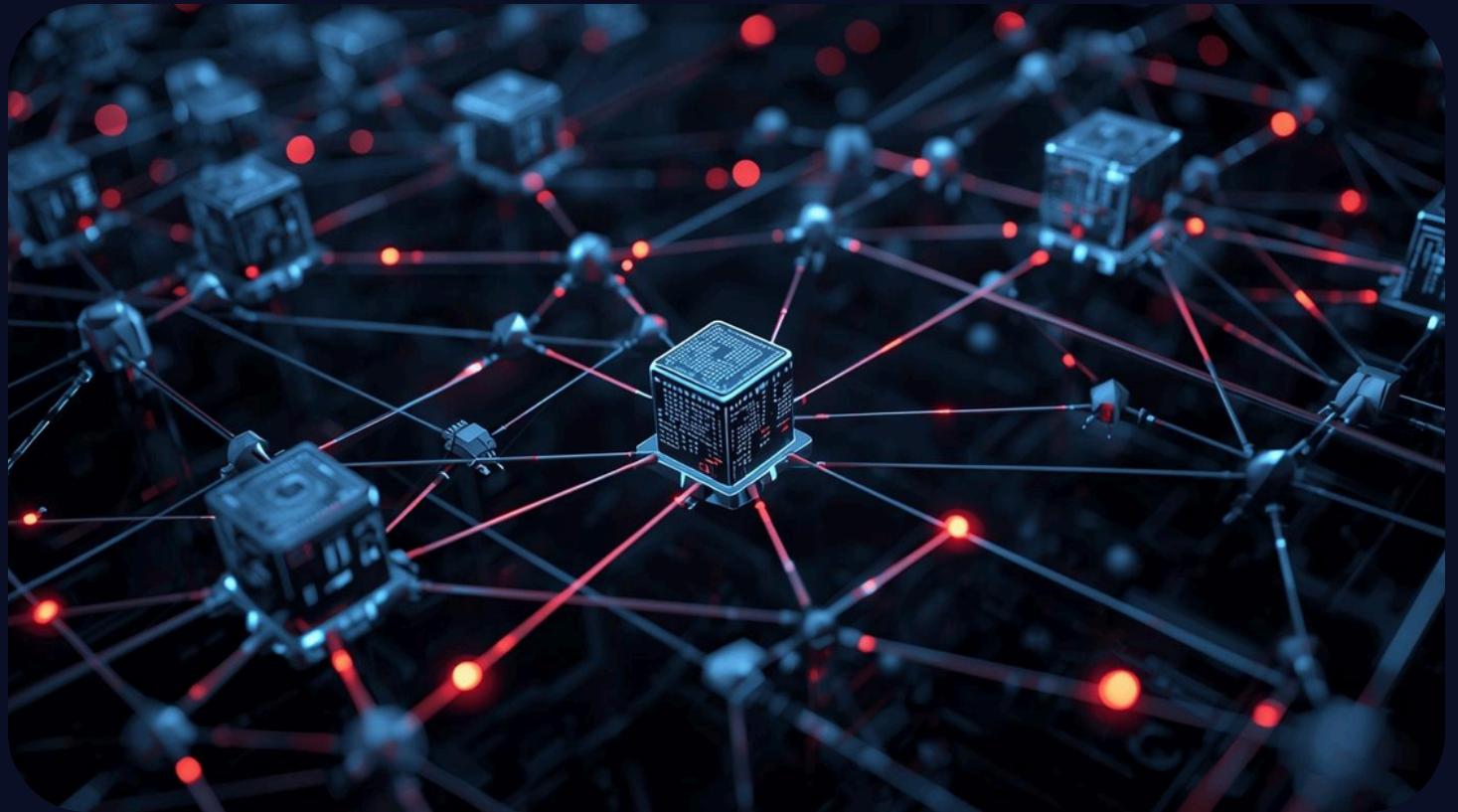
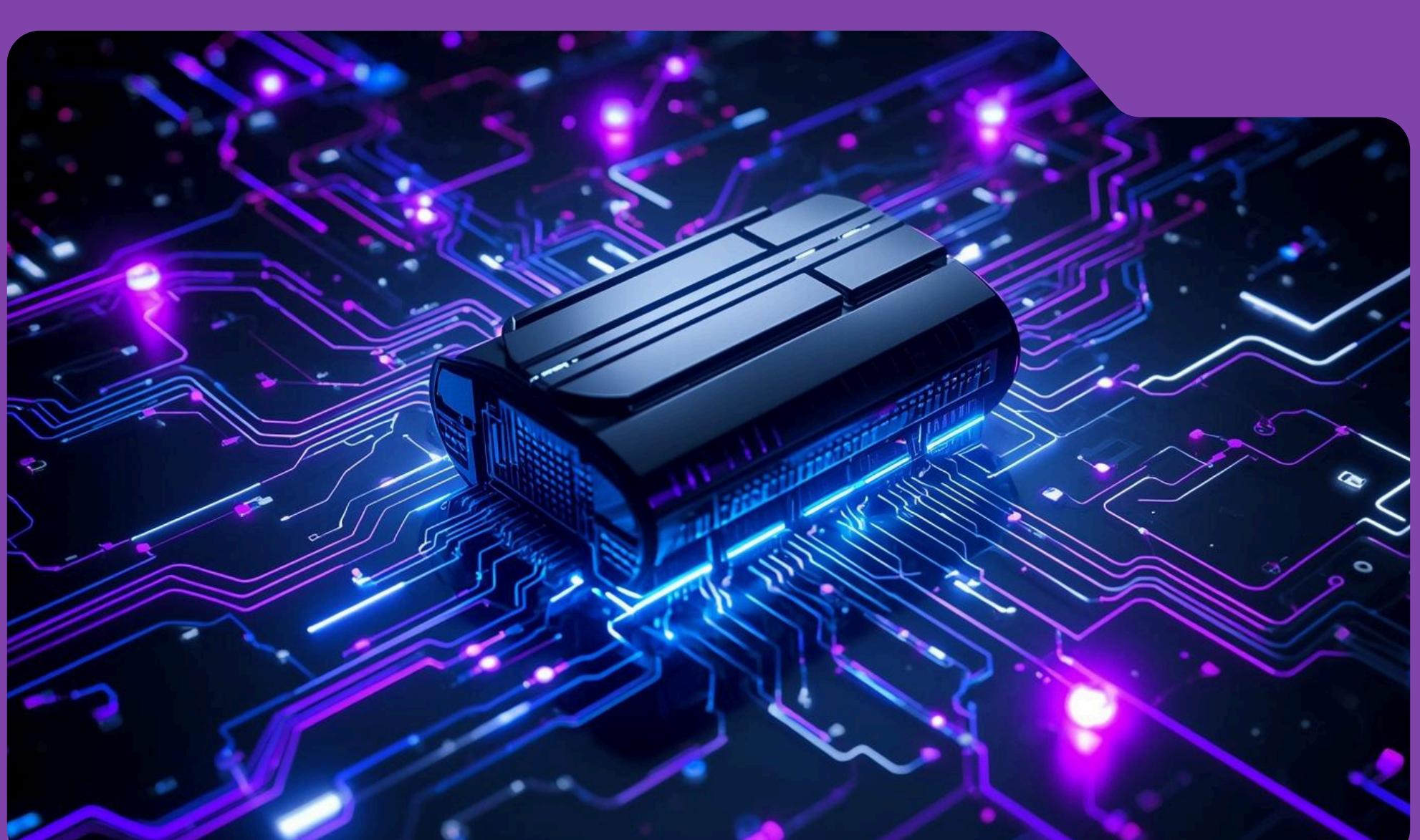
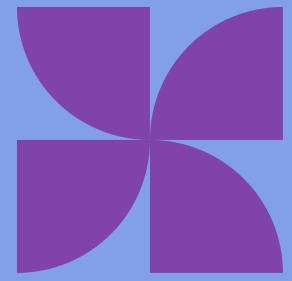


The Mistral Spectrum: Strategic AI Selection



Balancing Performance, Latency,
and Cost for the Enterprise

Lightweight Workhorses

Mistral 7B deliver high-speed efficiency for low-latency tasks. Specialized in routing and classification, they excel in single-GPU deployments, optimizing hardware capabilities while maintaining flexible context window sizes.

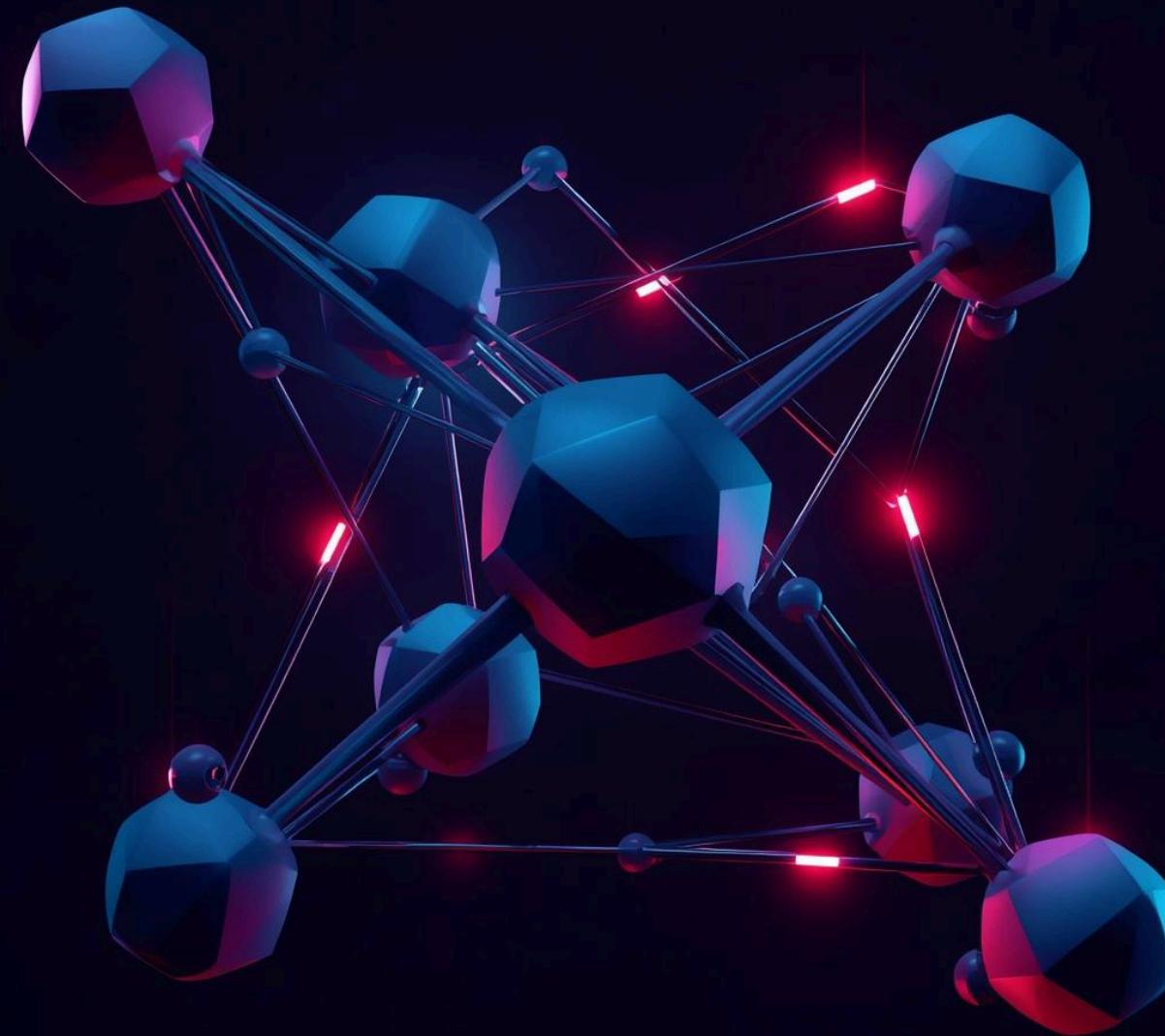
Mistral NeMo: Co-developed with NVIDIA; 128k context window for document-heavy RAG.



Efficiency Revolution

Mixtral 8x7B

Uses only 13B parameters for inference while maintaining 46.7B parameter intelligence.



Frontier Reasoning Model

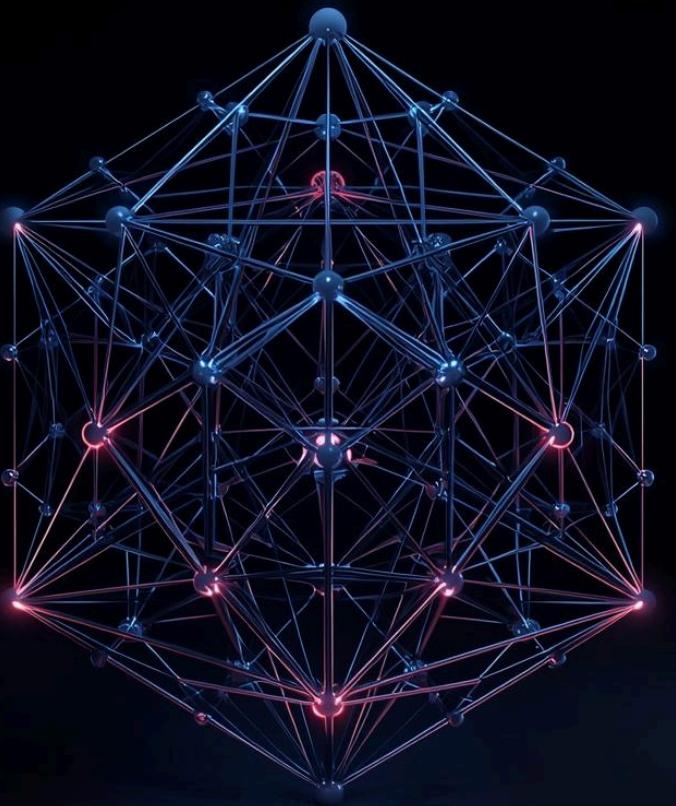
The pinnacle of computational capability

Mistral Large 2 boasts the **highest parameter count**, enabling complex logic, coding, and multi-step reasoning, serving as the orchestrator in sophisticated multi-agent systems for advanced applications.



The Architect's Trade-off

- High Speed / Low Cost: Use 7B or NeMo.
- Balanced Logic: Use 8x7B.
- High Reasoning / High Complexity: Use Large 2.



Contact Us



Right-Sizing

Match model complexity strictly to the business task.



Resource Management

Optimize VRAM usage on your Google Cloud GPUs.



Sovereignty

Maintain 100% control through open-weight deployment.