STRATEGIC GUIDE

Updated: August 29, 2026 · Zürich, CH

Model Routing, Not Model Selection: The Real AI Cost Lever

MANAGEMENT SUMMARY & CORE TAKEAWAY

Model routing means assigning the cheapest model capable of reliably handling each task—not automatically the strongest one. Simple tasks run on small models like Haiku, most cases on a workhorse model like Sonnet, and only the hardest five percent need premium models like Opus. This typically cuts costs by 60 to 80 percent.

CONTEXT & EXECUTIVE RELEVANCE

Model routing means deliberately assigning the cheapest model capable of reliably handling each task—rather than reflexively sending every request to the most powerful and most expensive option available. For simple, structured tasks such as classification or data extraction, a small model is almost always sufficient; only the hardest five percent of cases justify a premium model. Applied correctly, this approach cuts AI costs by 60 to 80 percent without any noticeable drop in quality.

Many organisations pick a single premium model and route every request through it—from sorting routine emails to analysing complex contracts. The instinct is understandable: the strongest model makes the fewest mistakes, and nobody wants to be blamed for a compromise. But it's the equivalent of taking a sports car to buy groceries. The car handles the trip effortlessly—but a compact car would have done the job just as well, at a fraction of the cost.

© 2026 KI-Outsourcing.ch · Digital Passion GmbH

CORE ACTION ITEMS

01 Entry tier (Haiku, GPT-4o-mini): classification, tagging, data extraction, simple summarisation—tasks with clear structure and little room for interpretation.

02 Workhorse tier (Sonnet, GPT-4o): the default for most business processes—customer communication, research, drafting, mid-complexity code review.

03 Premium escalation (Opus, o1/o3): the demanding five percent—multi-step reasoning, high error costs, strategic decisions, complex planning.

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