Trim overgrowing AI spending and decide what earns its budget

The first wave of AI enthusiasm is over, and the bills have arrived. Most companies now face a dilemma: cut back and risk losing real value or keep paying without knowing what for. But there is a better way. In this paper we discuss three lessons on how to approach your AI spend.

A rapidly growing garden of AI applications

It has only been four years since OpenAI pioneered the commercial breakthrough of LLMs with the release of ChatGPT to the public. Driven by the enormous potential to reshape our ways of working, AI has seen rapid adoption since then. In this early phase of our new AI world, use cases were often hastily explored to gain quick results and not miss out on the biggest opportunity in decades. The fear of missing out was real. Either you invested heavily, or you'd miss the party. This sentiment led to where we are today: many ongoing initiatives with little oversight of their outcomes. Following this initial build-up, the bills start coming in. Gartner forecasts worldwide spending on generative AI models to rise from approximately $13 billion in 2025 to $51.6 billion in 2027, nearly quadrupling within two years.[1] Naturally, this leaves leadership asking what they are getting in return, and they now must decide what earns its budget.

This paper proposes a value-driven approach to curb cost overruns and implement self-optimizing cost governance, without slowing down real value generation.

Understand where AI costs arise

To address AI costs, we first need to understand where they arise. AI spending spans four areas, but they differ widely in how fast they grow and how likely they are to cause waste. Figure 1 narrows them down step by step to where action is most urgent.

Vertical tools are usually part of larger AI initiatives with clear business ownership and a vetted business case, which makes waste less likely.[2] Horizontal tools are essentially the opposite: It is hard to evaluate what they are used for exactly or how effectively they deliver results. Additionally, a lack of cost ownership among end users hinders usage optimization in relation to

[1] Gartner, "Gartner Forecasts Worldwide AI Spending to Grow 49.5% in 2026", press release, 16 September 2026, https://www.gartner.com/en/newsroom/press-releases/2026-09-16-gartner-forecasts-worldwide-ai-spending-to-grow-49-point-5-percent-in-2026

[2] Assuming a broader, suitable, value-driven AI Operating Model within the organization

© 2026, AdEx Beratungs GmbH

the value generated. To resolve this tension, we focus our discussion of AI expenditure on where costs grow the fastest and where the likelihood of waste is highest: consumption-based costs for horizontal tools.

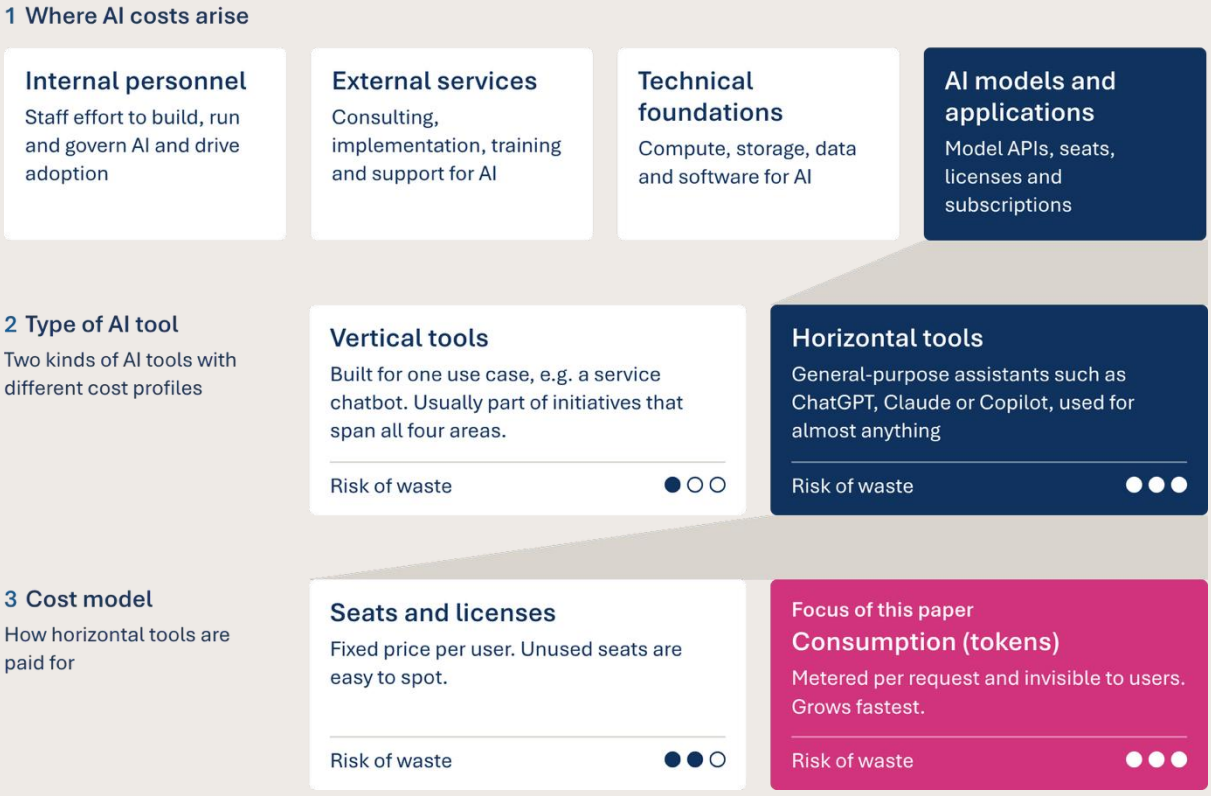


Figure 1: Where AI costs arise and where this paper focuses

Decide what earns its budget

Consider a company that has made a general-purpose AI assistant widely available. Employees are using it, token consumption is growing, and costs are rising. Leadership now faces the question introduced at the beginning: What are we getting in return, and how can we control spending without getting in the way of valuable work? Four years of working with clients on their AI-related challenges taught us three lessons that stand out, each building on the one before.

Lesson 1: Make the cost visible

For an employee submitting a request, the cost is usually invisible. Without knowing the economic differences between available models and processing options, there is little reason to consider whether a less expensive option would do the job. Making those differences understandable and easily accessible, together with practical guidance on when additional capability is worthwhile, gives employees a basis for more deliberate choices. This combination of transparency and enablement helps employees use AI more economically, even without personally paying the bill.

People can only spend wisely what they can see.

Lesson 2: Cap the few, free the many

© 2026, AdEx Beratungs GmbH

Awareness alone gives the company no reliable spending ceiling, so the obvious next step is a user-based spending limit, which raises two immediate questions: What is the right limit? And how do we make sure it does not hold back valuable work? To answer the first, we need to look at how token usage is distributed across the organization. In our studies we have found that consumption across users of general-purpose assistants follows the Pareto principle: The top 20% of users account for roughly 80% of token consumption.

This gives us a practical starting point. If we set the initial allowance at the consumption level of the 80th percentile, only the top 20% of users (those accounting for 80% of spending) will ever reach it, while the bulk of the organization is not affected at all. The same pattern shows where the question of value is worth asking at all. Most everyday work, such as researching a topic, drafting a text or summarizing a document, uses few tokens and costs very little. For this kind of work, it is hardly worth asking whether it pays off, because it almost always does. The real cost sits in the token-heavy usage of a few, such as long agentic tasks, very large documents or premium models running at full depth. That is where the limit takes effect.

> ***A well-placed limit leaves most employees untouched and focuses attention on the few who drive most of the cost.***

Lesson 3: Let those who benefit decide

That leaves the second question: What happens to those who reach the limit? An employee might run out of their allowance halfway through an important task, interrupting work that would easily justify the additional expense. Having capped consumption, we now need to decide which work deserves more room.

Different allowances for different roles or departments may seem like the next logical step. Yet who determines how much each group needs, and who adjusts the allocation when tasks change? A central decision maker would need detailed knowledge of everyday work to make those decisions well. Department leaders are far more familiar with the tasks, priorities and expected benefits. Consequently, this is where we propose placing responsibility for additional spending.

The company centrally funds a common baseline, initially guided by the threshold described above. Departments can then fund justified additional capacity from their own budgets, subject to centrally agreed rules. Reaching the baseline therefore triggers a very specific business decision: What will this additional spending enable, and is it worth funding alongside our other priorities? The department leader responsible can approve additional capacity and later review the result.

The central function sets the framework and provides cost visibility. Departments assess additional needs and take responsibility for funding them. Simply assigning consumption to a cost center achieves little unless the people responsible also have the authority to make these decisions.

> ***Spending decisions belong with those who can judge the value.***

Beyond tokens: what truly earns its budget

© 2026, AdEx Beratungs GmbH

This paper is not a full manual on how to handle the end-to-end costs of AI in an organization. It provides a starting point on how to tackle the rapidly increasing costs that widespread AI use incurs, but without restricting valuable usage. Our three lessons build on each other: make the cost visible, cap the few to free the many, and let those who benefit decide on additional spending.

However, as described in the beginning, token spending remains only one part of the bigger picture. A complete assessment, one that shows the true cost of each initiative, needs to consider all four cost areas discussed, including the spend for internal personnel, external services and technical foundations. Spending less on tokens offers little benefit if achieving it creates greater effort elsewhere. Governing the token spend is a fast first win, but the bigger lever is judging each initiative by the value it creates relative to its full cost. Achieving a full understanding of the return on investment regarding AI requires more than the cost side. It requires a clear view of how AI contributes to the generation of business value in the organization. The journey to get there starts with a business-anchored AI strategy that actively contributes to the long-term vision. Furthermore, it requires a suitable operating model, a product-focused use-case delivery engine, and finally, value and portfolio management that truly enables ROI-based decisions.

This is where AdEx helps. We combine deep AI expertise with years of experience in product and value management, from high-level strategy to hands-on implementation. Whether you want to get your token spend under control today or build the operating model and portfolio management that decide tomorrow's AI budget, let's talk!



Figure 2: Three lessons for governing AI spend, and the next step beyond tokens (illustrative data)

© 2026, AdEx Beratungs GmbH

AdEx Partners | ADvisory EXcellence

Your Contact



Dr. Michael Holzapfel
Partner | AdEx Strategy
michael.holzapfel@adexpartners.com



Dr. Philipp Strietzel
Senior Manager | AdEx Intelligence
philipp.strietzel@adexpartners.com

About AdEx Partners

As a Trusted Advisor for IT strategy, digital transformation and AI, AdEx Partners supports leading companies in shaping Europe's digital future. AdEx Partners was founded in Hamburg in 2012 and now operates throughout the DACH region. The team of around 350 consultants with an average of 13 years of industry experience implements tailor-made solutions in the consulting areas of ERP, Data Value Management, AI and Business Transformation.

The services of AdEx Partners are not limited to strategic consulting and project planning, but the Trusted Advisors also remain involved during implementation and contribute to the successful implementation of complex transformation programs. More than 60 percent of DAX40 companies from various industries, including pharmaceuticals, healthcare, industry, trade and financial services, rely on this comprehensive expertise.

With CONSULTING AG, we have further strengthened our presence in Austria and expanded our industry expertise in the public sector and healthcare.

AdEx Partners has been named „Hidden Champion“ four times in a row by the business magazine „Capital“, has received seven „Best of Consulting“ awards from WirtschaftsWoche and has been awarded the „Great Place to Work®“ seal of approval for years.

www.adexpartners.com

© 2026, AdEx Beratungs GmbH