ENTENTIA

ENTERPRISE GEN AI MODEL SELECTION FRAMEWORK

Part II – Beyond Technical Capabilities

Abstract

This whitepaper outlines an enterprise-specific framework to selecting a Gen AI model (or model family) as an enterprise standard across a range of use cases. The framework goes beyond model quality and performance to incorporate various other factors critical to successful and accelerated deployment of Gen AI solutions at scale.

Ententia LLC
info@ententia.ai

ENTENTIA

PART II: BEYOND TECHNICAL CAPABILITIES

In Part I, we introduced the Enterprise Gen AI Model Selection Framework and explored the key considerations for model performance and service infrastructure. These foundational elements help guide enterprises in assessing the capabilities and the technical ecosystem required to effectively deploy Generative AI solutions. In Part II, we will continue by addressing additional categories critical to the selection process, including Commercial Considerations, Sustainability, and Governance.

First, a quick recap of the proposed framework.

Ententia Model Selection Framework

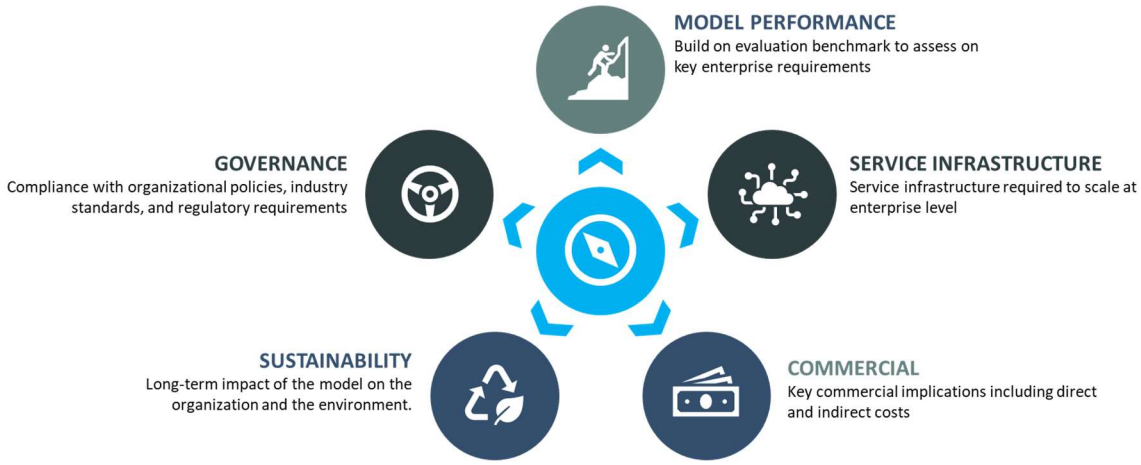

Figure 1: Ententia Gen AI Model Selection Framework

Now, let's look at each of the remaining categories and specific assessment criteria.

3.3 Commercial Considerations

Commercial considerations are equally important in the model selection process, as they directly impact the financial and operational viability of the chosen model. Key factors include:

1. Individual / Enterprise Licensing

Licensing structures can significantly impact the total cost of ownership and the scalability of AI solutions within an enterprise. It's important to consider:

> ***Licensing Models***: Understand whether the licensing is structured on an individual basis (per user) or for the entire enterprise. Individual licenses may be cost-effective for small teams, but enterprise licensing typically offers better scalability and value for larger organizations.

Flexibility and Scalability: The licensing model should offer flexibility to scale up or down based on the organization's needs. For enterprises with fluctuating demands, the ability to adjust licenses as needed can prevent overpaying or underutilizing resources.

2. Usage-based Cost

Usage-based cost models can provide a more accurate reflection of the resources consumed by AI solutions, particularly in environments with variable workloads. Key considerations include:

ONLY 19% OF THOSE SURVEYED VIEW COST AS THE TOP CONCERN WHEN CHOOSING AN AI/GENAI SOLUTION.
BCG

Per Token Cost: Cost per unit of Input and Output tokens will constitute most of the usage-based expense for an enterprise. Vector database usage and storage is typically an additional cost.

Cost Predictability: While usage-based pricing can be advantageous in terms of paying only for what you use, it can also introduce unpredictability in budgeting.

Cost Optimization Strategies: Opportunities to optimize costs, such as off-peak usage discounts or volume-based pricing that rewards higher usage with lower rates.

3. Availability with Cloud Platform of Choice

The ability to deploy and manage AI models on a preferred cloud platform is crucial for maintaining consistency with existing IT strategies and infrastructure. Consider the following:

Cloud Platform Compatibility: Ensure that the AI model or service is fully compatible with your organization's preferred cloud platform (e.g., AWS, Azure, Google Cloud). Compatibility minimizes the need for additional integrations or workarounds and supports smoother deployment and management.

Multi-Cloud Strategy: For organizations adopting a multi-cloud strategy, it's important to evaluate whether the AI model can operate across multiple cloud environments. This flexibility enhances resilience and allows for cost optimization by leveraging different providers.

Vendor Lock-In: Consider the risk of vendor lock-in if the AI model is tightly integrated with a specific cloud platform. While integration can offer performance benefits, it may also limit the organization's ability to switch providers or adopt new technologies in the future.

1. **Forward Compatibility**

Forward compatibility ensures that the selected AI models and their associated infrastructure evolve in a manner that does not require wholesale changes to deployed solutions. This requires looking at the vendor's track record to-date and their commitment to serving enterprise customers.

2. **Development and Maintenance Cost**

The total cost of developing, deploying, and maintaining AI models can vary widely. Key factors to consider include:

Initial Development Costs: This includes the cost of developing solutions, including customization of pre-trained models where needed, integrating them into existing systems, and any necessary infrastructure upgrades.

Ongoing Maintenance: Maintenance costs can include regular updates, bug fixes, performance monitoring, and technical support. It's important to assess whether the vendor provides adequate support and whether the enterprise has the internal resources to manage ongoing maintenance.

Customer Support: Effective customer support is a crucial component of the commercial considerations. Availability, responsiveness, documentation, ecosystem and size of development community are some of the factors that impact quality of customer support.

3.4 Sustainability

Sustainability considerations focus on the long-term impact of the model on the organization and the environment. This includes:

1. **Energy Consumption**

As AI models, particularly large language models, can be resource-intensive, energy consumption is a critical factor. Enterprises should consider the environmental impact and explore models that are optimized for energy efficiency, in terms of pre-training as well as inference.

2. **Model Size (# of Parameters)**

While energy consumption data may be hard to obtain for all models under consideration, size of the model can be a reasonable proxy as it impacts both the power usage and hardware requirements. Smaller, more efficient models are often preferable as they consume fewer resources and are easier to deploy across different environments, including edge devices.

3. **Long-term Viability**

Funding for AI labs developing models is not a significant constraint in today's environment, but it is important to consider whether the company has capacity for sustained investment in the medium to long run and whether they will be able to self-fund when external funding becomes scarce.

4. **Innovation**:

Sustainability is also about staying ahead with innovation. It is worth looking at the history (admittedly short) and ability of the company to either lead or continue to keep up with leaders.

3.5 Governance

Governance is a key area of focus for model selection. Effective governance mitigates risks associated with data handling, bias, and compliance, thereby protecting the enterprise's reputation, minimizing legal exposure, and fostering trust among stakeholders. Below are the key criteria to consider:

1. **Data Privacy Policy**

   Enterprise Data Protection: A comprehensive data privacy policy is critical for safeguarding personal and sensitive information across the organization. The AI model must align with the enterprise's data privacy framework, ensuring compliance with industry regulations and corporate policies on data collection, usage, sharing, and anonymization.

   Stakeholder Transparency: The enterprise must ensure that AI models operate with full transparency regarding data practices. This includes clearly communicating how data is collected and used, obtaining informed consent where necessary, and maintaining robust records of data management practices to support transparency and accountability.

2. **Data Retention and Security**

   Understanding what data is retained, if any, by the model service provider and how the data is secured (in transit and at rest) are an important consideration, especially for regulated industries and enterprises that regularly work with sensitive internal or customer information.

DATA PRIVACY (57%) AND TRUST AND TRANSPARENCY (43%) CONCERNS ARE THE BIGGEST INHIBITORS OF GENERATIVE AI.
IBM Newsroom

3. **Compliance with Regulations**

Regulatory Compliance: Adhering to regulations such as GDPR and HIPAA is crucial for enterprises operating in regulated industries.

Global Data Compliance: For enterprises operating across multiple jurisdictions, ensuring compliance with international data transfer regulations is essential.

4. **Bias Mitigation and Content Moderation**

   Bias mitigation is a key governance concern for enterprises, particularly in high-stakes applications such as hiring, lending, and customer service. The AI model must be equipped with strategies to identify, monitor, and correct biases, and enable enterprises to exercise appropriate control over how solutions are designed.

   The AI model must ensure that content is appropriate, legal, and aligns with the enterprise's brand values and public commitments.

5. **Auditability**

   The ability to audit model decisions and processes may be necessary when applied to certain use cases, to ensure transparency and accountability.

4. Framework in Practice

While the proposed framework is straightforward, implementing it effectively requires deliberate effort.

First and foremost, it is essential to have a deep understanding of the **organization's overall vision, priorities, and the key challenges** associated with adopting Generative AI. This understanding allows decision-makers to appropriately assign relative weight to each category and specific criteria outlined in the framework, ensuring that the model selection process aligns with the organization's strategic goals.

The next step involves conducting a **thorough review of publicly available information**. Many of the criteria listed in the framework necessitate an in-depth examination of existing research, case studies, terms and conditions, and the exploration of available services and infrastructure. Gathering this information and performing a comparative assessment across different models can be time-consuming, but it is critical for making an informed decision.

Certain selection criteria, particularly those related to data handling for high-frequency time series data (which is common in industrial settings) and document processing, require **hands-on testing**. This practical assessment helps determine how well each model performs within a specific organizational environment.

While many of the selection criteria are quantitative, some require qualitative assessment and the application of sound judgment to ensure that the evaluation is aligned with the organization's needs and priorities.

The good news is that, at Ententia, we have conducted extensive research and information gathering as part of our journey to select the most suitable model family for our platform and products. We are ready to collaborate with you to tailor our assessment to your organization's unique needs and help streamline your model selection process.

Contact us to learn more about how you can customize and quantify your assessment using Ententia's proprietary tools and services.

5. Edge Cases

While this framework is designed to offer a standardized approach to model selection across the enterprise, it acknowledges that there will always be edge cases requiring specialized models. These cases should be regarded as exceptions rather than the norm, ensuring that the enterprise can still reap the benefits of a standardized approach while addressing unique requirements.

Specialized Models:

Edge cases may necessitate the use of specialized models tailored to specific use cases or data types. For instance, a model designed for natural language processing in legal contexts may differ significantly from one used for image recognition in healthcare. These specialized models should be selected with careful consideration, ensuring they meet the specific needs of the edge case while remaining aligned with the overall enterprise strategy.

Exception Management:

Effectively managing exceptions is crucial for maintaining a balance between standardization and flexibility. This involves establishing clear criteria for when and how specialized models can be deployed, as well as developing processes for integrating these models into the broader enterprise framework.

Learning and Development:

Treating edge cases as opportunities for learning and development enables the enterprise to continually refine its model selection process. By documenting and sharing insights gained from these edge cases, the organization can build a knowledge base that informs future decisions and accelerates the deployment of AI solutions.

6. Conclusion

In conclusion, this whitepaper outlines a structured approach to model selection, underscoring the critical importance of aligning Generative AI initiatives with the broader strategic objectives of the enterprise. The framework presented here is designed to guide organizations in making informed decisions that extend beyond technical performance, ensuring that the chosen models not only meet current needs but also support the organization's long-term vision.

By integrating a wide range of factors—ranging from technical performance to sustainability, governance, and beyond—this framework offers a holistic and enterprise-centric methodology for model selection. We hope you found this information useful.

What's Next?

We hope this whitepaper has provided valuable insights as you explore how to choose the right models for your enterprise's needs. We welcome your feedback and suggestions for future topics. Stay tuned for more!

Ready to take the next step? Contact us today to set up a meeting and discover how Ententia can be your partner in transforming the future of your operations using Generative AI at scale.