

AI Agents: Ongoing Costs Scale with Tasks, Not Users

MANAGEMENT SUMMARY & CORE TAKEAWAY

AI agents don't scale per user—they scale per completed task. Inference costs rise with every agent task, while maintenance, monitoring, and compliance often make up two to four times the build cost over 24 months. The operating curve is the real lever, and it can be made predictable.

CONTEXT & EXECUTIVE RELEVANCE

Classic software costs scale with the number of users. AI agents work differently: once an agent independently runs research, checks documents, or takes over a process, compute is consumed for every single execution. Altimeter Capital projects installed AI compute capacity to rise from roughly 18 gigawatts in 2025 to around 115 gigawatts in 2028—with the largest jump of about 43 gigawatts in 2027 alone. For your budget, that means the operating cost curve follows completed tasks, not the number of logged-in employees.

Inference—the actual execution of a model—already accounts for about two-thirds of all AI compute, according to Bloomberg Intelligence via ClearML, and continues to grow. At the same time, the unit price per inference has dropped by 99 percent over the past two and a half years. That sounds like relief, but it is a trap: inference-time reasoning has driven token consumption parabolic. You buy the unit cheaper, but your agents consume far more of it. Total cost becomes a function of workload.

SWISS COMPLIANCE & ARCHITECTURE GUIDELINES

- ☑ Full compliance with revised Swiss Data Protection Act (revDSG).
- ☑ Zero vendor lock-in on closed proprietary models.
- ☑ Seamless integration into Abacus, Bexio and M365 ecosystems.

Non-Binding Process Analysis

Let us review in 20 minutes where AI agents create immediate measurable efficiency in your workflows.

ki-outsourcing.ch