ENTENTIA

ENTERPRISE GEN AI MODEL SELECTION FRAMEWORK

Part I - Going beyond the Benchmarks

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 I: INTRODUCING FRAMEWORK

1. CONTEXT

If you are involved in your company's Generative AI initiatives, you understand that selecting a model or model family is a strategic decision that extends beyond technical and performance considerations. This choice can significantly influence your enterprise's competitive edge by affecting the speed and cost of innovation, which in turn impacts operational and financial performance over the long term.

ABOUT HALF OF REPORTED GEN AI USES WITHIN RESPONDENTS' BUSINESS FUNCTIONS ARE UTILIZING OFF-THE-SHELF, PUBLICLY AVAILABLE MODELS OR TOOLS, WITH LITTLE OR NO CUSTOMIZATION.
McKinsey & Company

While model evaluation is critical, it often focuses on short-term, technical performance metrics such as accuracy, precision, and latency. It's easy to become overwhelmed by rapidly changing leaderboards across various benchmarks. However, model selection is a more nuanced process that must be tailored to the specific needs of an enterprise. It requires a comprehensive assessment of factors beyond performance, including service infrastructure, commercial considerations, sustainability, and governance.

This whitepaper aims to bridge the gap between model evaluation and model selection by introducing a comprehensive, enterprise-centric framework. This framework is designed to guide organizations in making informed, strategic decisions regarding their model selection.

2. MODEL EVALUATION BENCHMARKS

Let's begin by reviewing the landscape of model evaluation and benchmarking, as it remains a crucial factor in the decision-making process. Our intention is not to provide an exhaustive list of available benchmarks but to focus on key benchmark categories that enterprises should consider.

Text Generation Benchmarks

Benchmarks in this category evaluate models' capabilities across various tasks and domains. They typically encompass a wide range of topics, including humanities, social sciences, STEM fields, and more specialized areas like law and medicine. Notable examples include GLUE, SUPERGLUE, MMLU, HELLASWAG, MMLU-PRO, ARC, and SQuAD. This is a rapidly evolving landscape, with new benchmarks frequently introduced to keep pace with model improvements and to prevent benchmark saturation.

Multi-modal Benchmarks

As the name suggests, multi-modal benchmarks assess a model's ability to understand and reason across multiple modalities. For example, a model might be required to interpret text that describes an image or answer questions about a video. These modalities typically include text, images, audio, video, and sometimes other data types like structured tables or sensor data. As enterprise use cases become more complex and multi-modal, these benchmarks will play an increasingly important role in evaluating model performance. Examples of multi-modal benchmarks include VQA, MME, CLIP, and BLINK. As model capabilities advance, we anticipate rapid evolution in these benchmarks.

Code Generation Benchmarks

Code generation is a significant use case for language models. Even when models are not explicitly used for code generation, many enterprise applications rely on a model's ability to generate code to accomplish various tasks. These benchmarks evaluate a model's proficiency in understanding programming tasks, generating correct and efficient code, and handling various programming languages and paradigms. Examples include APPS, HumanEval, ClassEval, SWE-Bench, and BigCodeBench.

Long-Context Benchmarks

Long-context benchmarks are designed to assess models' ability to understand, retain, and generate information over extended sequences of text or other data. These benchmarks test how well a model can maintain coherence, memory, and reasoning across long contexts, which is essential for enterprise use cases such as summarization, long-form conversations, and analyzing lengthy articles and reports. Examples include NarrativeQA, Long-form QA, HumanEval+, and ArXiv Longform QA.

Why does this matter?

The primary reason for highlighting these different types of benchmarks is to guide decision-makers away from focusing on individual benchmarks and the hype surrounding leaderboards. Instead, the emphasis should be on selecting a mix of benchmarks that most closely align with current and future enterprise use cases.

3. Proposed Model Selection Framework

Now that we have a good sense of the capabilities being evaluated across benchmarks, let's build on that to incorporate additional considerations from an enterprise point of view.

We propose a structured, comprehensive framework that categorizes the selection criteria into five key areas: Model Performance, Service Infrastructure, Commercial Considerations, Sustainability, and Governance. Each of these categories addresses different aspects of the decision-making process, ensuring a holistic approach to model selection.

Ententia Gen AI Model Selection Framework

In order to quantify the assessment, we recommend assigning weightage to each of the outlined categories and specific criteria identified below.

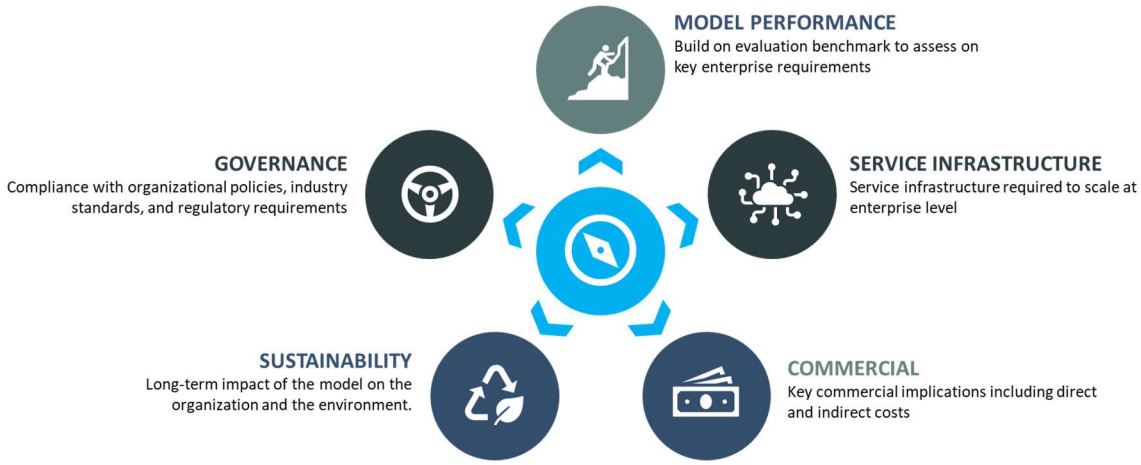

Figure 1: Ententia Gen AI Model Selection Framework

The following sections outline specific assessment criteria within each category.

3.1 Model Performance

We have already discussed in-depth the biggest component of model performance i.e. the benchmark scores. In addition, enterprises should also consider the following criteria:

1. **Industry and/or Domain-specific Capabilities**

   Some models may perform better than others when it comes to knowledge of specific industries or domains, depending on the pre-training dataset or training approach. As a proxy, you may also consider scores on specific topics like math, physics, engineering, etc. if those areas are of importance (e.g., in the industrial sector).

2. **Knowledge cut-off Date**

Depending on the pace of change in specific areas that an enterprise may be interested in, knowledge cut-off date for the model may be an important consideration.

3.2 Service Infrastructure

Service infrastructure encompasses the technical ecosystem required to develop, integrate, deploy, and maintain solutions that leverage language models. The robustness and flexibility of this infrastructure are critical for ensuring that the selected models can be effectively utilized and scaled within the enterprise. Below are the key considerations:

1. **Custom Chatbots and API-based Service Offerings**

Custom chatbots and API-based services are increasingly becoming central to enterprise AI deployments. The ability to integrate language models into these services efficiently and effectively is crucial. This involves:

Customization: The infrastructure should support the development of custom chatbots that can be tailored to the specific needs of the enterprise, including the ability to integrate seamlessly with existing communication platforms, customer service systems, and knowledge bases.

Scalability and Reliability: The API infrastructure must be scalable to handle varying levels of demand, particularly in customer-facing applications. It should also be reliable, ensuring minimal downtime and robust performance across different use cases.

Latency: This is particularly important in use cases that require real-time decision-making. Even when real-time decision-making is not critical, latency can have a significant impact on user experience and lead to adoption challenges.

Support and Documentation: The availability of robust support and comprehensive documentation is essential for ensuring that the model can be effectively leveraged. Enterprises should consider the level of vendor support available, including access to technical experts, training resources, and user communities.

2. **App-dev Playground Environment**

An app-dev playground environment allows developers and data scientists to experiment, iterate, and innovate with language models in a sandboxed setting. Key considerations include:

Ease of Use: The environment should provide user-friendly tools that enable rapid prototyping and testing of AI applications without the need for extensive coding or complex setup. This includes the ability to switch models easily.

Model Switching: Playground environment should also offer ability to easily switch and experiment with different models (including models in preview) – to better assess which model fits the needs of a specific use case better.

3. Data-handling, High-frequency Time-series Data and Documents

Handling diverse and complex data types is a common requirement in enterprise settings. The service infrastructure must be capable of:

High-frequency Time-series Data: Many industrial applications, such as those in manufacturing or energy, generate high-frequency time-series data. The infrastructure should support the efficient ingestion and processing of this data.

Document Handling: Enterprises often deal with large volumes of unstructured documents, such as contracts, reports, and technical documentation. The infrastructure should facilitate the processing of these documents, including OCR capabilities, natural language processing, and integration with enterprise content management systems. Ability to leverage embedding model and a variety of vector databases with different features (e.g., hybrid search) may also become important, especially when dealing with documents that contain images (e.g., engineering design documents). When working with sensitive data, ability to use embedded vector database may provide the necessary level of security.

Data Integration and Interoperability: The infrastructure must support integration with various data sources and ensure that data flows seamlessly between different systems, enabling comprehensive solutions.

4. Model fine-tuning

While RAG-based solutions are highly capable and may suffice for many use cases, model fine-tuning may be necessary to adapt pre-trained model to specific enterprise use cases. We recommend adjusting weightage for this criterion based on perceived needs of the enterprise. The service infrastructure should provide:

Support for Transfer Learning: The ability to leverage transfer learning, where pre-trained models are fine-tuned on domain-specific data, is useful. This reduces the time and resources required to develop high-performing models tailored to specific tasks.

Automation and Optimization Tools: Tools that automate parts of the fine-tuning process, such as hyperparameter optimization or model compression, can enhance efficiency and performance.

5. Private/Hybrid Cloud Deployment

Enterprises often require flexible deployment options to meet regulatory, security, and operational needs. Compatibility with enterprises' existing technology stack can reduce overall solution complexity. The service infrastructure should support:

Private Cloud Deployment: For organizations with strict data security requirements, the ability to deploy models in a private cloud environment is essential. This ensures that sensitive data remains within the organization's-controlled environment.

Hybrid Cloud Capabilities: Hybrid cloud deployment allows enterprises to balance the benefits of both public and private clouds.

Coming up in Part II

In Part II, we will extend the framework by exploring the remaining categories: Commercial Considerations, Sustainability, and Governance. We will also discuss how to apply this framework in practice, including handling edge cases and ensuring that the selected model aligns with the strategic goals of your organization.

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.