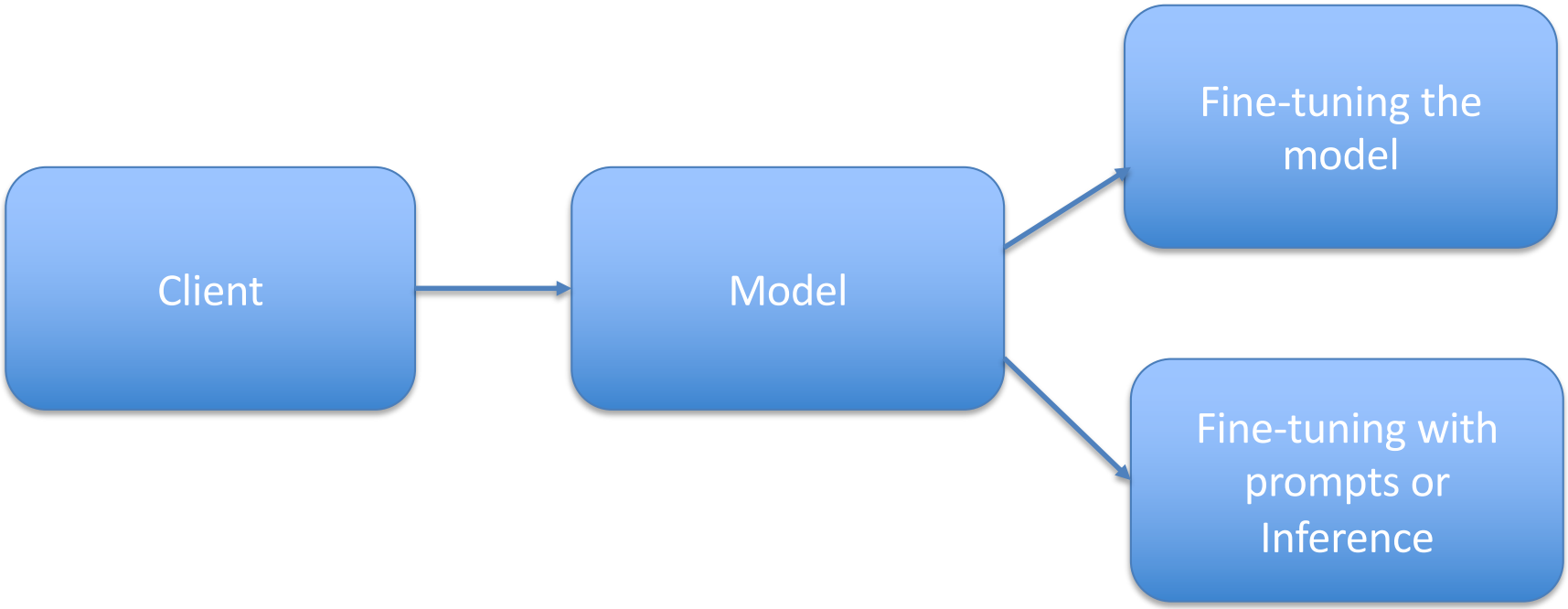
Use Case 2: Model Creators and Owners



- Boundary Considerations:
  - Include **entire data pipeline**?
  - Handle **multiple training iterations**?
  - **Versioning** and upgrades

SCI Use Cases in AI

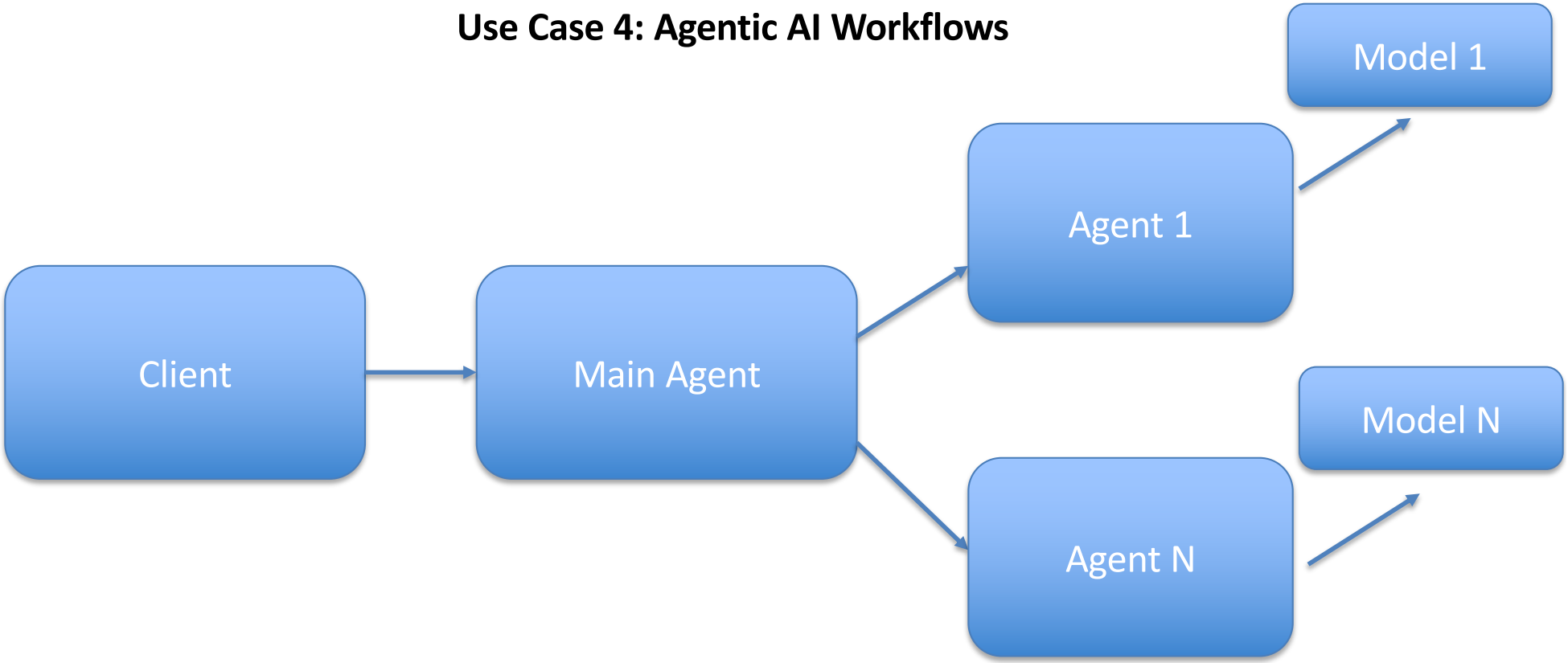
Use Case 3: Model Fine-Tuners (Pre-trained + Custom Data)



- Sits between full training and inference
- Requires **distinct SCI accounting**

SCI Use Cases in AI

Use Case 4: Agentic AI Workflows



- Composite workflows using **multiple models**
- Combination of API-based inference + fine-tuning
- Requires **hybrid SCI strategy**

SCI in AI - Actions

Use case	Training of Model included	Actions for reduction
LLM API Consumers (Prompt-based Usage)	No	• **Contextual prompt design** to minimize response latency and reduce the number of interactions needed for accurate output. • **Prompt caching and batching** to avoid redundant computations and enhance efficiency at scale • **Dynamic model routing** to select the most efficient model based on task requirements—optimizing for carbon footprint, cost, performance, and safety • **Energy-efficient hardware deployment** for inference workloads, where organizations host or run LLMs internally.

SCI in AI - Actions

Use case	Training of Model included	Actions for reduction
Model Creators and Owners	Yes	• **Select energy-efficient model architectures** that balance performance and training cost. • **Utilize programming languages and frameworks** known for computational efficiency and low overhead • **Utilize custom energy-efficient chips** optimized for AI training workloads. • **Optimize supporting components**, including energy-efficient data pipelines, storage solutions, and data augmentation processes.

SCI in AI - Actions

Use case	Training of Model included	Actions for reduction
Model Fine-Tuners (Pre-trained + Custom Data)	Yes (only for the fine-tuned model, not for the original pre-trained model)	• **Select energy-efficient model architectures** that balance performance and training cost. • **Utilize programming languages and frameworks** known for computational efficiency and low overhead • **Utilize custom energy-efficient chips** optimized for AI training workloads. • **Optimize supporting components**, including energy-efficient data pipelines, RAG, storage solutions, and data augmentation processes.

SCI in AI - Actions

Use case	Training of Model included	Actions for reduction
Agentic AI Workflows	**No**	• **Choose energy-efficient languages, frameworks, and architectures** to optimize multi-agent execution • **Deploy custom energy-efficient chips** for running agentic workflows and distributed reasoning tasks. • **Design contextual prompts for effective reasoning**, minimizing unnecessary computation and agent-to-agent interactions. • **Implement prompt caching and batching** across agents to reduce duplication and improve execution efficiency. • **Optimize supporting components**, including energy-efficient data pipelines, RAG, storage systems, and augmentation workflows.

SCI in AI

To define SCI for AI:

- Recognize **distinct roles** and **use cases**
- Clarify **what to include** in SCI boundaries
- Drive **consensus** across stakeholders
- Make it **simple** to implement, **explainable** and **actionable** for each of use cases to lower carbon emissions and increase energy efficiency.