

Model-Routing statt Modell-Wahl: So sparst du bis zu 80% KI-Kosten

MANAGEMENT SUMMARY & KERNAUSSAGE

Model-Routing bedeutet, für jede Aufgabe das günstigste Modell zu wählen, das sie zuverlässig löst – nicht automatisch das stärkste. Einfache Aufgaben laufen über kleine Modelle wie Haiku, die meisten Fälle über ein Arbeitspferd-Modell wie Sonnet, nur die schwierigsten 5 Prozent brauchen Premium-Modelle wie Opus. Das senkt Kosten um 60 bis 80 Prozent.

KONTEXT & RELEVANZ FÜR ENTSCHEIDER

Model-Routing bedeutet, für jede Aufgabe bewusst das günstigste Modell zu wählen, das sie zuverlässig löst – statt reflexhaft immer zur stärksten und teuersten Option zu greifen. Für einfache, strukturierte Aufgaben wie Klassifikation oder Datenextraktion reicht ein kleines Modell fast immer aus; nur die anspruchsvollsten 5 Prozent der Fälle rechtfertigen ein Premium-Modell. Richtig umgesetzt, senkt diese Strategie die KI-Kosten um 60 bis 80 Prozent, ohne bei der Qualität spürbare Abstriche zu machen.

Viele Unternehmen entscheiden sich für ein einziges Premium-Modell und schicken jede Anfrage dorthin – vom simplen E-Mail-Sortieren bis zur komplexen Vertragsanalyse. Das ist verständlich: Das stärkste Modell macht am wenigsten Fehler, und niemand will für einen Kompromiss verantwortlich sein. Nur: Es ist, wie einen Ferrari zu nehmen, um Milch einzukaufen. Der Wagen schafft die Fahrt mühelos – aber für die Aufgabe war ein Kleinwagen völlig ausreichend, und deutlich günstiger.

KERN-HANDLUNGSEMPFEHLUNGEN

- 01 Einstiegsstufe (Haiku, GPT-4o-mini): Klassifikation, Tagging, Datenextraktion, einfache Zusammenfassungen – Aufgaben mit klarer Struktur und wenig Interpretationsspielraum.
- 02 Arbeitspferd (Sonnet, GPT-4o): der Standardfall für die meisten Geschäftsprozesse – Kundenkommunikation, Recherche, Textentwürfe, Code-Reviews mittlerer Komplexität.
- 03 Premium-Eskalation (Opus, o1/o3): die anspruchsvollen 5 Prozent – mehrstufiges Denken, hohe Fehlerkosten, strategische Entscheidungen, komplexe Planung.

SCHWEIZER UMSETZUNGSKRITERIEN (REVD SG & IT-ARCHITEKTUR)

- ☑ Vollständige Konformität mit dem neuen Schweizer Datenschutzgesetz (revDSG).
- ☑ Keine Abhängigkeiten: Zero Lock-In auf proprietäre geschlossene Modelle.
- ☑ Schnittstellen zu Schweizer ERPs (Abacus, Bexio, M365) im operativen Betrieb.

Unverbindliche Prozess-Analyse

Lass uns in 20 Minuten prüfen, wo KI-Agenten in deinem Unternehmen sofort messbare Entlastung schaffen.

ki-outsourcing.ch