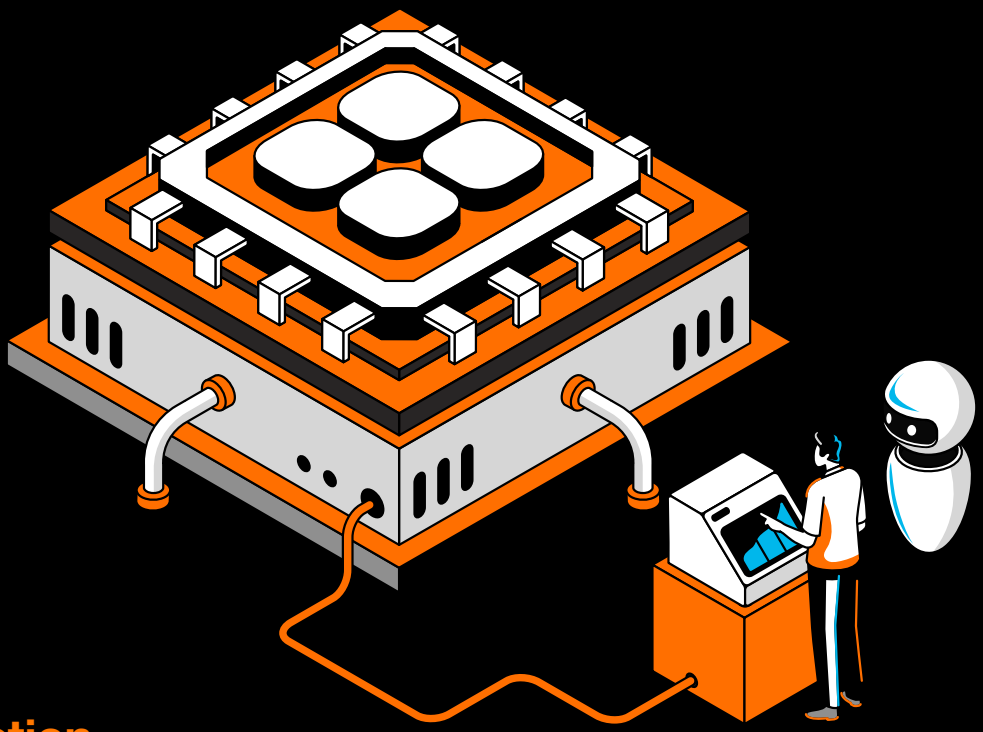
Business

Artificial Intelligence.
Real Wisdom.

Tomas's COO wants a unified approach to Gen AI. But Tomas knows a hybrid strategy is more cost-effective. Find out here what he and other Orange Business customers know.

A hungry mouth to feed:

Addressing the skyrocketing costs of AI services



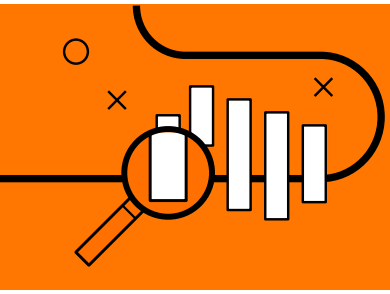
Introduction

Most GenAI Proof of Concept (PoC) projects start on public clouds like AWS, Microsoft Azure, and the Google Cloud Platform – this is the simplest and most logical place to host a fledgling service due to the readily available computational power, scalability, and access to advanced AI models offered by these platforms.

However, as you operationalize the PoC by scaling the GenAI use case into a production-grade service, you add users and data, and the amount of processing power necessary to service the needs of your AI model increases. And, as you learn more about your GenAI model, it will need to be retrained and fine-tuned – each step of which requires more compute power. As a result, your costs can start to climb very steeply.

It's also true that GenAI's ease of use is both a blessing and a curse. Its natural language interface means you don't have to be a data scientist or an AI engineer to access the technology. However, GenAI is the most compute-intensive and, therefore, expensive resource in your data center, so the fact that everybody can use it doesn't mean that everybody should.

At this point, smart leaders start to wonder if there is a more cost-effective way of delivering the same outcome. Leaders at small and mid-size companies will be wondering if they can do anything at all with GenAI on the budgets they have available. At Orange Business, we know that, at every stage of your GenAI journey, there are moves you can make that will make a huge difference on the final cost of your service. This will increase value delivery for large companies and bring GenAI services within reach of everyone else.



First of all, you must start with a careful evaluation of the use case: not every application demands GenAI as the answer; traditional AI and even data visualization technologies may be able to do the same job at a lower cost. Your GenAI service doesn't need to search every piece of data in the model for every query. You can generate huge efficiencies by training your users to be more efficient with their prompts. Finally, adopting a FinOps approach will provide the visibility you need to monitor your cloud spend and switch off unnecessary services.

1

Applying GenAI to the right use cases – identify where it can add the most value

As the most talked about – and potentially disruptive – technology since the smartphone, there is a certain excitement that applies to the use of GenAI to solve problems. But, while there is a temptation to use the technology to draft a two-line email, you must ask yourself if this would deliver any value.

The point about GenAI is that it generates – or infers – new outputs from existing data. Gartner defines inference as the process where an AI model applies the knowledge it is trained on to new, unseen data, generating outputs like text, images, or code. It makes predictions or draws conclusions based on the input it receives, effectively putting the model 'in action' to produce a result. This process – and the use of GPUs to process the necessary data – has both a financial and a sustainability impact.

With the GPU market growing at over 30% per year , many organizations are struggling to access the processing power they need for their GenAI services. If that sounds familiar, you may be interested to know that you can rent GPUs on an 'as-a-service' basis from Orange Business, which is another way to rein in cloud costs. This service is available via our Cloud Avenue platform.

Not every use case requires this level of computational intensity or will deliver value through using it. The hype around GenAI tends to obscure the fact that other, earlier technologies - traditional AI or business intelligence (data visualization) – may deliver the same results at much lower costs. So, you must determine who really needs GenAI. That's why starting with the use case is essential.

To take an example from our own experience, Orange Business trialled Microsoft Copilot across its entire business and eventually decided that there wasn't sufficient value generation to justify implementing it company-wide. Instead, we deployed it for two use cases: first, for the functions that create and manipulate content and, second, for project managers who must provide a lot of coordination and meeting summaries.

Gartner recently claimed that at least 30% of generative AI (GenAI) projects will be abandoned after PoC by the end of 2025. This was for reasons such as poor data quality, inadequate risk controls, escalating costs or unclear business value . A typical GenAI PoC can cost anywhere between $15,000 and $50,000+ depending on the extent of the desired functionality, data requirements, and the level of development expertise needed: the financial and opportunity costs of such failures are therefore huge, so backing the right horse at the start of the race is critical. These were the use cases where we could demonstrate real impact and for which there was a readily calculable ROI.

Actionable takeaways

According to Gartner, there are two classes of GenAI: 'everyday GenAI' is focused on productivity and enables workers to do what they already do faster and more efficiently; while 'game-changing GenAI' is focused primarily on creativity and may well disrupt business models and entire industries.

If you want to rapidly access everyday GenAI, Orange Business has applied the learnings from its internal adoption of this technology to the creation of Live Intelligence: a multi-LLM offering available on a Software-as-a-Service basis that democratizes access to GenAI. Through a straightforward and intuitive interface, employees have access to a library of pre-set prompts, enabling them to address the most common everyday use cases. These include analyzing or summarizing a document, extracting important information from an email chain, writing meeting minutes, drafting an agenda, preparing interviews, or editing articles.

For game-changing use cases, Orange Business has a bespoke methodology that addresses, in a holistic manner, the key challenges facing GenAI operationalization, such as value creation, security, infrastructure modernization, governance and change management.

2

On-premise versus cloud – selecting the right hosting strategy for your GenAI use case

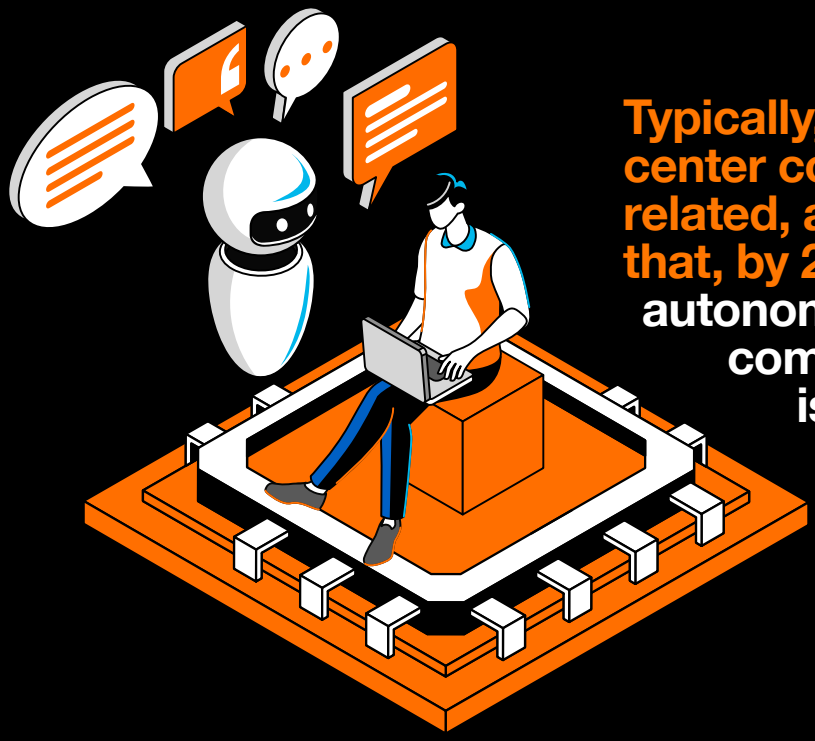
Starting with the use case will also help you determine whether you should host the GenAI service on the edge or in the cloud. The hyperscalers are pricing competitively for storage and compute power and, if cloud deployment is appropriate for your use case, this may be more cost-effective than building and managing your own infrastructure.

Generally speaking, edge deployment is recommended if there is a high requirement for real-time connectivity or if the basic infrastructure is either unavailable or unable to provide high-speed connectivity. For example, Orange Business is working on a safety-related AI use case for a mining company that demands real-time results and for which the underground location makes it hard to install an IT infrastructure, so this GenAI service is hosted at the edge. A use case with a high requirement for resilience and/or security may also suggest edge hosting – this would ensure that the service is available even if there is no or limited external connectivity and that sensitive data isn't transferred externally.

By contrast, if you have a Copilot use case in a well-connected office environment where there is no need for instantaneous results, then it makes perfect sense to host the service on the cloud. (You can learn more about this topic in our blog 'AI on the edge? Optimize performance and price by finding the perfect balance between cloud and on-premise hosting for your AI services'.)

However, one key question about this decision relates to compliance – and in particular to the issue of data sovereignty. Private clouds offer greater control over where data is stored, allowing you to physically locate your infrastructure within the specific country or region mandated by law. Public cloud providers may have data centers in various locations, but it can be complex to ensure that specific data remains within the required jurisdiction.

Data sovereignty legislation often favors private cloud solutions due to the greater control they offer over data location, access, security, and governance. However, the decision between private or public cloud also depends on other factors, such as cost, scalability, and specific business requirements. Organizations need to evaluate their needs and the implications of data sovereignty laws to choose the most appropriate cloud deployment model.



Typically, 80% of contact center costs are personnel-related, and Gartner claims that, by 2029, agentic AI will autonomously resolve 80% of common customer service issues without human intervention – this would lead to a 30% reduction in overall operational costs.

Once you have made this decision, you have a further choice concerning where you host the data, which will have a critical impact on your overall costs. Take, for example, a GenAI-based chatbot in your contact center. There are huge benefits to both the business and customers for implementing this use case. Typically, 80% of contact center costs are personnel-related, and Gartner claims that, by 2029, agentic AI will autonomously resolve 80% of common customer service issues without human intervention – this would lead to a 30% reduction in overall operational costs. Also, because chatbots are under no time pressure and have access to all the information they need, they can respond more fully to customer queries, increasing both customer satisfaction and NPS.

However, for the chatbot to provide the relevancy that customers demand, it must have access to the latest interactions with them. This information may be stored in different clouds and subject to egress charges: these are fees charged by cloud providers to move data out of their cloud and are in addition to storage and computer charges. These charges can mount up substantially.

One way to address this would be to move the data to the edge if this is appropriate. Alternatively, you might consider adapting your cloud strategy by creating a proxy of the customer interaction data and storing it on a private cloud that attracts no egress charges.

This strategy is part of a FinOps approach (see below) that provides complete transparency over your cloud costs.
As Orange Business works closely with all the leading cloud providers, customers often ask us to help them compare the cloud business models, manage the different stakeholders and optimize their cloud costs.

Actionable takeaways

You should focus on providing secure and reliable GenAI solutions with a strong emphasis on data sovereignty and compliance. Having a fixed notion of where you want to host your GenAI service – all on the cloud or all on-premise – is unlikely to deliver good results across all use cases. We recommend a hybrid approach that realizes the benefits of both hosting strategies through a clear decision matrix based on latency, data sensitivity, costs, and regulations. You should also build modularity into your architecture to future-proof the service and enable the straightforward integration of emerging technologies.

Orange Business's Live Intelligence solution provides a comprehensive framework that addresses these challenges while offering the flexibility to adapt to evolving needs and technological advancements. Regular monitoring and assessment of performance across environments remain essential.

Using the right language model for your use case

All GenAI language models are not created equally. What components are necessary, and which are the costliest?

Different large language models (LLMs) have different resource needs – some are more efficient than others. In one example, Orange Business evaluated the use of two different generations of the same LLM for a particular use case. What we found was that the results with each were the same, but the older version of the LLM was 90% cheaper than the new one. However, there is no 'rule of thumb' you can apply here – despite being released after ChatGPT-4, ChatGPT-4o Mini is the cheaper of the two models. (The same can be said for DeepSeek.) So, shop around. Ask yourself if you need the most powerful LLM for your use case.

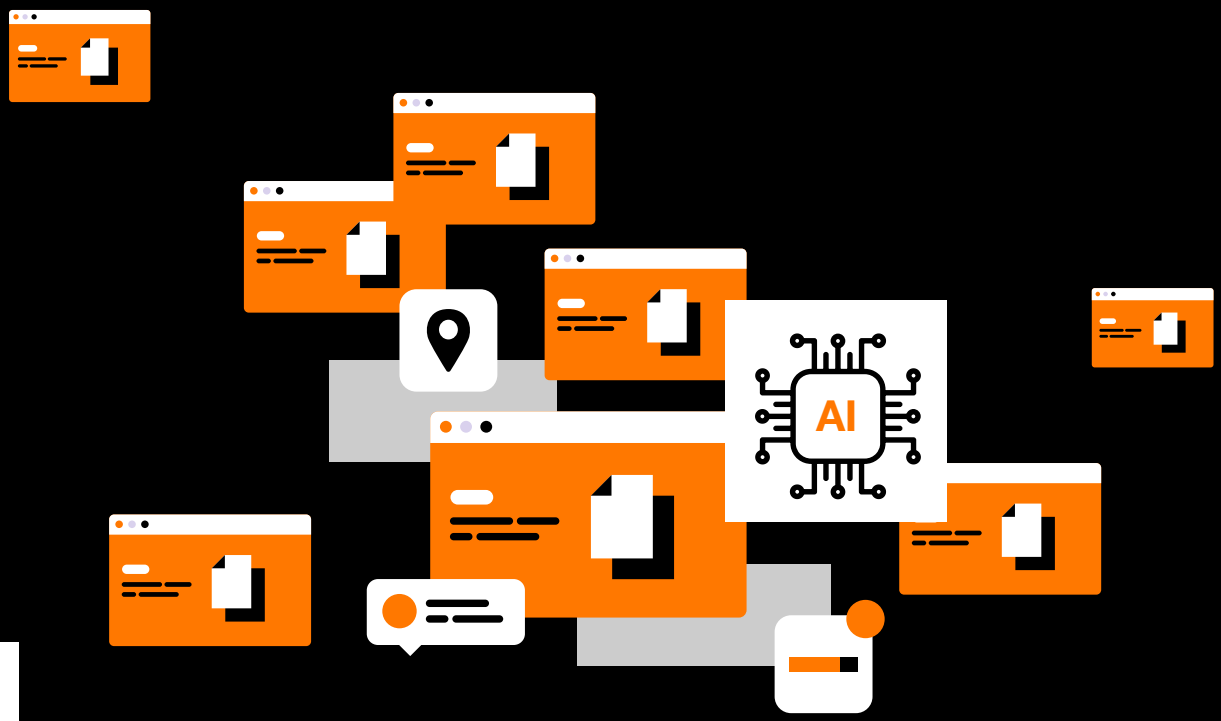
Also, can you use a small language model (SLM) instead? Some of these have been carefully optimized to have a lower impact through distillation. This is the process by which knowledge is transferred from a large, complex model (teacher) to a smaller, more efficient model (student or pupil). As a result, the cost impact is much less – it requires less compute power, and the infrastructure needed to use the SLM is much lower, and this brings down the overall cost a great deal.

You should be aware that if an AI model is repeatedly trained on data generated by other AI models, it loses its quality and ability to produce accurate, diverse, and relevant outputs: this ultimately leads to a phenomenon called 'model collapse'.

Actionable takeaways

To evaluate if an LLM or SLM is right for your use case, consider factors like the complexity of the task, required response speed, available computational resources, data size, and budget. You will find that SLMs are generally better suited for simpler, focused tasks requiring quick responses and limited resources, while LLMs excel in complex, nuanced tasks that demand large datasets and higher computational power. You should also conduct thorough benchmarking tests between different models before making your choice and ensure performance metrics, costs, and output quality are all properly documented.

Because GenAI technology is evolving so rapidly – and the number of available models is proliferating – you may wish to consult with independent experts as part of the model selection process. Orange Business's Live Intelligence platform already includes a curated selection of LLMs/SLMs that evolves over time, ensuring access to the most efficient models without any maintenance being required of the customer – the platform automatically handles model selection and updates while maintaining flexibility and cost optimization.



4

Document governance – focusing your model on the most relevant documents

Logic suggests that the more information your LLM has access to, the better it will perform. Somewhat counterintuitively, this is not the case – the more data you give an LLM, the more it will use – and storing many low-value documents can lead to over-polling of data with poor results and high costs. In reality, it's not about the volume of the data available to your LLM but its relevancy.

So, while data governance is an essential feature of information management, we must also consider document governance. While AI relies on structured data, GenAI has natural language capabilities that allow it to ingest unstructured data like emails, presentations, and other types of documents. Document governance is therefore necessary to ensure that only the most relevant information is polled to reduce costs and improve the quality of results.

Document governance is an essential underpinning of retrieval-augmented generation (RAG), which many companies are using to reduce their GenAI costs. RAG works by combining an information retrieval system with a generative language model: based on a user query, the retrieval system first searches for relevant information from an external knowledge base and then feeds that information to the language model to generate a more accurate and contextually relevant response. RAG is itself more cost-effective than retraining an LLM on your internal data as it leverages existing pre-trained models and allows them to access and incorporate information (this can include unstructured data like emails, presentations and documents) beyond their initial training data.

Document governance is an essential underpinning of retrieval-augmented generation (RAG), which many companies are using to reduce their GenAI costs.

However, this still involves large volumes of data to be queried for each prompt and – as the compute power required varies directly with the volume of data being searched – this is a big driver of cost. A different way of approaching the management of GenAI costs is to reduce the amount of data that is queried. **There are two ways of doing this – using either technology or governance.**

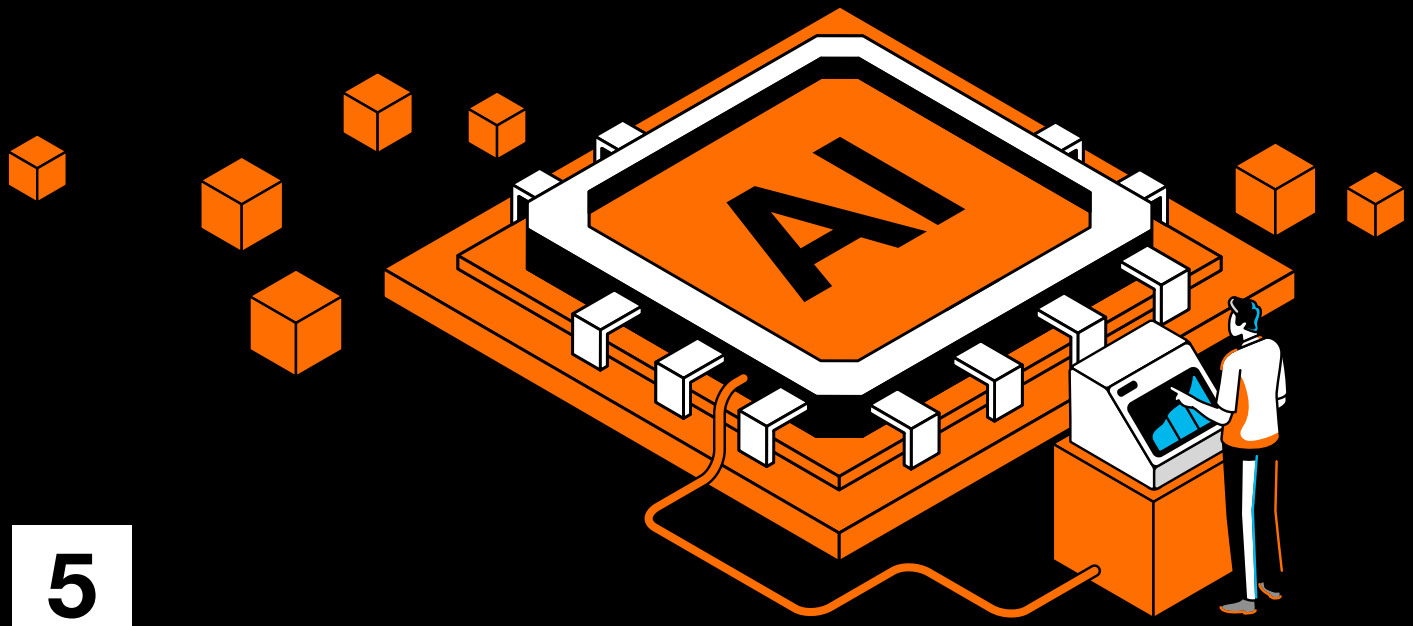
- The technology-based approach is to have the GenAI model create a subset of the overall data by preselecting only the documents most relevant to the search and then setting the LLM to work on this reduced data set
- The second approach has to do with governance. Take, for example, a salesperson using GenAI to automate the response to an RFP. In this instance, the model has retained information relating to every live contract (so, for example, legal teams can update any changes to terms and conditions), but a completely new pricing model was introduced six months ago. In this case, any prompt relating to pricing should exclude any contract predating the introduction of the new pricing model – this would not only provide more accurate information but, by reducing the number of documents, would also dramatically reduce associated costs

Actionable takeaways

For all types of data, traditional 'garbage in, garbage out' rules apply – and it is much more cost-effective to address document governance and data quality from the start than it is to implement it once the project has begun. (This will also maximize the value you create from your GenAI service.)

To ensure that only relevant documents are included in your RAG repositories, it is vital that those responsible for creating the data work transversally with those who will ultimately consume it. This is about putting the right responsibilities in the right hands and creating a feedback loop that will ensure high-quality results. You should establish a data/document governance committee to oversee document quality and relevance. You should implement standardized metadata tagging for better categorization and retrieval and set up regular review processes to archive or delete obsolete documents. Clear guidelines for document lifecycle management and regular audits of your document repository are both important to ensure relevance. Orange Business has considerable experience in helping customers manage this process and would be happy to advise.

Robust, structured data cleaning involves identifying and removing duplicates, outliers, inconsistencies, and irrelevant information and ensuring proper data formatting while also regularly monitoring for emerging biases or outdated facts within the training data.



5

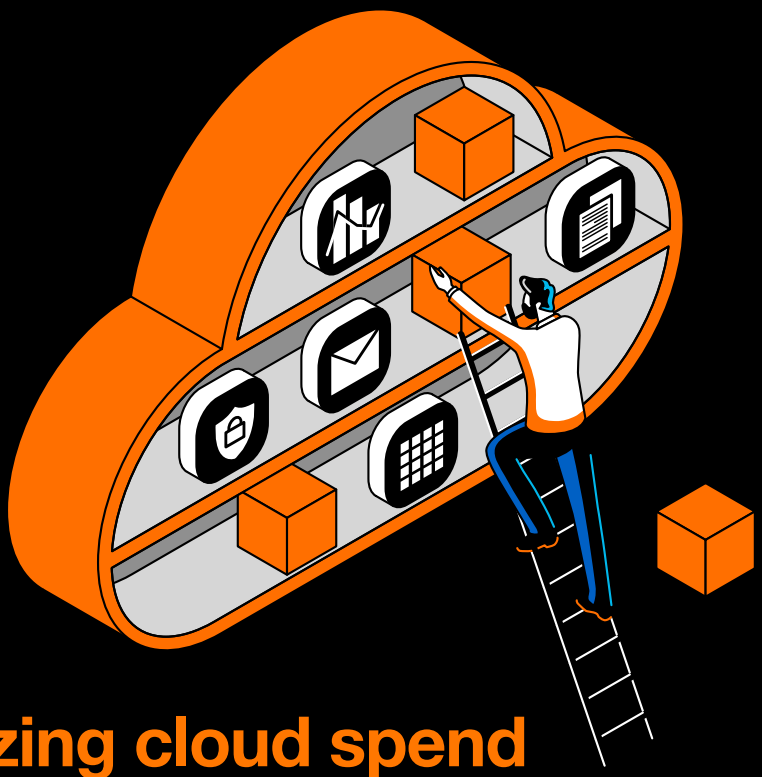
Prompt libraries – every query has a cost, so give users access to the right prompts

As we've explained, the large volumes of data queried by an LLM come with a significant cost. So, if a user submits five or six prompts before alighting on one that provides the right answer, compute costs can increase substantially. Equally, reducing the number of attempts can generate significant savings. You should, therefore, consider a prompt library created by pre-building prompts that are efficient and allow users to get good results with a minimum number of attempts.

Actionable takeaways

Once again, collaboration is the key. Prompt engineers should work closely with end users to create a feedback loop that continuously improves computational efficiency by identifying and validating prompts for the most common queries. This can be the heart of a center of excellence for prompt management that centralizes best practices to provide a ready-to-use directory of frequently asked prompts that leverages internal expertise and optimizes resource usage. You should also consider implementing a rating system to identify and promote the most effective prompts.

Orange Business's Live Intelligence comes with a pre-built, industry-tested prompt library that customers can use immediately and extend based on their specific needs. This foundation significantly reduces implementation time while ensuring prompt efficiency.



6

FinOps – optimizing cloud spend through greater visibility of GenAI activity

The direct costs of your GenAI service – the compute and storage costs – are no secret. The invoices land in your inbox. But the indirect costs, those relating to the enhanced infrastructure needed to run that service effectively, are harder to determine and often surprisingly higher than expected.

FinOps is a process providing financial and operational control for cloud and cloud-related budgets. It is a practice and framework that helps organizations optimize their cloud spending by ensuring transparency and collaboration between technical teams, finance, and business units. It enables them to make data-driven decisions about cloud resource usage, maximizing the business value from their cloud investments while controlling costs effectively.

An obvious example would be the introduction of a GenAI service hosted by a hyperscaler, which becomes wildly popular internally and results in skyrocketing costs. In just such a scenario, when we at Orange Business first enabled our employees to use ChatGPT, we put a value limit on usage volumes to control the costs.

Actionable takeaways

There are several FinOps-based strategies you can employ to reduce the cloud costs of your GenAI services. Firstly, you should establish a robust cloud cost governance framework and optimize resource utilization by switching off unused resources. You can also leverage cost management tools like AWS Cost Explorer and Azure Cost Management to gain visibility into your spending and implement chargeback models in which cloud costs are allocated to individual departments.

However, these tools will not address the wider scope of GenAI costs that encompass data, usage, infrastructure, cybersecurity and cloud, many of which are beyond the control of the hyperscalers. As always, you should rely on independent expertise that will ensure you can take a holistic approach to cost management.

Business

Conclusion

As you can see, there is no silver bullet that will suddenly minimize your cloud costs. Instead, there are a series of small decisions that can add up to a significant reduction in your expenditure – and a concomitant increase in the value delivered by your GenAI services.

What we can say for sure is that the infrastructure you need for GenAI tomorrow will be different from what you have in place today. A recent survey commissioned by Orange Business found that less than half of respondents had or will have the necessary IT infrastructure to support the GenAI projects they are trying to operationalize. DeepSeek has also demonstrated that it is possible to deliver GenAI services with far fewer resources than was previously thought possible.

Efforts to right-size your digital infrastructure in the future may mean reducing your cloud footprint and enlarging on-premise capabilities (or vice versa). So, by controlling your cloud and edge infrastructure, you gain the agility you need to find the right balance between the volume of data you must manage and the locations in which it is stored. This will ensure you future-proof your business and manage your costs effectively.

We've designed our Evolution Platform to deliver precisely that flexibility. It is a composable platform with a high degree of interoperability between our best-of-breed ecosystem partners so you can quickly and cost-effectively make changes to your digital infrastructure. It is also provided on a cloud-like 'network as a service' basis with a consumption-based pricing model so you can flex the size of your network as circumstance or strategy dictates.

We can all agree with Peter Drucker's famous maxim that, 'The best way to predict the future is to create it'. At the same time, we should all bear in mind the Swedish proverb that states, 'He who buys what he does not need, steals from himself'. By following the advice in this white paper, it should be possible to fulfill them both.