



The rapid adoption of AI coding assistants is transforming software development. At the same time, the industry is shifting from seat-based to usage-based pricing, causing AI inference costs to grow significantly — coding sessions run longer, requests become more expensive, and premium models are frequently used for routine development tasks. **Engineering organizations need a way to control AI costs while preserving developer productivity and existing development workflows.**

## SOLUTION OVERVIEW

MajorDomo optimizes AI requests generated by coding assistants before they are processed by AI models. The platform combines AI Request Optimization, Intelligent Model Selection, and Enterprise Governance to reduce AI costs while maintaining the developer experience. Organizations can achieve up to **50–60% reduction** in AI coding costs, depending on workload characteristics, coding patterns, and model usage. **No application changes. No developer workflow changes.**

## KEY CAPABILITIES

### › AI Request Optimization

Optimize AI requests to improve efficiency, reduce AI costs, and deliver a seamless developer experience.

### › Coding-Agent Agnostic

Works seamlessly with leading AI coding assistants — Claude Code, GitHub Copilot, Cursor, Codex, OpenCode, Cline, and other OpenAI-compatible coding assistants.

### › Flexible Deployment

Deploy in SaaS, Private Cloud, On-Premises, or Air-Gapped environments to meet enterprise security, compliance, and operational requirements.

### › Intelligent Model Selection

Automatically select the most appropriate AI model for each coding task to balance capability, latency, and cost.

### › Enterprise Governance

Centralize AI policies, usage visibility, cost analytics, and governance across engineering teams to ensure secure, compliant, and efficient AI adoption.

### › Zero Workflow Changes

Integrate seamlessly into existing development environments — developers continue using their preferred coding assistants without changing tools or workflows.

## HOW IT WORKS



## BENEFITS

### Engineering Teams

- Continue using existing coding assistants
- Reduce AI coding costs without disrupting developer workflows
- Improve efficiency during long-running coding sessions
- Support multiple coding assistants through a common optimization platform

### Platform Engineering

- Centralized AI governance and policy enforcement
- Enterprise-wide visibility into AI usage and costs
- Optimize AI model utilization across engineering teams
- Apply consistent optimization policies across coding assistants

### Business Leadership

- Up to 50–60% reduction in AI coding costs
- Predictable and governable AI spending
- Better utilization of enterprise AI investments
- Enterprise-ready deployment for large-scale engineering organizations

## AT A GLANCE

| METRIC                      | VALUE                                                              |
|-----------------------------|--------------------------------------------------------------------|
| AI Coding Cost Reduction    | Up to 50–60%                                                       |
| Developer Workflow Changes  | None                                                               |
| Supported Coding Assistants | Claude Code, GitHub Copilot, Cursor, Codex, OpenCode, Cline & More |
| Deployment Options          | SaaS, Private Cloud, On-Premises, Air-Gapped                       |