AI at Scale: Cost-Efficient AI for Development, Agents and Enterprise Automation

Leading models, compatible APIs and optimized costs to scale enterprise AI usage without losing control of token budgets.

KOMETASOFT + HUAWEI CLOUD

Enterprise AI has moved from pilots to massive consumption

Companies are integrating AI into development, operations, support, document analysis, automation and agents. Token usage is no longer marginal; it is becoming a recurring IT cost.

1. Individual use	Prompting, basic copilots and personal assistance.
2. Development teams	Code assistants, refactoring, test generation and documentation.
3. Business processes	Automation, document analysis, extraction and classification.
4. Automation at scale	Agents, multi-agent systems, enterprise batch workloads and connected workflows.

Token cost is becoming a critical IT budget line

Consumption grows non-linearly when autonomous agents, long contexts, complete repositories and API-connected tools enter the workflow.

Agentic AI does not consume tokens like a chatbot. It consumes tokens like infrastructure.

Long contexts	Complete repositories, long documents and working memory.	**Agentic iterations**	Reasoning, tool calls, error correction and retries.
Multi-agent	Additional multiplier from orchestration and coordination.	**Massive batch**	RAG, classification, document processing and technical translation.

Not every AI use case needs the most expensive model

AI maturity is not about always using the premium model. It is about knowing when it is not needed and routing each workload to the right model.

High sensitivity + low volume	Controlled premium model.	**High sensitivity + high volume**	Specific architecture, private model or dedicated environment.
Low sensitivity + low volume	Flexible use depending on quality and availability.	**Low sensitivity + high volume**	Cost-efficient MaaS, batch and optimized models.

Priority entry points: high usage, low risk and fast ROI

The natural starting point is not personal data or critical processes. It is high-value, low-risk work with clear measurement.

Development and software engineering	Code generation and review, refactoring, legacy migration, tests, technical documentation and DevOps scripts.
Technical agents	Development, QA, documentation, internal support and non-sensitive tool-connected agents.
Anonymized data and repositories	Classification, extraction, summarization, technical translation, code analysis and documentation generation.

Huawei Cloud MaaS: leading models as a managed service

MaaS removes the complexity of deploying GPUs, managing clusters and maintaining models. The client consumes intelligence through APIs, measures results and compares real costs.

- Access to leading LLMs and coding models such as GLM, DeepSeek and others.
- Token-based consumption with compatible APIs.
- Managed platform without operating proprietary infrastructure.
- Integration with development tools, agents and workflows.
- Options for tuning, evaluation and inference.

User / IDE / Agent	Compatible API	Huawei Cloud MaaS	Optimized inference

Designed to integrate with the current ecosystem

Adoption does not require rebuilding the full technology stack. If a tool allows a compatible endpoint and API key, a POC can start in days.

IDEs and coding assistants	VS Code + Cline, Cursor, CodeArts.
Automation and agents	Dify, n8n, WorkBuddy, OpenClaw / NanoClaw and similar tools.
Compatible APIs	OpenAI API compatibility and Anthropic compatibility in selected scenarios through base URL + API key.

The right model for each workload

Optimization is not just about lowering price. It is about choosing the right model for the task, context, reasoning need, latency and token cost.

Workload	Recommended model	Main reason
Complex development and agents	GLM 5 / GLM 5.2	Coding, long context, agents, refactoring.
Development agents	GLM 5.2 / DeepSeek V3.2	Balance between capability and cost.
Routine coding tasks	DeepSeek V3.2 / V3	Cost-efficient automation and scripting.
Document processing	DeepSeek V3.2 / V4-Flash	Extraction, summary, translation and batch.
Massive batch	DeepSeek / GLM depending on case	Throughput and cost per million tokens.

Cost efficiency: meaningful savings versus premium models

When usage moves from thousands to hundreds of millions of tokens, token price becomes a strategic variable, not a technical detail.

Model	Indicative output price per 1M tokens
Claude Opus 4.6	$25
GPT-4 Advanced	$22.5
GLM 5.1 Huawei Cloud	$3.77
DeepSeek R1 Huawei Cloud	$2.16
DeepSeek V3.2 Huawei Cloud	$0.40

Indicative values subject to availability, region, commercial terms and provider changes. Final analysis must use current prices and real workloads.

Example: monthly impact in a development team

A team adopting agentic AI intensively can generate hundreds or thousands of millions of tokens per month. The economics change completely.

Indicative scenario	25 developers · 20 working days · 4 agentic sessions per developer per day · 500,000 tokens per session · estimated consumption: ~1,000M tokens/month.
Expected outcome	Compare current cost against GLM, DeepSeek or other available models, measuring quality, acceptance, retries and cost per completed task.

The advantage is not only price: it is architecture

Sustainable savings do not come from swapping one API for another without control. They come from designing architecture, governance and observability before calling the model.

Applications and agents	Real business tools, IDEs, pipelines, documents, workflows and internal users.
AI Gateway	Provider management, keys, policies, consumption, traceability, rate limiting and observability.
Security and sanitization	PII detection, secret masking, prompt validation and workload controls.
Intelligent routing	Premium for critical cases, MaaS for cost efficiency, local or private for sensitive data.

No-cost POC: measure performance, quality and savings

The best way to validate AI at Scale is to test a real, bounded and measurable workload. We do not ask the client to believe a promise; we propose measuring their own use cases.

Week 1 - Discovery and security	Use-case selection, data cleanup, removal of PII and secrets, prompt and metric definition.
Week 2 - Integration and tests	MaaS setup, comparison with the current model, quality, latency and cost-per-task measurement.
Week 3 - Business case	Estimated monthly usage, current cost versus optimized cost, monthly and annual savings, risks and recommendation.

Three high-impact POCs to start

Concrete, measurable POCs connected to real costs, with clear value and avoiding personal data where possible.

AI Coding Cost Optimization	Reduce AI development tooling cost while maintaining productivity. Metrics: cost per task, output quality and acceptance rate.
Repository Analysis Agent	Analyze repositories or complete modules without using premium models for everything. Metrics: cost per repository and technical quality of the report.
Anonymized Document Processing	Process business documentation without PII at lower cost. Metrics: cost per document, precision and savings versus current model.

How we calculate real savings

Savings must become cost per task, cost per team and tangible annual savings. The key metric is not only price per million tokens; it is cost per correctly completed task.

Consumption	Input/output tokens, iterations per task, users and frequency.	**Performance**	Latency, TTFR, tokens/second, errors, retries and stability.
Quality	Acceptance rate, human review, precision and rework.	**Economics**	Cost/task, cost/user, monthly and annual cost, savings potential and payback.

Reference formula: real cost per task = tokens + retries + human review + integration + operations.

From POC to an AI at Scale strategy

The goal is not an isolated test. It is to build a sustainable, governed and cost-efficient AI architecture that can scale with the business.

1. Discovery	Inventory of AI tools, models, costs, use cases and constraints.
2. POC	Workload selection, fast integration, anonymization and comparative measurement.
3. Business case	Monthly and annual savings, recommendation by model, workload map and risks.
4. Controlled production	AI Gateway, usage policies, spend control, monitoring and data governance.
5. Scaling	More teams, more agents, batch processing, DevOps/document integration and AI FinOps.

Kometasoft + Huawei Cloud: technology, architecture and enterprise adoption

Huawei provides MaaS capacity and infrastructure. Kometasoft turns that capacity into architecture, integration, measurement and real business cases.

Huawei Cloud contributes	MaaS platform, leading models, optimized AI infrastructure, compatible APIs, competitive cost and global cloud capacity.
Kometasoft contributes	Use-case identification, AI Gateway architecture, anonymization, integration with existing stacks, cost measurement and support from POC to production.

The client is not only buying tokens. The client is buying a way to adopt AI with lower risk, clear numbers and a partner that grounds technology in real measurable cases.

Let us identify a real workload and measure the savings

The opportunity does not require a massive migration or an immediate strategic decision. It requires starting with a well-chosen POC, executing it in 2-3 weeks and deciding with real data.

1. Select workload	Assisted development, technical agent, concrete repository, anonymized documentation or batch process.
2. Execute POC	Model comparison, estimated real cost, output quality, latency and performance.
3. Decide with data	Monthly/annual savings, identified risks, adoption recommendation and integration roadmap.

AI at Scale is not about using cheap AI. It is about using AI intelligently, sustainably and ready to scale.