

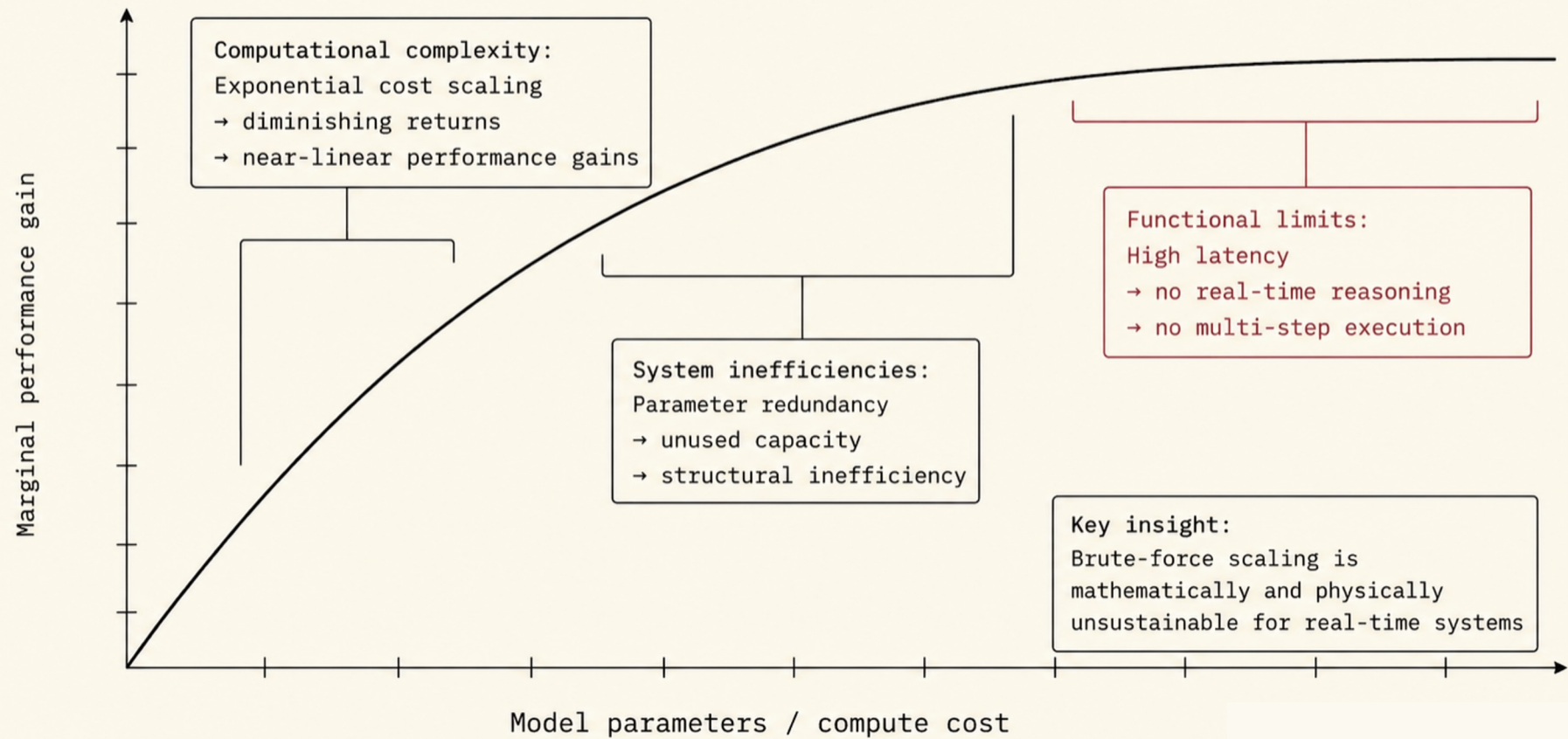
Structural Approaches to Model Compression and Orchestration

Moving beyond LLM scalability limits
to efficient, multi-agent execution.

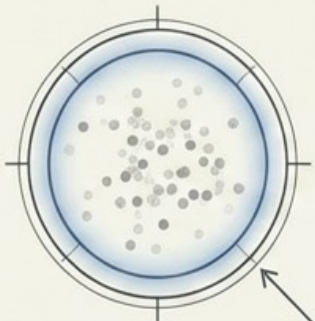
HPC AdminTech 2026

Dr. Daniel López

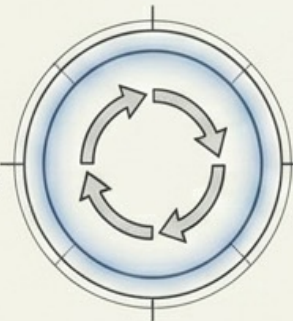
THE LIMIT OF BRUTE-FORCE SCALING



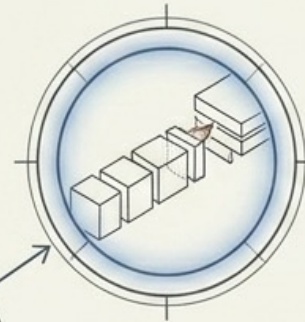
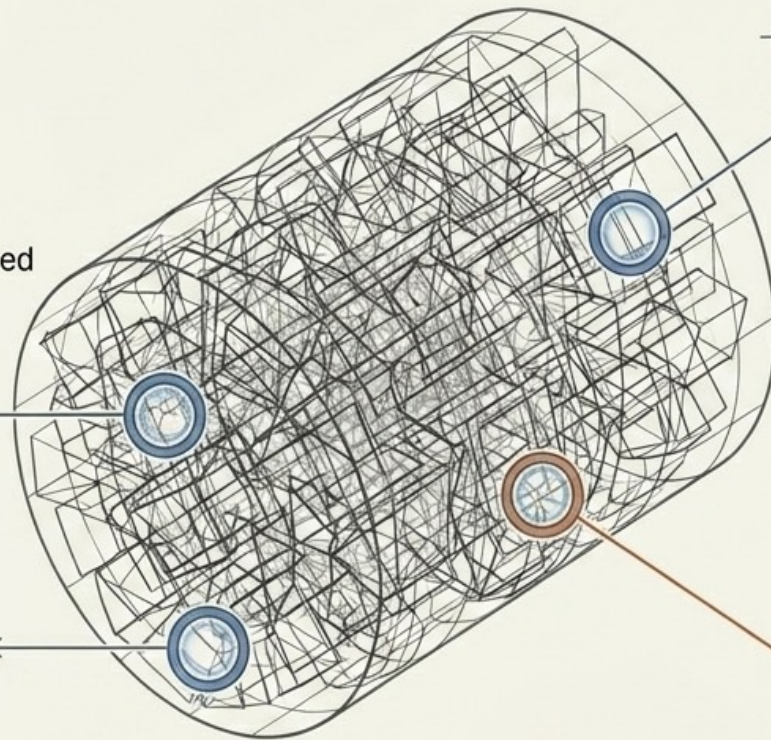
The Structural Inefficiencies of Brute-Force Scaling



Massive inactive node clusters representing unused overcapacity.



Closed-loop generation lacking execution pathways.



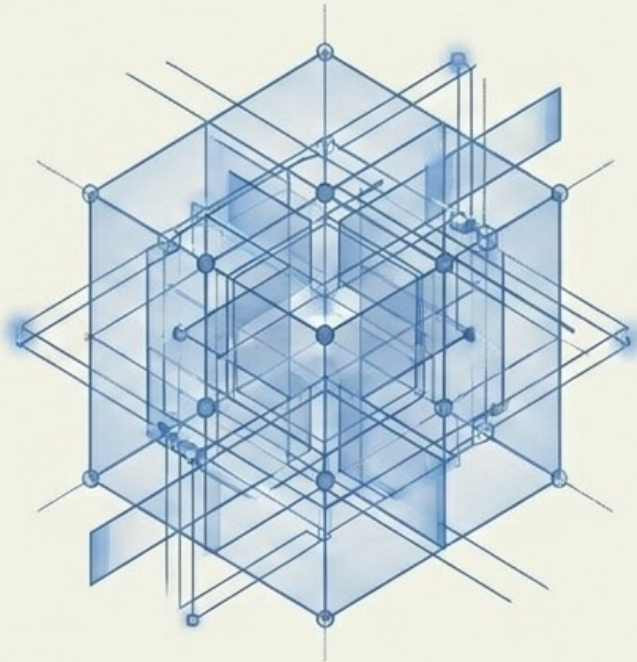
Sequential processing gridlock leading to delayed inference.

[SYSTEM LIMIT]
Unused parameter overcapacity drives up hardware dependency without proportional gains in functional reasoning.

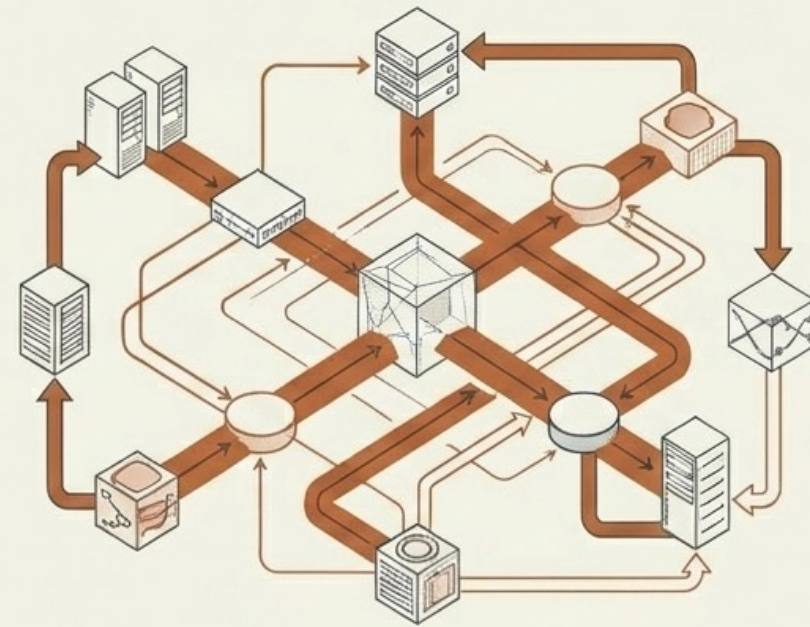
$[-0 \ 0 \ 0 \ 0 \times 2]$

0	0	0
1	2	0
1	0	0
0	0	0
0	0	0
0	0	0

The problem is not the model. It is the entire architecture.



Layer 1: Structural Compression
Re-evaluating the mathematical foundation of the model itself.



Layer 2: Execution Architecture
Re-evaluating how models interface with complex processes.

True scalability requires simultaneous innovation at both the tensor level and the orchestration level.

Compression is Not Trivial Reduction

	Standard Reduction (Pruning / Quantization)	Structural Re-expression (CompactifAI)
Mechanism	Trimming weights & lowering precision	Tensor networks and low-rank decomposition.
Size vs. Precision Trade-off	High parameter drop causes sharp precision degradation	Maintains precision through mathematical re-expression.
Inspiration	Heuristic algorithms	Complex systems physics.

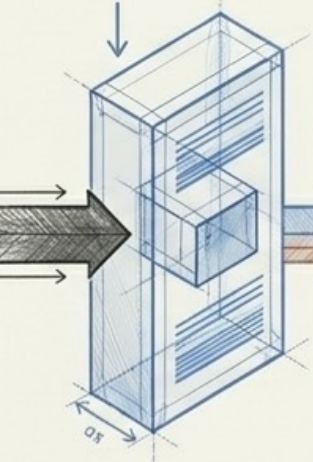
The Three-Stage Structural Re-expression Pipeline

Gate 1: Model Profiling
Analyzing parameter density
and identifying optimal
decomposition zones.

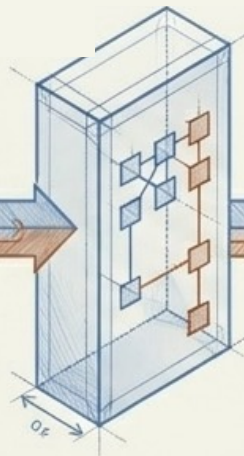
Bloated Parameter Mass



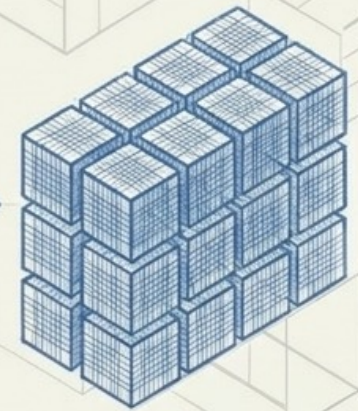
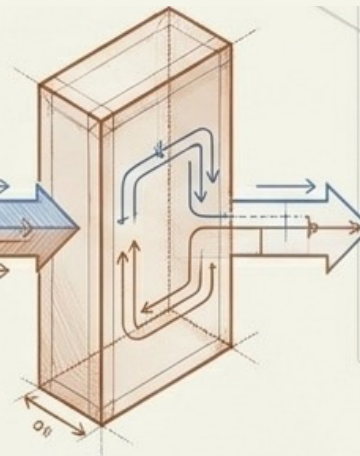
UNOPTIMIZED MODEL:
Parameter Tanglement



(Measures baseline
information retention)



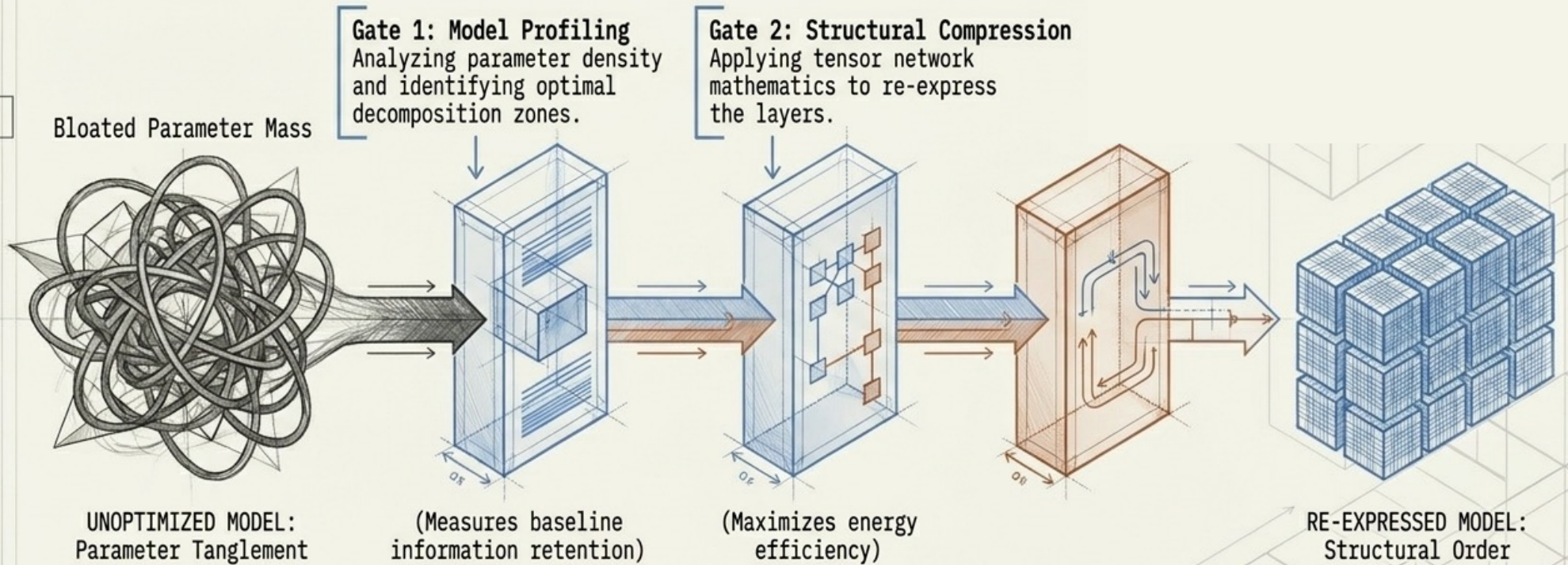
(Maximizes energy
efficiency)



RE-EXPRESSED MODEL:
Structural Order

Open-weight models:
LlaMA, Mistral, DeepSeek, Gemma, GPT,...

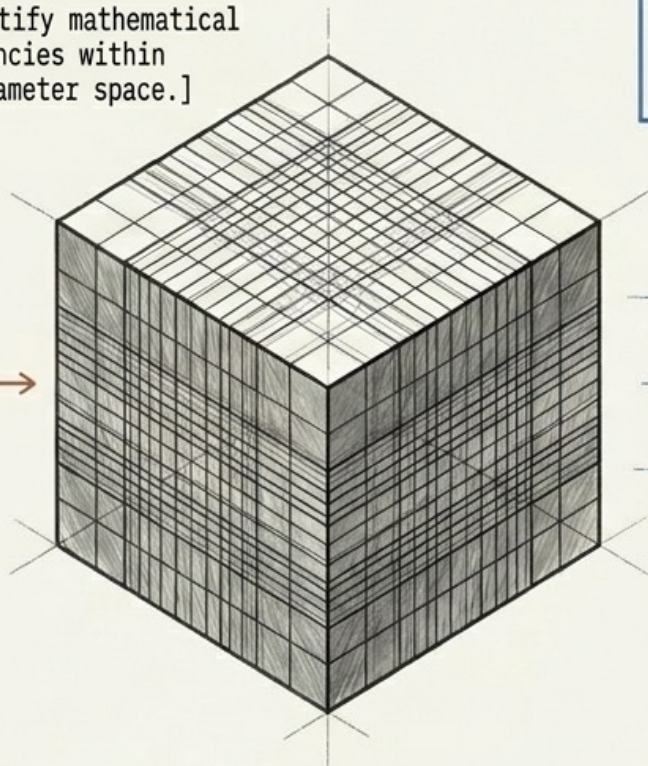
The Three-Stage Structural Re-expression Pipeline



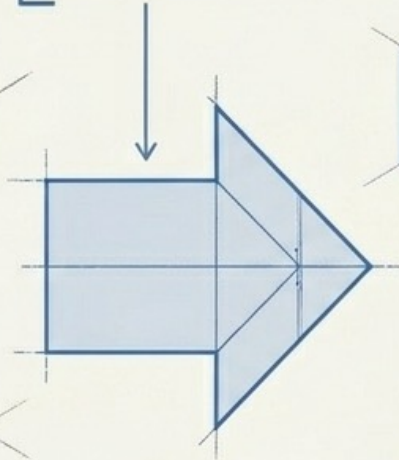
Open-weight models:
Llama, Mistral, DeepSeek, Gemma, GPT,...

The Mathematics of Structural Re-expression

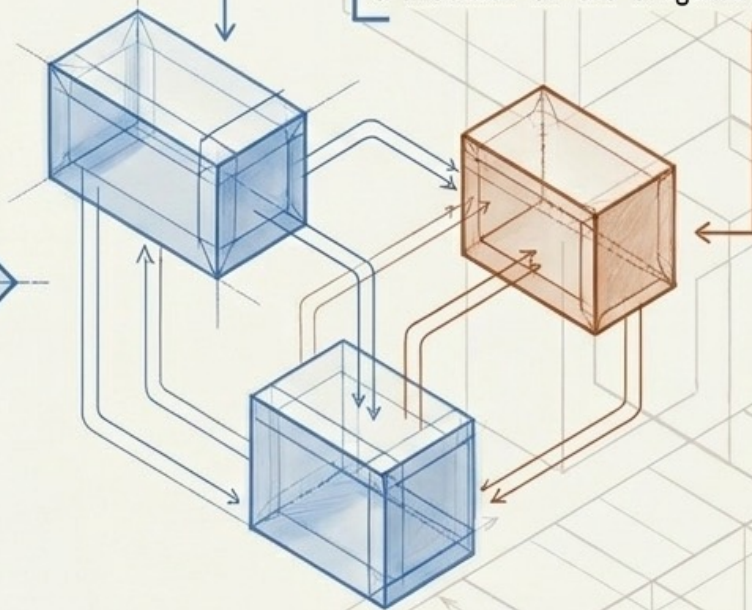
[1. Identify mathematical redundancies within the parameter space.]



[2. Decompose the architecture into low-rank tensor components.]



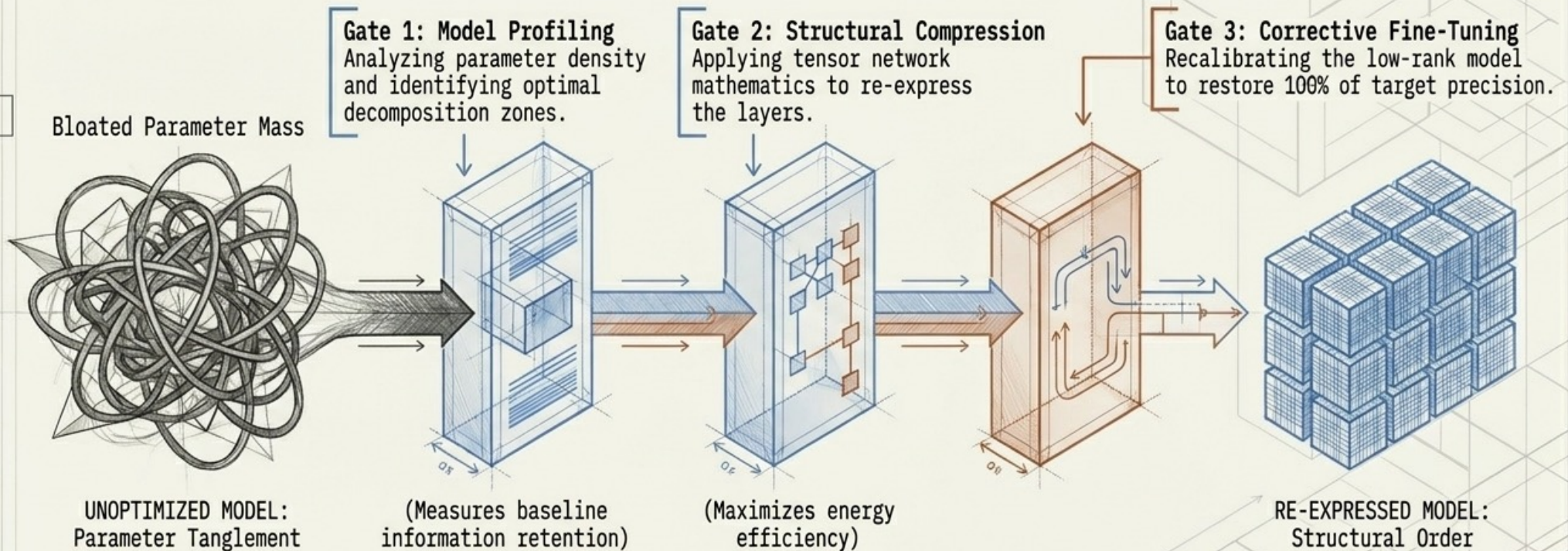
[3. Reconstruct the model to execute the identical logic with a fraction of the original volume.]



**We do not compress models.
We re-express them.**

Efficiency depends on model structure: more structure and redundancy enable better compression with minimal accuracy loss.

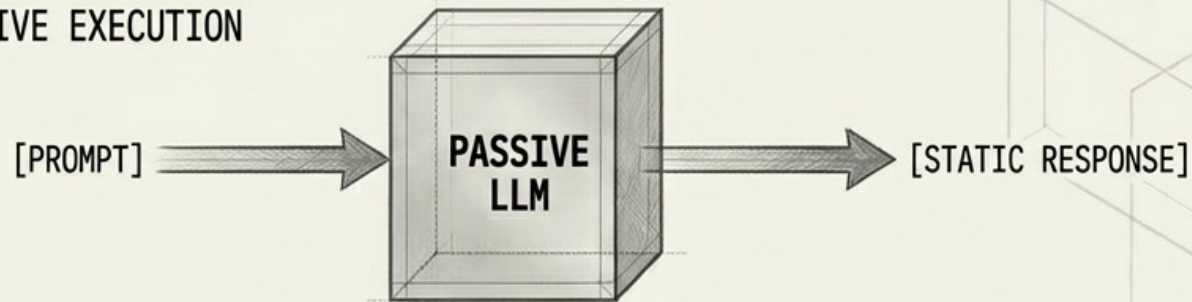
The Three-Stage Structural Re-expression Pipeline



Open-weight models:
Llama, Mistral, DeepSeek, Gemma, GPT, ...

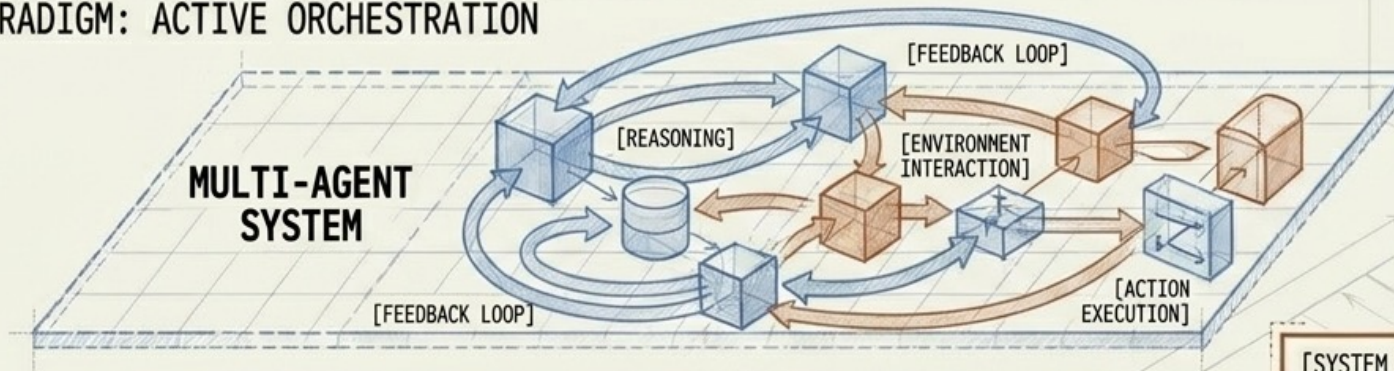
The Capability Shift: From Models to Systems

THE OLD PARADIGM: PASSIVE EXECUTION



Incapable of executing complex processes.


THE NEW PARADIGM: ACTIVE ORCHESTRATION



From generating responses to executing actions.

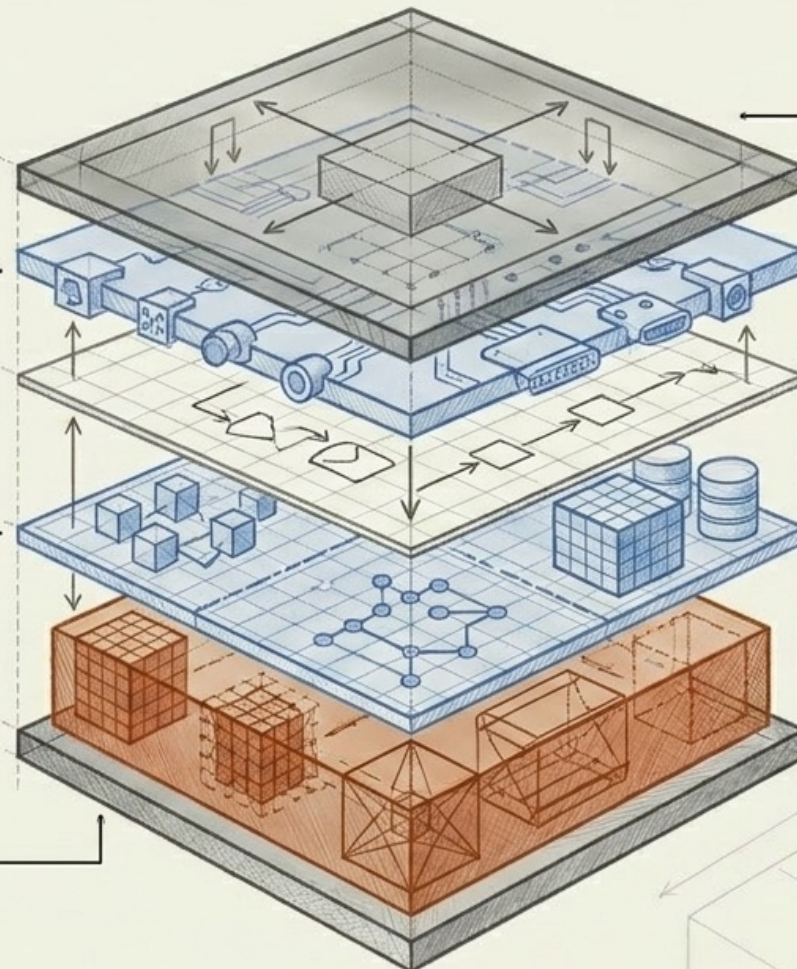
[SYSTEM CONSTRAINT]: Multi-agent coordination requires non-linear control logic and robust error handling protocols; computational scales with interaction complexity.

Multi-Agent Architectural Wrappers (CRAFT)

External Integration Layer: API  hooks and system interfaces.

Memory Layer: Split into Short-Term (Context) and Long-Term (RAG/Vector).

Core Engine:
The highly compressed LLM (CompactifAI).

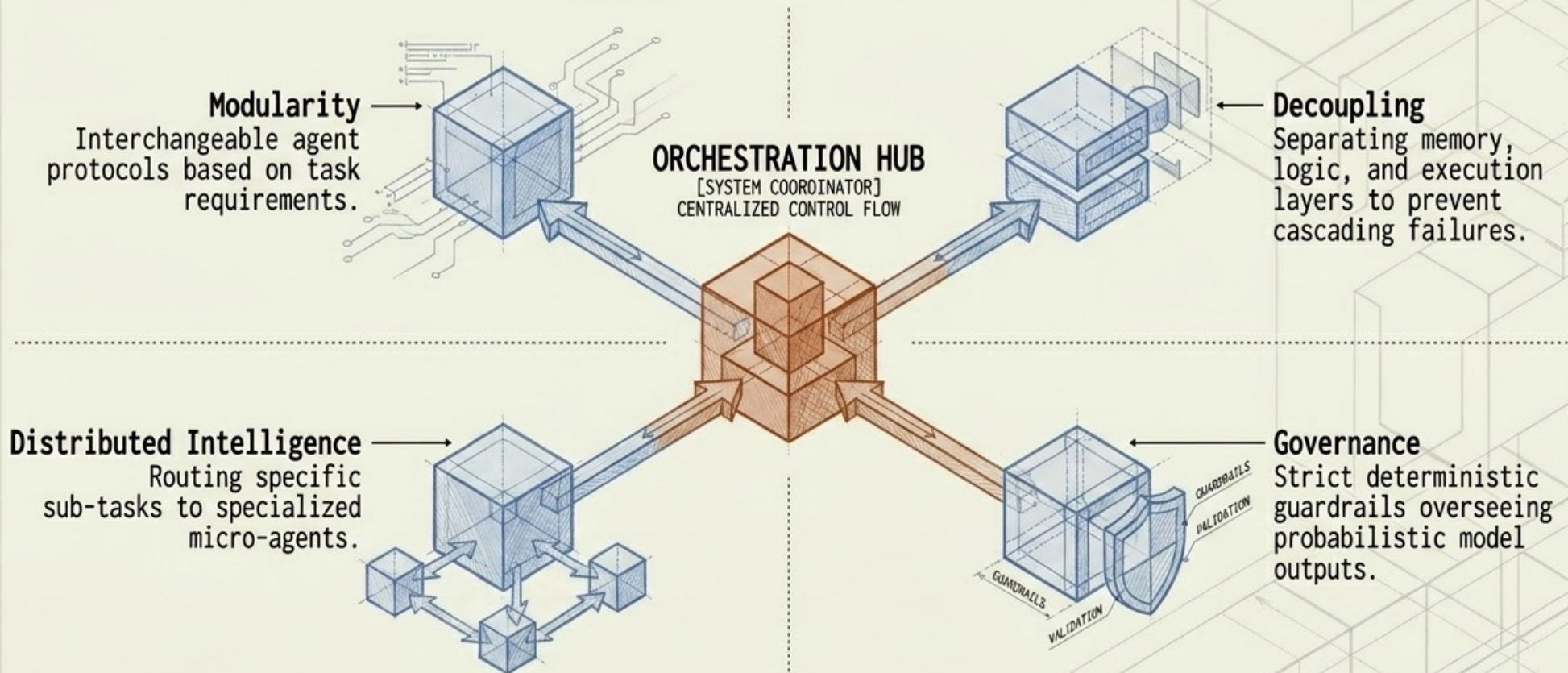


Orchestration Shell:
Global governance and agent coordination.

Planning Layer:
Task decomposition and logic routing.

The ultimate value lies not in the underlying model, intelligent coordination of specialized capabilities.

System Principles: Core Design Pillars of Multi-Agent Architectures



[VALIDATION PROTOCOL] Q: How are multi-agent systems validated?

A: Through isolated unit testing of decoupled agents before integration into the overarching governance shell.

WHAT IS CRAFT?

CRAFT: Coborg Responsible AI Framework & Toolkit

CRAFT is an enterprise-grade platform for building, orchestrating and governing AI agents in a secure, responsible and scalable way.

A MODULAR, ENTERPRISE-READY PLATFORM

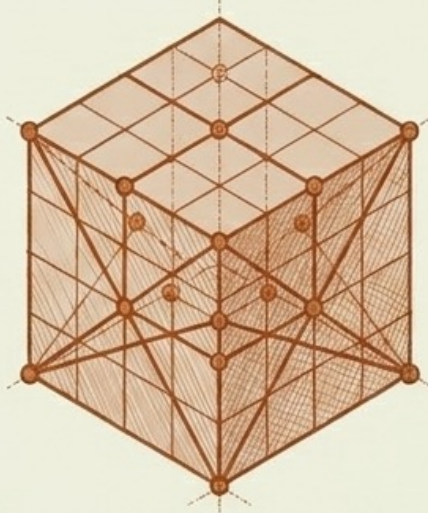
- 9 specialized microservices working in harmony
- Built from the ground up for Responsible AI
- Cloud agnostic: Azure, AWS, GCP, private cloud
- OpenAI-compatible API interface
- Security integrated across all layers

9 SPECIALIZED MICROSERVICES

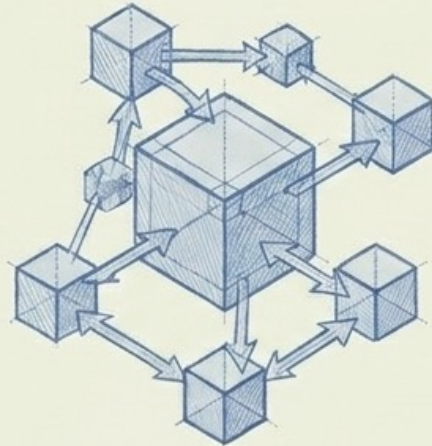


The ultimate value lies not in the underlying model, intelligent coordination of specialized capabilities.

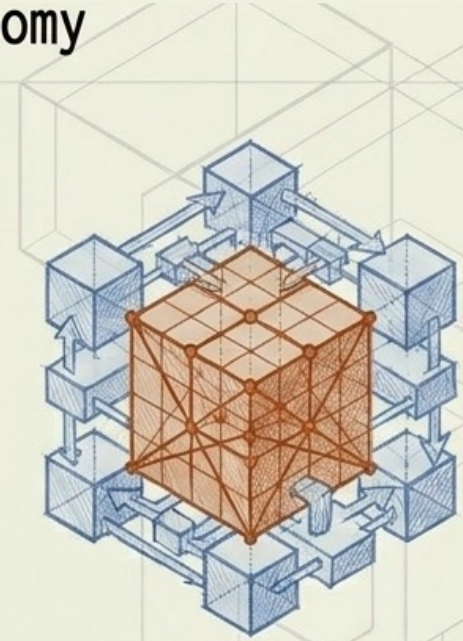
Synthesis Equation: The Path to Sustainable Autonomy



(Low Latency /
Low Compute)



(High Orchestration /
High Compute Demand)



**Sustainable
Autonomous Systems**

Efficiency Enables Agency.

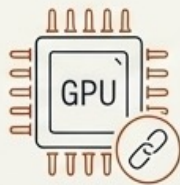
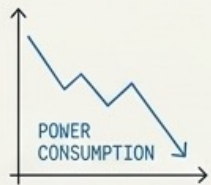
Without structural efficiency, multi-agent systems do not scale due to prohibitive compute costs. Without agentic architecture, efficient models fail to generate real-world impact.

Implications for System-Level Deployment



Point 1: Computational Sustainability

Drastic reduction in energy expenditure per executable task, breaking the dependency on massive GPU clusters.



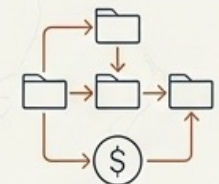
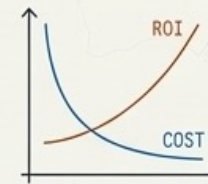
Point 2: Deployment Democratization

Enabling complex, multi-agent architectures to run on Edge devices and secure On-Premise environments.

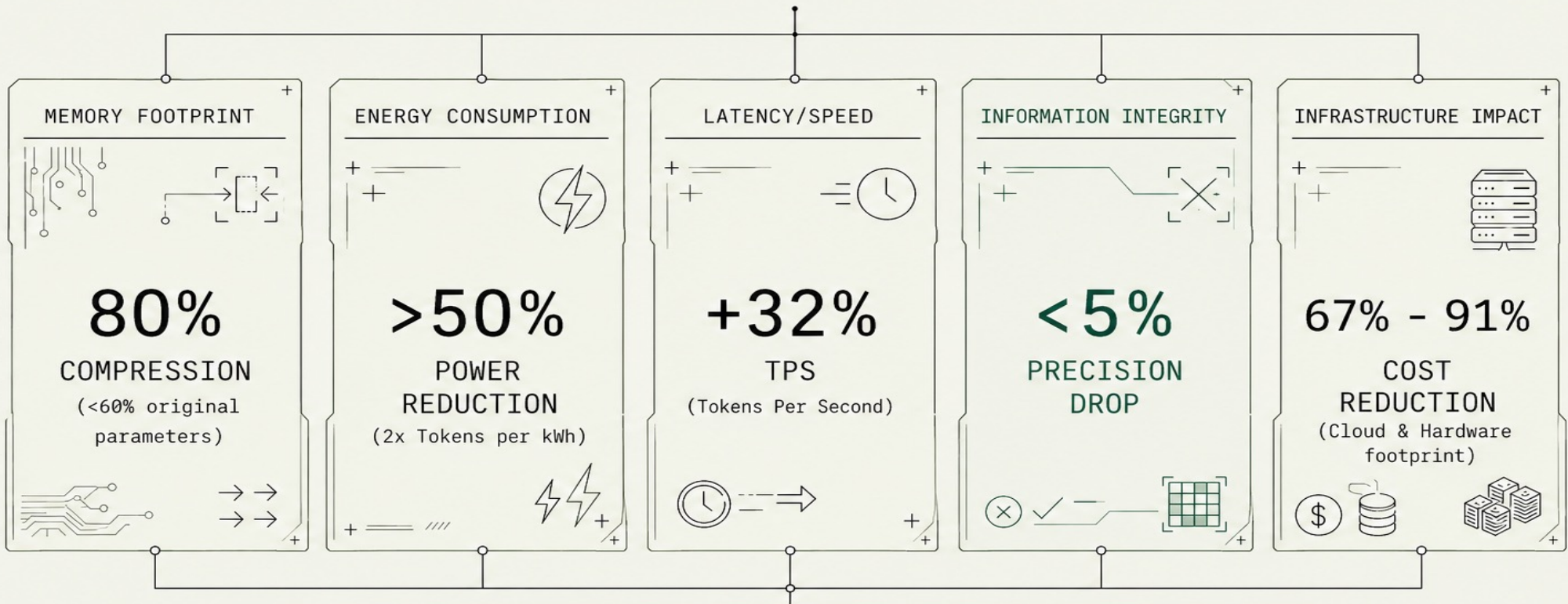


Point 3: Economic Viability

Flattening the exponential cost curve of AI integration, delivering scalable ROI for enterprise workflows.

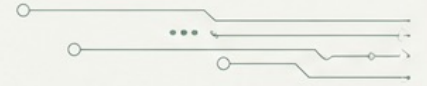


PHYSICAL AND COMPUTATIONAL OUTCOMES



Conclusion: Real-time edge deployment is unlocked purely through physics and mathematics.

Happy clients ;)



sopra  steria

"The research confirms that the model compressed via **CompactifAI provide substantial reductions** in power consumption, operational costs, and carbon emissions while maintaining competitive accuracy."

[Read More →](#)



"We conducted extensive technical benchmarking and were very impressed with the results: decrease in time to first token, increase in token throughput, and models that are cheaper to run **(50–80% cost and energy savings)**"

[Read More →](#)

 Telefónica

"The compressed models developed can be deployed directly on Telefónica's network, including local facilities, making it possible to **reduce energy consumption by up to 75%** compared to uncompressed models."

[Read More →](#)

Luzia

"Integrating CompactifAI's compressed models into our customer support chatbot has been a game changer. We have **reduced our model footprint by over 50%** while maintaining high response quality with lower latency and cost."

[Read More →](#)

HPC Admintech 2026

daniel.lopez-fernandez@inetum.com



MULTIVERSE
COMPUTING

Compactif AI

inetum.

CRAFT