How to evaluate costs in relation to Agentic AI solutions ?

Dec 2025

Consulting | Digital Finance | Agentic AI

Agenda

1 Introduction

2 Cost drivers

3 Best Practices

4 Conclusion

5 Contacts

Complexity and costs of implementing AI solutions are unknown to most people.

All seems as simple as inputting a prompt and getting outcomes presented in seconds freely.

But the reality is quite different.

AI models			Price per API call for 1M Input/Output tokens	Price difference
	Open AI	GPT-5.1 GPT-5-nano	10.1000$ 0.4050$	25X vs 5-nano
	Anthropic	claude-opus-4.5 claude-haiku-4.5	30.0000$ 6.0000$	5X claude-haiku-4.5, **80X** vs gemini-2.5-flash
	Google	gemini-3-pro-review gemini-2.5-flash	14.0000$ 0.3750$	37X vs gemini-2.5-flash
	Deep Seek	deepseek-chat deepseek-reasoner	2.2400$ 0.7000$	3X vs deepseek-reasoner

Costs of sophisticated vs cheaper AI models can be 80X more expensive

For simple searches, expensive and cheaper AI models can give similar results.

For complex, specialized tasks, expensive AI models are required.

Increases in API calls rise costs enormously, especially when using AI agents.

The more input/output tokens, API calls, the higher the costs, so engineer smartly.

It's key to engineer Agentic AI solutions, using most optimal LLM selections, finding a sweet spot in costs, accuracy & speed

Agenda

1 Introduction

2 Cost drivers

3 Best Practices

4 Conclusion

5 Contacts

For each Agentic AI project, consider all costs per workflow and ROI:

- **Hardware:**
  - Storage cloud or on-premise, GPU/CPU instances
- **LLMs / Agents / APIs / LLM Ops Platforms:**
  - Own model builds, or costs per LLM / Agents (costs per input /output tokens, number of API calls).
  - Enterprise licenses or per seat fees (hosted platforms)
  - 3rd party APIs (e.g. RAG or "Retrieval Augmented generation" to fetch external data)
  - Vector database (or memory to store, index and retrieve semantic meaning and past interactions)
- **Data requirements:**
  - Data cleaning, data integration
- **Regulatory, ethical, data protection compliance:**
  - A governance framework is extremely important to check, monitor results, security and compliance
- **Humans:**
  - Defining scope, timeframe, committed costs/benefits, openly communicated / aligned with strategy
  - Accountability: define clear responsibility by IT & the business. Ensure models are understood.
  - Adoption: start with AI use-cases that bring concrete benefits. Though, avoid business critical tasks.
  - Design, analysis, agent set-up, maintenance, monitoring, training.
  - Change management: The more processes change, the higher the risk of not realizing promised ROI.
- **Environmental costs (energy consumption & other environmental costs)**

And consider risks

Include 20-30% safety margins for testing, explorations…

Agenda

1 Introduction

2 Cost drivers

3 **Best Practices**

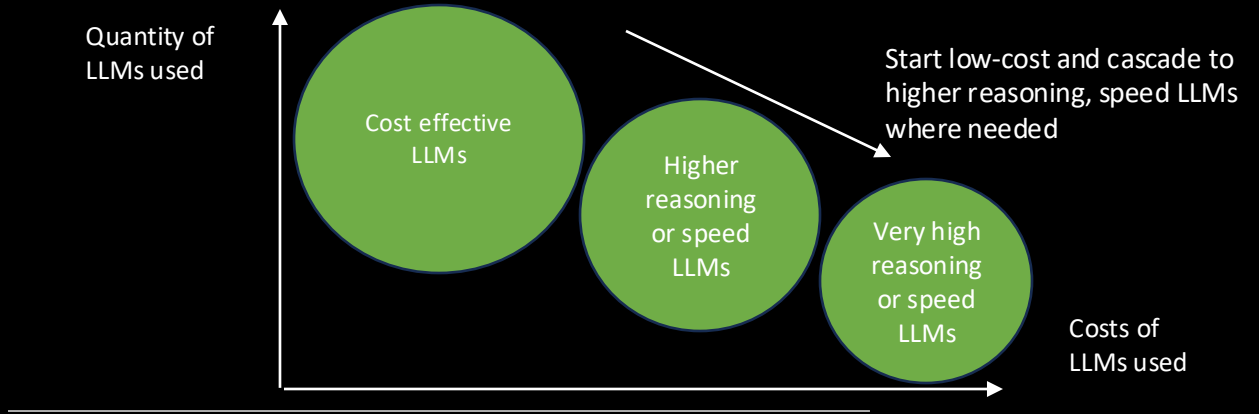
4 Conclusion

5 Contacts

When engineering your Agentic AI solution, think about which LLMs to use.

- **Minimum viable cost models** that attain your goal can help **save huge costs.**
  - Executing a prompt search (188 input tokens, 2426 output tokens) by 1000 persons, once per day for one year (365'000 API calls): Costs vary between:
    - 323.18$ (Open AI gpt-5-nano) a more cost-efficient LLM
    - 22'571.6$ (Anthropic claude-opus-4.5), a higher reasoning LLM

    But expensive higher reasoning LLM might not always generate better outcomes.
- Monolithic, expensive "high reasoning" LLMs at scale can jeopardize ROI.
- **Multi-model LLMs** at scale are proven to be **superior and with lower costs**
- **Check first for cost-effective LLMs, cascade up when speed / accuracy is key.**



Scaling will significantly increase costs.

Start cost-effective, scale for value.

Agenda

1 Introduction

2 Cost drivers

3 **Best Practices**

4 Conclusion

5 Contacts

- Optimise costs by **using efficient, lower-cost priced models for routine queries**, where performance is comparable to premium options.
- Ensure accuracy & reliability by serving **advanced, higher cost models for complex, specialised or high-stake tasks.**
- **Control AI spend by monitoring API usage** – especially in agent-based workflows – where call volumes can grow rapidly. Tracking real usage – e.g. avoiding unnecessary long prompts, limiting unused licenses and controlling call volumes **can reduce unplanned costs by up to 40%** and significantly improve ROI delivery.
- **Governance** is essential to ensure quality (e.g. avoiding hallucinations), avoiding costly pitfalls or reputational damages.
- **Stress-test the business case**. AI models carry inherent uncertainties.

  Running "what-if" scenarios - e.g. API volume spikes, accuracy drops or higher-then-planned model usage – is crucial to validate assumptions, anticipate risk, and secure a resilient ROI-positive investment case.
- **Start** with Agentic AI solutions that bring **concrete benefits** but avoid business-critical tasks.
- Do not under-estimate **Change Management.**

Stress-test with "what-if"

And adopt with concrete winning benefits

Agenda

- Design & selection of LLMs makes or breaks your ROI.
- There is a huge risk on the P&L bottom line, as scaling can explode variable costs with Agentic AI solutions.
- Define a clear approach to estimate and track costs and ROIs
  - Select and prioritize use-cases and agent workflows
  - For each use-case, classify key requirements (costs, accuracy, speed)
  - Map MML options with strengths and weaknesses.
  - Estimate costs, based on input/output tokens, number of API calls.
  - Cascade to more expensive models, when higher accuracy levels or speed are required.
  - Perform sensitivity analysis determining break-even points on ROI.
  - More expensive models are justified when going from break-even to ROI.
  - Monitor usage, actual costs and benefits attained.
- Monolithic LLMs can be quick to implement, but…
- Hybrid Multi Agent LLMs can become far more superior in performance & lower costs.

Keep a close eye on costs & value.

The coin shifts when moving from cost to value optimization.

Agenda

1 Introduction

2 Cost drivers

3 Best Practices

4 Conclusion

5 **Contacts**

Nathalie Ruttia-Pedretti

Founder, SUVERT AG

nathalie_ruttia@yahoo.com

Cristina Berges

Advisor, SUVERT AG

cbergesdelamo@gmail.com

"Develop a passion for learning. If you do, you will never cease to grow."

(Anthony d'Angelo)