W69 AI Consultancy

LLM Orchestration & Integration.

Multi-model architectures deploying the right LLM for every task.

Service Brief | AI Engineering

MARKET INSIGHTS 2026

One LLM isn't enough. Orchestration is key.

The LLM landscape is exploding: GPT-4, Claude, Gemini, Llama, Mistral — each with unique strengths. **78% of enterprise AI teams** already use multiple LLMs. But without intelligent orchestration, this leads to fragmentation, vendor lock-in and unpredictable costs. The solution: an orchestration layer that automatically selects the right model for each task.

78%

Use Multi-LLM

Of enterprise teams work with multiple LLM providers (Forrester)

40%

Cost Savings

Through intelligent model routing vs. GPT-4 only (McKinsey)

3.7x

Better Output

Multi-model pipeline vs. single-model approach for complex tasks

60%

Less Latency

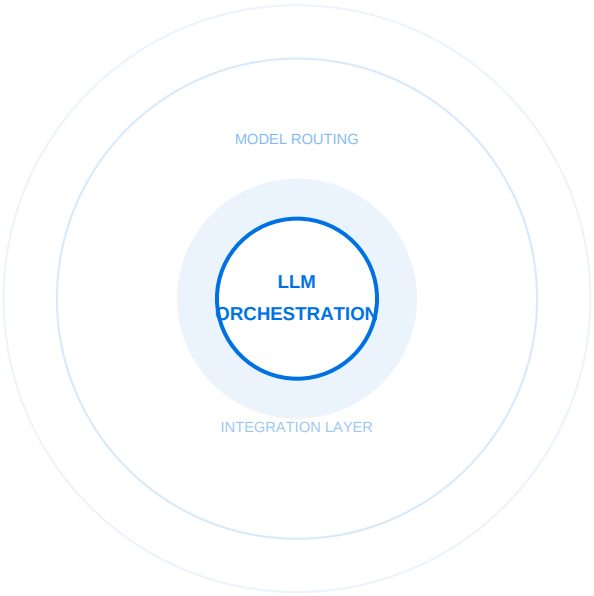
Through edge caching and model-specific optimization

> *“The future isn't one AI model — it's an orchestra of specialized models.”*

W69 FRAMEWORK

The LLM Orchestration Suite

Five components for intelligent multi-model AI architectures.



- **Model Router**

  Intelligent routing automatically selecting the optimal model based on task complexity, latency and cost.

- **RAG Architecture**

  Retrieval-Augmented Generation seamlessly integrating your business knowledge with LLM capabilities.

- **Context Engineering**

  Context assembled per step: instruction, knowledge layer, tools, memory and boundaries.

- **API Gateway**

  Unified API abstracting multiple LLM providers behind one consistent interface.

- **Cost & Performance**

  Real-time monitoring of model performance, token usage and cost per request.

Cross-cutting: Model Routing and Integration are interwoven for optimal cost-performance balance.

INTELLIGENT ORCHESTRATION

The right model for every task

How W69 designs multi-model architectures that reduce costs and increase quality.

1 Adaptive Model Selection

Not every task requires GPT-4. W69 implements intelligent routing: simple tasks to efficient models, complex reasoning to premium models. Automatic, transparent and cost-optimal.

2 RAG Pipeline Design

Your business knowledge + LLM intelligence = superior answers. W69 designs RAG architectures with semantic chunking, hybrid search and re-ranking for maximum relevance.

3 Fine-Tuning Strategy

When is fine-tuning worthwhile? W69 evaluates costs vs. benefits and implements targeted fine-tuning where it measurably adds value.

ENTERPRISE INTEGRATION

LLMs seamlessly in your enterprise

How AI models integrate with your existing systems and workflows.

1 System Integration

Connecting LLMs with SAP, Salesforce, ServiceNow and custom systems. API-first architecture making AI capabilities available to every business application.

2 Streaming & Real-Time

Real-time AI responses for customer-facing applications. Server-Sent Events, WebSocket architectures and edge deployment for sub-second latency.

3 Enterprise Security

Secure LLM integration: data filtering, PII redaction, prompt injection protection and audit logging for every AI interaction.

AGENTIC WORKFLOWS

From static pipelines to self-optimizing orchestration

Traditional LLM pipelines are hard-coded. **Agentic Orchestration** optimizes itself. W69 implements systems that automatically evaluate models, adjust routing and optimize costs — continuously and autonomously.

1 **Model Evaluation**

Automatic benchmarking of new models against your specific use cases. Adopt better models as soon as they become available.

2 **Dynamic Routing**

Self-optimizing routing learning which model performs best per task type, and automatically routes accordingly.

3 **Human-in-the-Loop**

Model selection decisions for critical workflows escalate. Routine optimizations run autonomously.

4 **Cost Intelligence**

Continuous optimization of token usage, caching and batching. Your AI costs decrease while quality increases.

AI READINESS

Your Orchestration Readiness Score

How mature is your LLM architecture? The AI Scan™ gives the answer.

Exploration — 10 – 20

Single LLM via API. No orchestration. Hard-coded prompts. No cost monitoring or performance tracking.

Integration — 20 – 35

Multiple models in use. First RAG pipeline. Prompt templates. Costs being tracked.

Strategic AI — 35 – 50

Intelligent multi-model orchestration. Self-optimizing routing. Enterprise-grade RAG. Fully automated cost-performance balance.

Start your AI Scan™

Discover the orchestration maturity of your LLM landscape.

w69.ai/en.html#scan

W69 PROOF™

✓ Enterprise-Grade

✓ GDPR Compliant

✓ NDA-Protected

Every system W69 delivers carries the W69 Proof™ seal — your guarantee for architecture, performance and measurable results.

What you receive:

Strategic Assessment · Architecture Roadmap · Implementation Plan

■ **We respond within 24 hours**

“We don't build for one model, but for a future-proof orchestra of AI intelligence.”

NEXT STEPS

Ready to get started?

Scan for AI Scan™

1. **Scan the QR code**
   Or go to w69.ai/en.html#scan
2. **Complete the AI Scan™**
   5 minutes, no registration required
3. **Receive your AI Readiness Report**
   Including personalized recommendations
4. **Strategic conversation**
   We discuss your results and opportunities
5. **Implementation roadmap**
   Concrete steps toward enterprise AI

W69 AI Consultancy

Enterprise AI Architecture & Strategy
LLM Orchestration & Integration

w69.ai · hello@w69.ai · Amstelveen

+31 6 2797 3800