GAI Insights

AI Sovereignty: Evaluating the Case for Local Model Ownership

Comparing Inference Costs from GPT 4o on Azure, Llama 3.1 405B on AWS, & Inflection AI in the Enterprise Environment

Commissioned by Inflection AI

February 2025

Table of Contents

Executive Summary

In a September 2023 HBR article entitled, *Where Should Your Company Start with Generative AI?*[1]" We created a new category of work we call WINS work. That is, work that is made up of creating or improving Words, Images, Numbers and Sounds: WINS. We predicted that those tasks, functions and **firms that are highly WINS intensive and already digital, will be radically disrupted by 2030 due to GenAI**. And in the past year and a half we have seen significant adoption of GenAI to increase productivity in WINS work.

Companies face a bewildering array of deployment options all with different cost and feature tradeoffs. To better understand these tradeoffs, we detail 3 scenarios which explore the cost of inference, to partially, or fully automate WINS tasks using frontier size large language models[2]. The relative cost of inference compared to the employee's salary is strikingly small. For example, in a 45,000 person call center with an estimated fully loaded cost of employees at $75,000 per year, the labor cost alone is over $3 billion. **A 1% improvement in labor productivity would represent over $30,000,000 per year in savings and would more than cover our highest estimate of token use to support 45,000 workers.** Even if we are off by 10X in our cost estimates it is an obvious choice for any enterprise with significant WINS work to invest now to begin the process of productivity improvement

This document outlines **three scenarios** as of February 2025:

- Buying OpenAI's ChatGPT 4o from Azure
- Sourcing Llama 3.1 405B from AWS
- Purchasing Inflection AI's large model deployed within the enterprise environment

The financial model shows that the cost of hosting a proprietary model can be less expensive on a comparative basis, ranging from 10% less expensive to as much as 60% in a given year. The job of bringing the model in house will mean additional effort and management. Moreover, because the economics of hardware, software, models and algorithms are moving very quickly and dramatically, all economic assumptions have significant uncertainty in them. Nevertheless, we believe that the additional investment in management and effort is worth it for firms who are critically dependent on inference for their current costs and future growth. Also, we must always keep in mind the much larger costs will be associated with tuning the model, mapping the new processes and implementing the new work flows.

Benefits of deployment within the enterprise environment are:

- Development of **internal capability** with GenAI
- More ability to **assess** IP & security **risks**
- Better ability to **negotiate price** with suppliers
- Superior control for business **continuity** and regulatory **compliance.**

[1] https://hbr.org/2023/09/where-should-your-company-start-with-genai
2 In this report we have made a number of simplifying assumptions and focused heavily on inference costs and provision. Appendix A outlines some of the volatile dynamics of this market.

Our Recommendations:

Assess your organization's reliance on GenAI now and in the future using our GenAI Strategic Grid (see the Strategic Grid in the next section of this report) and decide if you are in the Strategic, Innovation, Factory, or Supply quadrant.

- Firms in the **Strategic** or **Innovation** categories (e.g. firms like JP Morgan Chase or Novartis), we recommend a **hybrid** approach to sourcing inference with a combination of models within the enterprise environment and others purchased from cloud providers. For both types of firms a hybrid strategy enables more control over IP risk, security, regulatory compliance and business continuity.
- Firms in the **Factory** category, (such as Ensemble Health, whose cost base is made up of significant creation and improvement of WINS work) should consider having **two or more suppliers** of inference and consciously architect the solution to make it easy to switch between suppliers.
- Firms in the **Support** category (such as McDonalds), we'd recommend **choosing one** of the large providers of inference and to learn their overall environment to see where they could apply tools and techniques for incremental benefit.
- Given the vastly lower cost of machine inference compared to human inference, we encourage **all firms to begin a process of exploration and exploitation,** and continue with their existing efforts as the potential labor productivity[3] is vast and leverage of innovation efforts well documented.
- Perhaps most importantly, we believe that there are two types of **compound learning** with these models.
    - The **model can be improved** with time and tuning
    - The **organization's capacity** to design and implement entirely new processes with a combination of human and machine agents is only beginning and we are already starting to see superior ability to deploy and run models in firms who have begun their journey. **Learning effects are notoriously difficult to compete against if first movers keep investing in their lead.**

[3] https://www.forbes.com/sites/jackkelly/2023/03/31/goldman-sachs-predicts-300-million-jobs-will-be-lost-or-degraded-by-artificial-intelligence/

Introduction

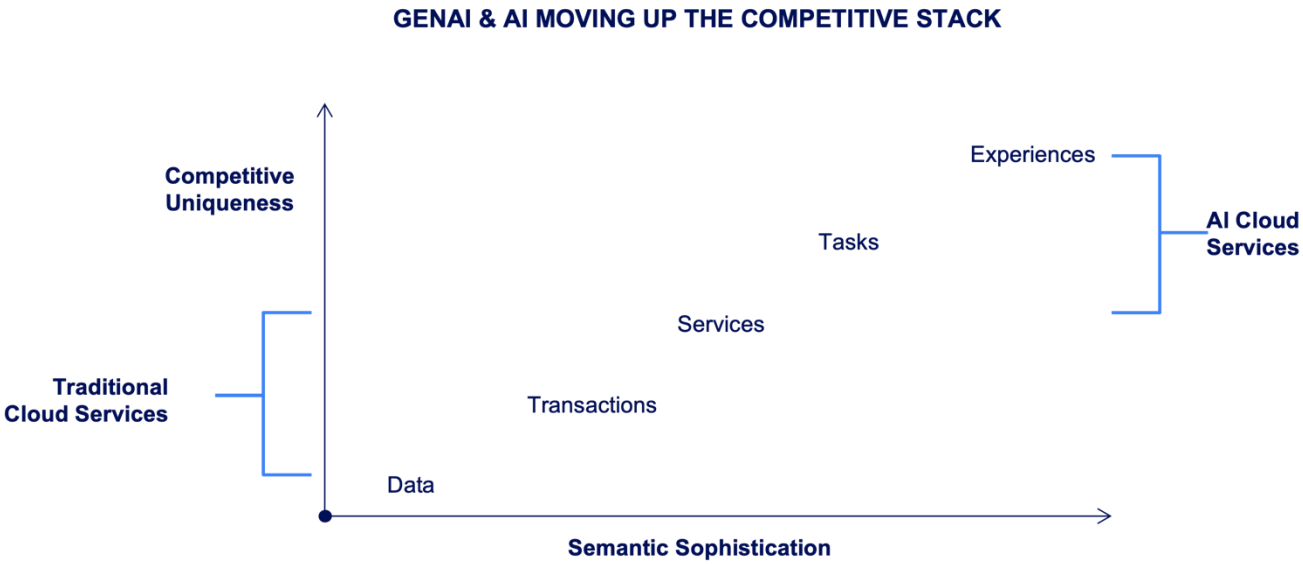
> **Prediction is hard, especially about the future.**
>
> Niels Bohr

Where are we with GenAI/AI?

We are in the early stages of GenAI adoption. Since November 2022 when ChatGPT 3.5 set the world record for a new technology adoption, we have seen tremendous growth in investment and early use. A recent survey of Data and AI leaders has seen the percentage of firms in **production with GenAI move from 5% of respondents to 24% in 2024.** Other estimates are in a similar range. In our work with these early adopters we see they tend to create a hybrid model of purchasing some inference from within a provider environment and building some capabilities internally through a combination of specialty suppliers and upskilling internal talent.

The Difference Between GenAI Cloud and Traditional Cloud Services

Cloud-like services have been around since at least the 1960s with the advent of timeshare computing and commercial cloud services were offered at scale from Amazon AWS beginning in 2002 – so just over twenty years ago. With the advent of GenAI and cloud inference **we are now moving up the stack in terms of an organization's cognitive assets.**



The figure above illustrates the move from data and transactions to services to tasks and experiences. This increase in semantic sophistication and increasing role in customer experience means that the supply strategy for inference capabilities in the firm is ever more important.

The central questions to ask when thinking about your inference supply strategy is:

1. How important is GenAI to our current and future cost base?
2. How important is GenAI to our current and future revenue growth?

These are the most important strategic variables to consider when investing in GenAI capacity and supply. The next figure lays out this idea: we call it the GenAI strategic matrix[4].

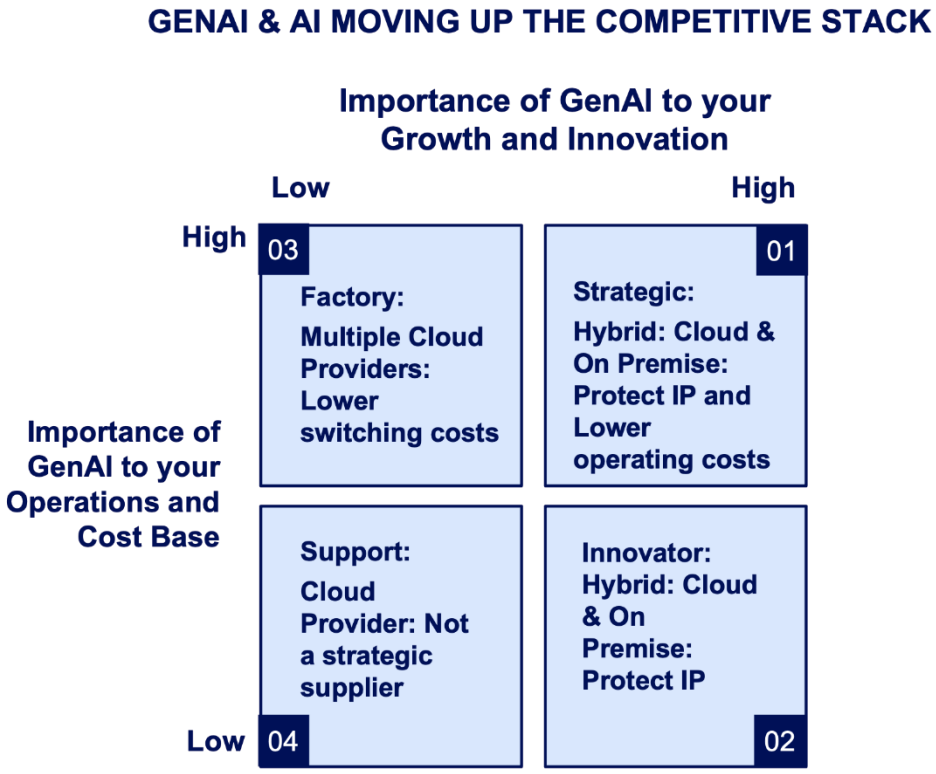


Figure 2: GenAI Strategic Matrix: Strategic Sourcing

The **GenAI Strategic Matrix** provides a clear framework for organizations to determine their optimal approach to adopting GenAI based on two critical dimensions: the importance of GenAI to their current operations and cost base (vertical axis) and its significance for future growth and innovation (horizontal axis). By assessing where they fall on this 2x2 grid, organizations can align their AI adoption strategies with their business priorities, ensuring both cost-effectiveness and strategic advantage.

[4] I'd like to acknowledge my debt to F. Warren McFarland's Strategic Grid, as the inspiration for this framework. McFarlan, F. W. (1984). *Information technology changes the way you compete*. *Harvard Business Review*, 62(3), 98–103. https://hbr.org/1984/05/information-technology-changes-the-way-you-compete

The Support Category: Low Operational Importance, Low Growth Importance

For firms where GenAI is not currently critical to operations or future innovation, the most practical strategy is to adopt a wait-and-see approach. These firms can rely on cost-effective, standard services from any reputable cloud provider, minimizing upfront investment while keeping the option to scale AI capabilities as needs evolve.

The Factory Category: High Operational Importance, Low Growth Importance

Organizations for whom GenAI is essential for current operations but less significant for driving innovation should focus on resilience and cost control. These firms are best served by partnering with multiple cloud providers, ensuring redundancy, competitive pricing, and uninterrupted service. This approach safeguards operational efficiency while avoiding overcommitment to a single supplier.

The Innovator: Low Operational Importance, High Growth Importance

Firms prioritizing GenAI as a driver of future growth and innovation, but for whom it is not yet a major operational necessity, can benefit from hybrid deployments where they may have some on premise capability as well as being familiar with the development environments of the major providers: Azure for MSFT/OpenAI, Bedrock for AMZN, and Vertex for Google. By maintaining capability in developing their AI infrastructure, these firms can experiment, innovate, and build unique capabilities that position them for long-term competitive advantage. In this domain the potential leakage of cognitive and data assets is a risk that needs active management.

The Strategic User: High Operational Importance, High Growth Importance

For organizations where GenAI is critical to both current operations and future innovation, a hybrid strategy is essential. By leveraging a combination of cloud-based inference and on-premise solutions, these firms can optimize costs, maintain operational agility, and develop innovative applications that drive growth while protecting vital data and cognitive assets. This dual approach ensures that both immediate and long-term business objectives are met effectively.

The GenAI Strategic Matrix helps firms navigate the rapidly evolving AI landscape, aligning their investment decisions with their strategic needs and maximizing the value of GenAI across the enterprise.

Key Drivers of LLM Adoption

Functions and organizations whose core value creation depends on WINS work—such as software development, customer service, financial analysis, marketing, and entertainment—are poised to be early and intensive adopters of LLM-based solutions. Their large volumes of knowledge work, combined with the potential for rapid innovation and cost savings, make them highly motivated to invest in this technology.

Market projections indicate that computing costs may rise by nearly 89% from 2023 to late 2025, largely due to GenAI workloads. Moreover, the proliferation of AI agents—expected to handle vast volumes of transactions, each potentially requiring 300,000 to 500,000 tokens or more—underscores the scale and complexity of future AI operations. We see three key reasons driving adoption.

Reason 1: Automation

To date the most rigorous empirical work shows productivity increases of about 13%[5]. But, with further job redesign, additional training of the model, and task redesign we believe it will easily reach 50% or more of many jobs. In this report we expect 50% or more of a call center operator's job can be automated. We base this prediction on the fact that Jerry Insurance (see below) was able to automate 89% of the chat and SMS customer traffic in a mere 4 months[6]. The fully loaded cost of a call center worker's salary in the U.S. is about $300 per day. That means **for each day's salary one could buy over 250,000,000 tokens of inference capacity. Our model indicates that we could automate about 50% of the worker's job with 100,000 tokens a day**. That leaves massive token capacity for further automation. As in the industrial revolution, the AI age is ushering in a massive substitution of capital for labor. Many studies have shown this powerful effect, and every executive must take a long hard look at their own organization because of it.

> **..for each day's salary one could buy over 250,000,000 tokens of inference capacity. Our model indicates that we could automate about 50% of the worker's job with 100,000 tokens a day.**

How big will this trend be? **We expect in 2025 a third or more of firms will deploy labor-saving GenAI/AI at scale in at least one major function.** These hybrid organizations will lead the way in cost performance as they also learn how to optimally create and work with digital workers. Developing a new productive capacity takes time, so senior executives who have a wait-and-see attitude may find themselves on the wrong side of an experience curve that will be hard to compete with.

We predict the capability to deploy a truly hybrid organization will become better and better with

[5] National Bureau of Economic Research (NBER). (2023). *Working Paper Series*. https://www.nber.org/system/files/working_papers/w31161/revisions/w31161.rev0.pdf
[6] https://sloanreview.mit.edu/article/turbocharging-organizational-learning-with-genai/

more volume and transactions. (See Bruce Henderson and BCG's classic work[7] on this powerful strategic concept of the experience curve and why first movers can gain competitive advantage if they apply it rigorously.)

Reason 2: Lower Cost of Innovation

Chip Hazard, co-founder of Flybridge Ventures, a successful early-stage venture capital firm, has said, "With this new wave of AI, **startups that used to ask for $5-10 million dollars are now asking for $2-4 million** and using AI to make up the difference."

> **With this new wave of AI, startups that used to ask for $5-10 million dollars are now asking for $2-4 million and using AI to make up the difference.**
>
> - Chip Hazard, co-founder of Flybridge

AI, well applied, creates organizational capacity to do new things. Think of Google's policy of letting people work on their own innovations one day a week, which has helped them create many new things including the well-known Google Finance product. GenAI can help deliver the slack to incubate any type of innovation. Digital workers can help with the innovation process itself, too. Large firms like Coca-Cola have used AI[8] to collaborate on images and branding, and many academic papers[9] demonstrate increased innovation while using AI because it is a great collaborator on ideation, simulation and explication. For example, creating optimal customer profiles, analyzing trends in markets, laying out detailed marketing execution plans are trivially easy to draft, review and improve. **This means hybrid organizations will have more capacity to innovate and a lower cost of experimentation than traditional firms.**

Reason 3: Agile Scalability

One of this report's authors has run large people-intensive businesses and he knows from experience just how difficult it is to scale up and scale down personnel in the most effective and humane manner. **Digital workers and the hybrid organization can deliver the promise of massively scalable services of higher quality and less volatility.** For example, when getjerry.com, the popular car insurance and refinancing site, created GenAI tools to help serve its 5 million customers in the chat and text channels, it was able to automate all but 11% of the customer questions with higher satisfaction and almost instantaneous response. Not only did this create a $4 million per year ROI[10] for the company, but it also provided scalability that will enable it to grow 3X or more without adding any additional customer service staff. A hybrid organization that uses digital workers to augment tasks with peaks and valleys of demand such as customer

[7] https://www.bcg.com/publications/2013/growth-business-unit-strategy-experience-curve-bcg-classics-revisited
[8] https://www.coca-colacompany.com/media-center/coca-cola-creations-imagines-year-3000-futuristic-flavor-ai-powered-experience
[9] https://www.emerald.com/insight/content/doi/10.1108/ejim-02-2024-0129/full/html
[10] https://gaiinsights.substack.com/p/genai-customer-care-project-now-saving

service, contract review, etc., will have better service with less risk.

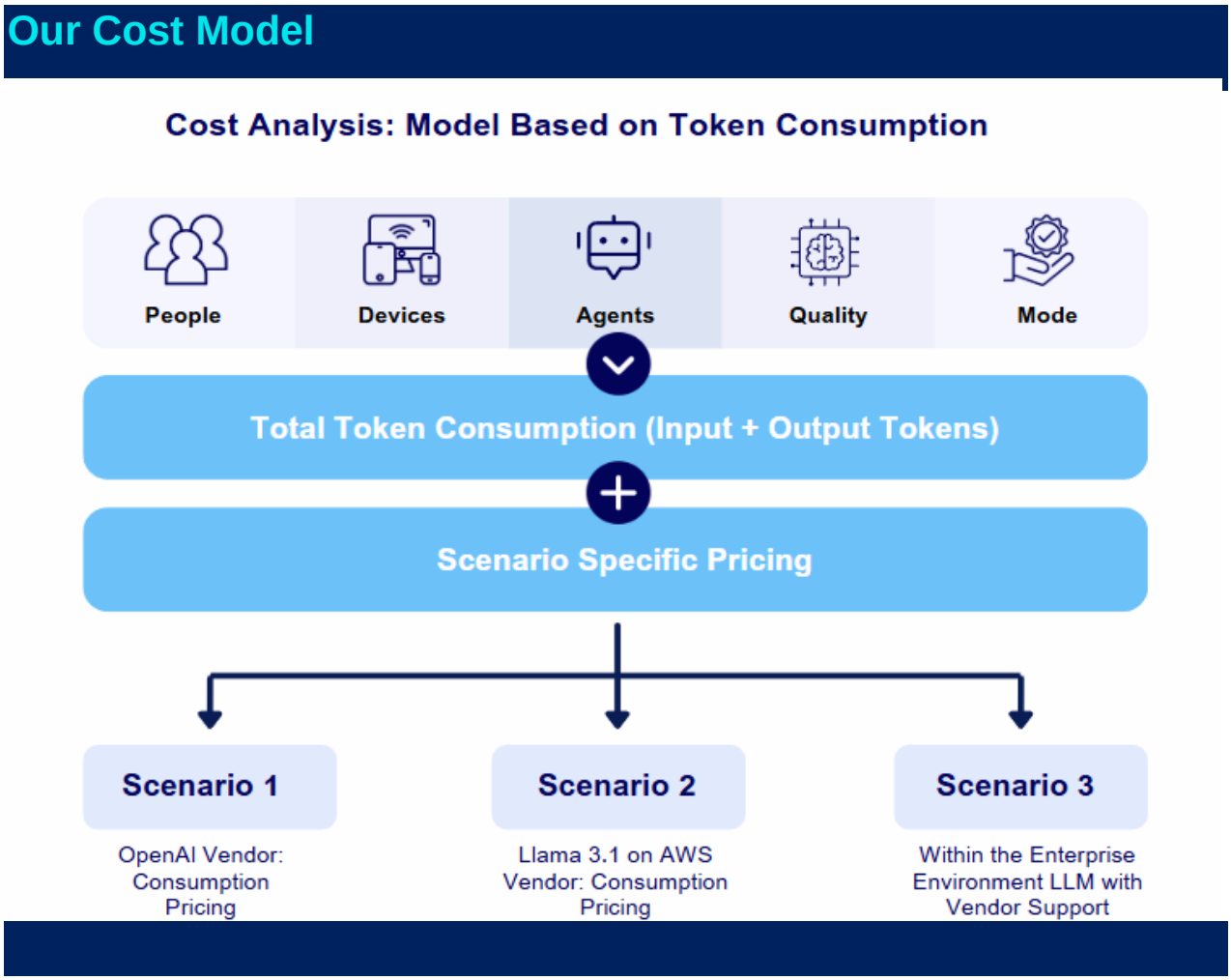


Figure 3: Inference Supply Analysis Based on Token Consumption Across 3 Options

The core of this cost model is to look at two large functions within a firm for whom GenAI is strategic. In this case we are looking at a large call center with 45,000 employees. Large telecommunications operators like AT&T have about this many call center personnel. The second function is the role of a banker in a large firm like JP Morgan Chase. They have approximately 50,000 bankers in a firm of over 300,000 employees. These two domains provide the "demand" side of our cost analysis. Of course, the profile of the demand for inference is related to consumption and adoption levels. For this model we have assumed that there will be gradual increases in usage for each year across the two models. See the table below.

We also assume that the organization already has a data center; that any model tuning costs would be the same across the different options, and that local data storage, security, communications, etc. are either already in place, or the incremental needs of the different

models' approaches would not be materially different[11]. Put another way we are interested in the major, material incremental costs of building a supply of inference.

Call Center and Banker Models: GenAI Adoption Speed Over 3 Years

Year	AI Supported Call Center Percentage of Workers Using GenAI	AI Supported Banker Percentage of Workers Using GenAI
1	20%	50%
2	50%	75%
3	75%	90%

We have created an estimate of likely adoption speed (see above) and the number of tokens that will be used per shift by the call center workers based on the typical length of call, number of words in a call, and number of calls per operator per day. For the bankers, the second demand generation scenario, we created an estimate of how many tokens bankers would use in their day-to-day work – answering calls, analyzing reports, etc. and created an estimate of the number of tokens a banker would use per day. For the call center daily worker use is 48,000 and for the banker's token use goes from 150,000 in the first year to 225,000 in the third.

We do have evidence from one leaving investment bank, that has approximately 10,000 bankers, that their AI platform has already increased banker productivity in "double digits" across the board and in some tasks, such as preparing for a high-net-worth client, productivity may be as high as 40% on that task, as reported by their CIO. We are confident we will see significant growth in token use. Even last year, work by the Wharton School which showed year on year increase of 50% in token use[12]by those adopting the technology.

On the supply side, we have three scenarios for provision of inference: scenario 1 is the purchase of OpenAI Chat GPT 4.o from Azure with Provisioned Throughput Units (PTU) pricing and an assumption of 25% of inference costs added for other needed Azure services. The second supply scenario is hosting Llama 405B on AWS using the per token pricing, and 25% for additional needed AWS services. The third is an enterprise environment model where the firm hires one of the large model firms, in this case Inflection AI, to provide the core model, set up

11 In this report we use the term: provider environment versus enterprise environment. On premise versus cloud is not a fine enough distinction. Is the model running in the provider environment – e.g. ChatGPT 4o on Azure or is the Inflection AI model running on a virtual private cloud with dedicated servers – either physically on premised within the enterprise, or in a provider's data center, but within the enterprise's Virtual Private Cloud.

[12]Korst, J. Purk, M. (2024). *Growing Up: Navigating GenAI's Early Years.* Wharton Business School. https://ai.wharton.upenn.edu/wp-content/uploads/2024/11/AI-Report_Full-Report.pdf

and priced on relevant servers to provide needed compute capacity. Each bar chart below represents one demand scenario and the three supply models.

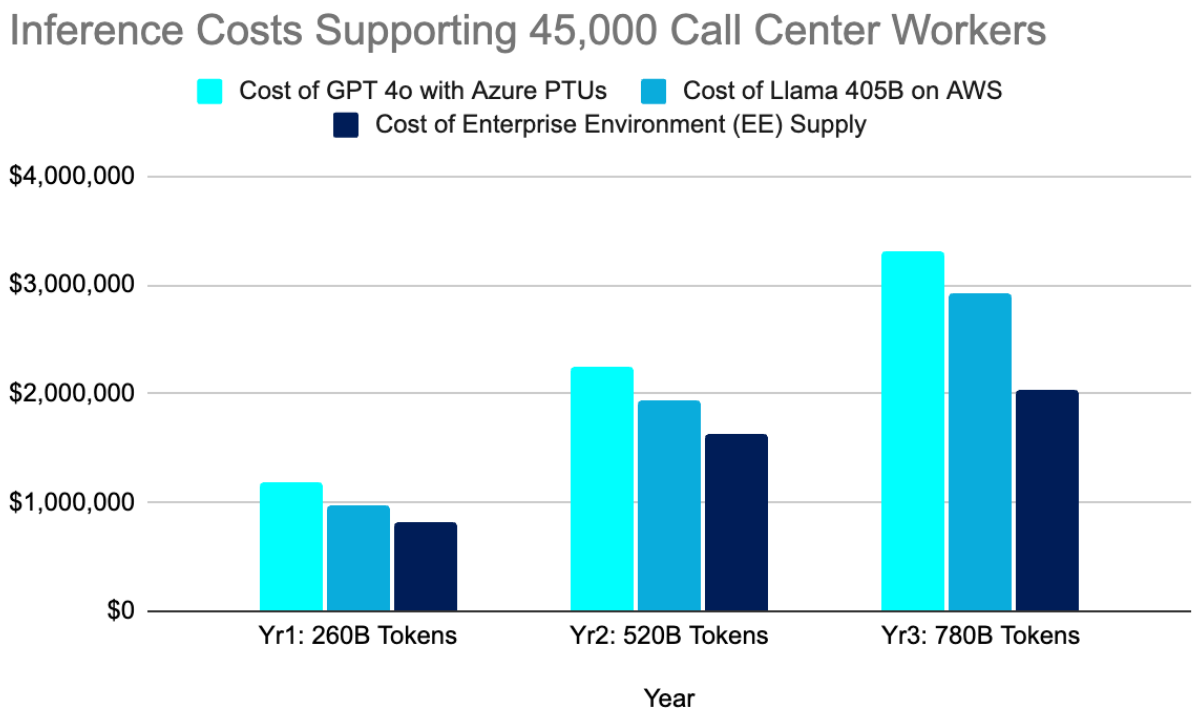


Figure 4: Financial Model of Call Center Deployments

Key Current Attributes Across Three Sources of Inference Supply

Figure 4 has a comparison of the three provisions of tokens for the call center and Figure 5 has it for the banker support. In both cases there is a cost saving associated with bringing inference into the enterprise environment. In the call center case, that saving ranges from $200,000 or so in year one to about $1.2 million in year three.

This enterprise environment installation would be supporting a significant portion of the 45,000 call center workers and the cost of the enterprise environment inference in year three of $2 million represents less than 0.1% of the yearly salary cost of those 45,000 workers. The much greater challenge will be in building and tuning the model and reengineering the workflows and processes to enable significant labor savings in the call center. Also, it is worth remembering that a study for the National Economic Bureau showed a 13%[13] increase in task level productivity for call center workers supported by a large language model.

The banker case begins with a token use of 500 billion in the first year, but with greater adoption it rises to over 2 trillion tokens per year. The enterprise environment savings begins at about $500,000 and rises to $2.3 million by year 3. This is a substantial difference, but again the total of this inference capacity is less than 1% of the total labor cost of the bankers (salary cost is about $7.1 billion per year). The GenAI cost of inference pales in comparison and early reports of increased banker productivity amply covering the investment.

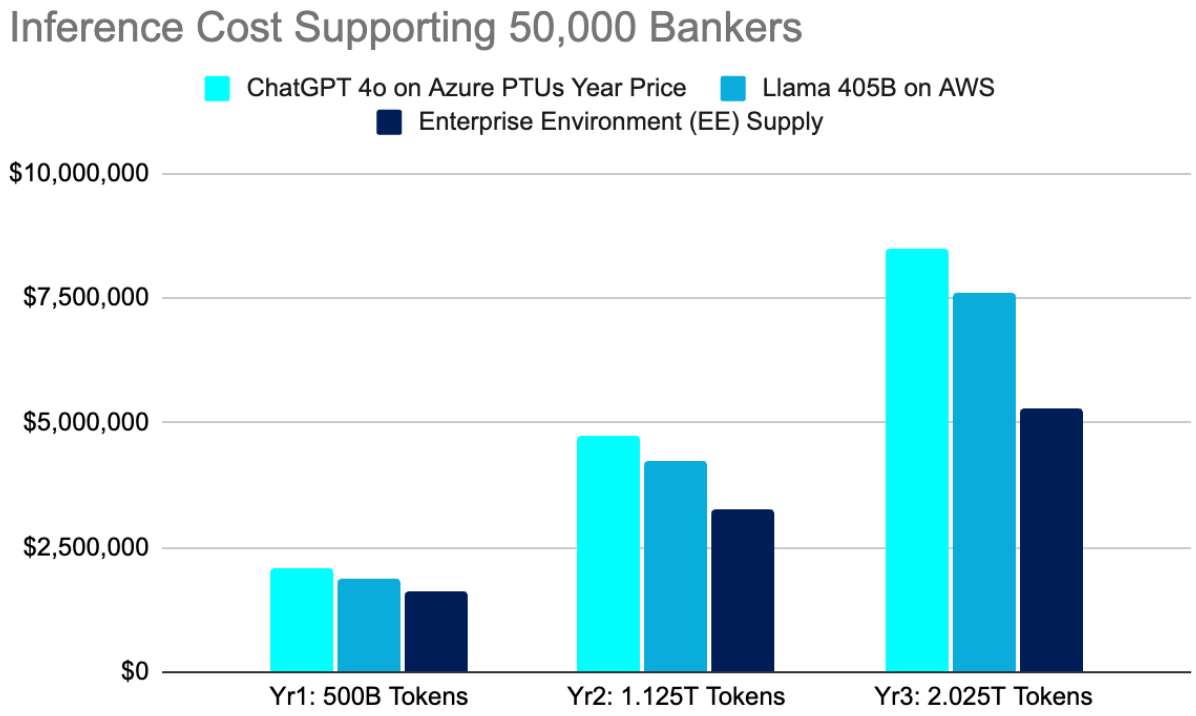


Figure 5: Financial Model of Banker Deployments

If you believe, as we do at GAI Insights, that due to the massive potential labor cost savings in WINS work there will be widespread use of agents and other techniques to improve labor productivity. This means that token use will grow significantly from where it is today.

[13]National Bureau of Economic Research (NBER). (2023). *Working Paper Series*. https://www.nber.org/system/files/working_papers/w31161/revisions/w31161.rev0.pdf

The No Brainer: Labor Costs versus Inference Costs

If we look at the cost of labor in these two scenarios, we have 45,000 call center operators which in the US cost approximately $75,000 per year, fully loaded with benefits costs, etc. That is a labor cost budget of $3.375 billion per year. The cost of the bankers is about $150,000 per year per banker, fully loaded, across 50,000 bankers which is a labor cost of $7.5 billion per year. **Even in the most expensive scenario with high token use and an enterprise environment implementation the total cost is under $8,500,000 per year which is well under the $75 million a year which is the value of a 1% increase in banker productivity.** It is important to remember that many other costs such as data acquisition, cleaning; model tuning, process redesign and retraining, and other costs will be significant, but still well below the cost of labor.

We believe there are five key non-economic reasons to consider implementing high quality reasoning behind your trust boundary:

1. Intellectual Property Risk
2. Risk of leakage of proprietary data and cognitive capital
3. Market power/supplier independence
4. Continuity and risk management
5. Creating internal capacity in GenAI/AI

Intellectual Property Risks

The current wave of lawsuits against major large language model (LLM) firms centers around complex issues of copyright and intellectual property (IP), raising questions about how AI models handle the data they are trained on. The potential remedies include financial settlements, licensing agreements, or judicial rulings that clarify the legal framework governing AI training data. Regarding indemnification, most LLM firms offer no or very limited protections to their users. For example, the Llama models from Meta clearly provide no indemnification of any type and recently Mark Zuckerberg[14] is accused of actively encouraging his team to violate copyright rules to access more training data. This may cause many issues for firms adopting this popular free model.

These lawsuits will likely take years to be resolved, and the uncertainty hangs over the users of these models. Therefore, careful consideration of the approach and care with which the model builders take on training data stewardship and ownership are a critical consideration whether one is bringing the model in house or buying inference from a provider.

Risks of Leakage of Proprietary Data and/or Cognitive Capital

When a firm contracts with OpenAI or other model providers there are several important security measures to consider. Firms can work with the vendor to make sure that their needs and

[14] Zuckerberg, M. (n.d.). [Public statements or internal memo regarding data usage]. Meta. https://edition.cnn.com/2024/02/29/tech/meta-data-processing-europe-gdpr/index.html

policies are reflected in the use of the model. Here are some of the most important issues:

- Basic encryption and data protection
- Training permissions
- Queries and responses
- Administrative functions that also keep data

Of course, the model provider can sign contracts affirming that they will use strong encryption for data transmission and storage, not train their models on the users' data, and delete relevant inputs and responses as the client wishes. There are two points that even the most secure relationship needs to open. **When a client uses a model hosted by a cloud provider, the data must be unencrypted to be used in the model itself[15].** In addition, many firms, including Microsoft, keep data, often thirty days worth, to make sure that the user does not violate the provider's use policies.

Many highly regulated firms, such as US health care insurers, some financial institutions, and others are wary of allowing their data to be available to the model that is outside their trust boundary. **Some believe they must have the models within their trust boundary to meet their risk and regulatory requirements.**

More broadly, as firms build their AI capabilities, they are taking data and processes that today are tacit or unstructured and structuring them and creating a cognitive asset that can perform tasks that could only previously be done by the humans in their organization. We at GenAI believe that **it is vital that firms protect their cognitive capital**. Many people who are buying tokens must utilize all appropriate measures to protect key cognitive assets. This issue is especially poignant for firms in quadrants 1 & 2 of our Strategic Supply Matrix covered earlier in this report.

Cost Control/Market Power/Vendor Lock In

Who do you trust? Certain vendors in the AI ecosystem have inordinate market power. Therefore we believe it is vital to consider how difficult it is to move to a new vendor if you need to. If your organization becomes critically dependent on inference and tokens by one provider, that vendor will have enormous ability to raise the cost of their services as some are seeing in the current market for traditional cloud services. **Building in-house expertise** and capabilities in LLM tuning, operations and maintenance equips enterprises to negotiate more favorable terms with external suppliers.

This strategic autonomy helps mitigate the risks of vendor lock-in, protects knowledge assets, and empowers firms to innovate rapidly without ceding too much influence to dominant market players.

Firms should also architect their efforts so that they can more easily access multiple models and vendors. Creation of prompt databases, keeping control of security software and the emerging semantic operating systems that do security management, model routing, and

[15] See Appendix C for an Illustration of this flow of data.

other cross model functions are key decisions to keep from becoming locked into a particular vendor due to massive switching costs.

Cost and Quality Control:

As the cost base increasingly shifts toward AI inference and model tuning, maintaining better visibility into total costs becomes essential. Owning or partially hosting LLMs within the enterprise environment can help firms understand the drivers of these expenses, reducing dependency on supplier pricing and aligning with long-term budgetary strategies. Also, because the supply base is so new and dynamic there are reports of lack of availability of models and higher latency in certain circumstances. If a major supplier cannot deliver needed inference and quality, it is useful to have your own internal capacity to use.

There are already reports that some firms who have contracted for access to the large models have been routed to smaller models on the supplier's platform, without notice. This is very hard to monitor in this fast-moving market.

Continuity and Risk Management:

For those firms in quadrants 1 & 2 where proprietary data and IP are critical to their value adding activities, they may want full access to model source code, weights, and other IP. The large cloud providers are not yet willing to share most of the inputs to model creation. Moreover, in the open-source world the challenge is getting the talent to build and run these gargantuan models. Hybrid capacity enables business continuity if for whatever reason the provider environment has an inability to perform.

Ensuring business continuity is paramount. Controlling the inference infrastructure—whether fully within an enterprise environment or as part of a hybrid model—reduces dependence on external providers. This resilience is critical for sustaining operations, meeting compliance requirements, and preserving service quality even amid market shifts or vendor disruptions.

Conclusion

In summary, our analysis indicates that there is likely to be significant GenAI adoption because of the labor cost savings, organizational slack for innovation, and easy scalability such capacity can deliver. These factors underscore the importance of carefully assessing the role and significance of GenAI in your organization's strategy.

- For companies in our GenAI Strategic Matrix who are in the **Strategic and Innovation** quadrants, leveraging GenAI to balance cost reduction and growth—such as retail giants like Walmart or energy companies like Shell—**a hybrid approach** is optimal. Sensitive and strategic operations should remain within the enterprise environment, while other workloads can take advantage of the scalability offered by the provider environment.
- If GenAI is not central to your growth or operational goals, aligning with a major provider environment solution for inference can provide a cost-effective and streamlined approach.
- Organizations who are large and relying on GenAI for both growth and cost management should **dedicate a significant portion of their resources to bring new capability**, such as at scale proprietary models like Inflection AI and others, into their enterprise environment. Such supply will increase the organizations security options, control over intellectual property, and increase their market power to allow for other suppliers and options in their ongoing quest for best quality and best cost supply of inference — implementing models with vendors who offer transparency, including sharing model architectures, weights, and even source code, to ensure greater control and adaptability.
- Finally, it is vital for these organizations to **invest in building internal expertise**. Cultivating a skilled team to configure, operate, and refine these models driving sustained competitive advantage and agility in a rapidly evolving GenAI landscape.

Appendix A: The Volatile, Uncertain, Complex and Ambiguous (VUCA) Nature of GenAI Dynamics!

Throughout history, the more powerful the technology is the more challenging it is to predict its impact. In the early days of cloud computing projections, the modeler had a few anchor points such as the cost of mainframe, minicomputer and personal computer access, and good data on many usage patterns such as the consumption of databases, etc. But, in the new GenAI/AI world, we see tremendous change in computing hardware, the nature of algorithms, the cost of inference, the size of models, sources of data, the evaluation of models, primary usage patterns, architecture of provision of supply, regulatory environment all amplified by the most profitable firms in the history of mankind. Also many countries, especially the US and China are spending hundreds of billions of dollars hoping to win the next great competitive and military battle for supremacy.

The graphic below gives some of the complex, interacting dynamics in this market.