ENTENTIA

HARNESSING GENERATIVE AI FOR ENTERPRISE EXCELLENCE

Bridging the Gap between Potential and Performance

Abstract

This whitepaper explores the unique challenges faced by enterprises in integrating Generative AI into their operations. Unlike digitally native counterparts, other enterprises require solutions that must focus on accuracy, consistency and operational safety. We outline the limitations of existing models and propose a systematic, multi-layered approach to addressing them. This whitepaper serves as a guide to understanding and addressing the specific needs, paving the way for Gen AI adoption to drive operational and financial efficiency.

Ententia LLC
info@ententia.ai

ENTENTIA

WELCOME

Welcome to the inaugural whitepaper from Ententia. We are an applied Generative AI startup focused on revolutionizing the way enterprises harness the power of large language models to enhance operational and financial performance across their core business workflows.

We invite you to explore how our focused, value-driven, modular approach combining domain, data and technology can turn Generative AI into a transformative force for your business operations.

CONTEXT

As industries continue to evolve, so must the tools they use.

In recent years, enterprises across various sectors have embarked on a journey to reinvent their core business workflows using AI based solutions and tools. They are delivering value as a result of sustained investments in Data, Technology and Talent. Rise of Generative AI presents a new challenge.

THE ADOPTION RATE OF GENERATIVE AI TECHNOLOGIES IN ENTERPRISES IS SIGNIFICANT, WITH A REPORTED 67% OF ENTERPRISES USING THESE TECHNOLOGIES. THIS MARKS THE FASTEST ADOPTION OF A TECHNOLOGICAL INNOVATION IN RECENT HISTORY

McKinsey & Company

HYPE VS REALITY

While the promise of Generative AI-driven innovation is enticing and expectations are high, the reality of its implementation has been mixed. Digitally native sectors and enterprises have taken an early lead, as expected. However, despite ongoing efforts across industries, the rapid evolution of technology, coupled with a lack of compelling, industry-centered solutions and a shortage of skilled talent, has created a challenging environment for leaders. Moreover, the complexity of integrating Generative AI into existing workflows and the need for significant customization have further slowed down widespread adoption, revealing a gap between the hype and the practical realities.

THE NUMBER OF MENTIONS OF AI IN EARNINGS CALL TRANSCRIPTS HAS INCREASED BY 6X SINCE THE RELEASE OF CHATGPT IN NOVEMBER 2022.

Accenture Research

This whitepaper aims to shed light on the unique challenges faced by these enterprises and how they can tackle these challenges through consistent use of key design principles in building solutions using Gen AI.

ENTENTIA

WHY ENTERPRISES ARE DIFFERENT

From multiple decades of our team's experience of working with small and large enterprises – especially in the industrial sector, we realize that there are distinct characteristics that set them apart from others, when it comes to application of Generative AI:

1. **Operational and Safety Impacts**: Enterprise business workflows are often intricate, involving numerous variables and interdependencies. For core workflows, the stakes are high, and precision is paramount. In this environment, controlling model hallucinations and consistently maintaining accuracy is a key hurdle for adoption – even mor so when personnel safety is at stake.

2. **Compliance**: Regulatory as well as procedural compliance is critical for safe operations in many enterprise environments. Any AI solution must prioritize these aspects to ensure that they operate within the required guidelines.

3. **Domain Expertise**: Enterprises rely heavily on domain-specific knowledge, which must be embedded within the models to ensure relevance and accuracy. This includes making sense of domain-specific information and meaning inherent in many sources of information.

4. **Data Complexity**: Enterprises not only generate vast amounts of data from diverse sources, but they also generate data in a variety of different formats. Data coming from sensors, documents (of wide variety), images (real and simulated), operational and safety procedures, and physics-based simulators are just some examples of that. Acquiring, organizing, managing and making sense of this data requires purpose-built tools.

5. **Data Readiness Challenges**: While investments in data platforms over the last few years have led to significant improvements in quality of structured datasets, unstructured data remains an unsolved challenge. Considering that many enterprise operate in sectors with long lifecycle assets, this becomes especially challenging to incorporate data from older media (e.g., scanned documents or seismic).

6. **User Experience**: Employees and contractors operating in varied work environments, e.g., engineers and operators working in the field or in plants, require tailored interfaces for a better user experience. These interfaces must be intuitive, robust, and designed to handle the unique challenges of their specific work environment.

7. **Geographic & Operational Context:** For industrial enterprises, physical movement is an essential part of business operations. Geographic location and associated operational context are important to incorporate into any Generative AI solution, for maximum impact.

STANDALONE LLMS ARE NOT SUFFICIENT

Large Language Models (LLMs), such as those directly accessible today from many of the leading AI labs, offer remarkable capabilities but fall short in addressing the nuanced needs of a typical enterprise. Here's why:

ACCORDING TO A 2023 MCKINSEY REPORT, 32% OF ORGANIZATIONS THAT HAVE ADOPTED GENERATIVE AI TECHNOLOGIES ARE ACTIVELY MITIGATING RISKS ASSOCIATED WITH INACCURACY, WHICH IS CITED MORE FREQUENTLY THAN CYBERSECURITY OR REGULATORY COMPLIANCE RISKS

McKinsey & Company

1. **Accuracy and Consistency**: LLMs are predisposed to answer user queries, which often leads them to rely on external and at times untrustworthy sources of information. This combined with inconsistency in response is a key point of failure in critical operational context.
2. **Generalization vs. Specialization**: LLMs are designed to be generalists, lacking the depth of domain-specific knowledge required for many core enterprise workflows.
3. **Industry and Company Context**: Understanding the specialized terminology used within different industries and individual companies is crucial for effective deployment. LLMs often lack the context and depth needed to interpret and utilize this terminology correctly.
4. **Data Integration Challenges**: Enterprises need AI that can seamlessly integrate with incredible variety and complexity of both structured and unstructured data sources, as outlined in the previous section. In addition, they sometimes lack ability to handle high-frequency datasets (e.g., time-series sensor data), which are very common in industrial settings.
5. **Security and Compliance**: Ensuring data security and regulatory compliance is non-negotiable in industrial settings. Generic LLMs do not inherently prioritize these aspects.
6. **User Experience**: Nature of operations in a non-traditional work setting (e.g., field work) dictates that out-of-the-box user interfaces are not effective and may lead to challenges with widespread solution adoption.

ENTENTIA

How do we tackle these challenges?

As we consider how to address the challenges outlined in the previous sections, it's crucial to adopt an approach that remains agnostic to any specific Large Language Models (LLMs), applies effectively across a wide range of enterprise use cases, and is sustainable from an application development and maintenance perspective.

At Ententia, we've spent the past several months experimenting with various approaches, particularly in applying Retrieval-Augmented Generation (RAG) and fine-tuning techniques. Through this iterative process, we've identified an approach that works best to achieve the goals outlined above.

Layered Design

The first key architectural principle we employ is a layered design. This involves creating distinct logical layers within the architecture, each responsible for specific features and capabilities, which collectively work to overcome the outlined limitations.

- **LLM Layer**: The first layer is your choice of LLM, which provides the foundational capabilities of a state-of-the-art model. While selecting the appropriate model or model family is critical, especially in an enterprise setting, we'll delve deeper into this topic in a future discussion.
- **Industry Context Layer**: The second layer integrates industry-specific knowledge, enabling the model to understand key terminology and recognize what's important to the users within a particular industry and sub-industry domain.
- **Use Case Context Layer**: The third layer focuses on understanding the specific problems users aim to solve and the typical approaches they might take. Here, we also identify the relevant internal and/or external datasets needed to provide accurate and meaningful responses.
- **Customer Context Layer**: The fourth layer introduces company-specific details such as organization structure, asset hierarchies, key data sources, available datasets (structured and unstructured), and samples of user queries and responses.

These layers collectively equip the model with the necessary tools to navigate enterprise environments effectively, comprehend user inquiries, and leverage information from diverse sources. We refer to this as the full operational context, enabling users to extract maximum value from LLM-based solutions.

GOVERNANCE & ORCHESTRATION

While technically, this is the next layer in our design, we thought it was important to call it out as a design principle, considering the distinct role it plays in the overall architecture.

Governance & Orchestration layer performs many responsibilities, including but not limited to:

1. Defining and managing boundaries of a solution, which helps prevent hallucinations while also managing user expectations
2. Parsing complex user queries into manageable tasks and orchestrating their execution within the solution
3. Personalizing responses based on user roles and managing access control
4. Handling transitions in multi-agent scenarios

And more...

This layer acts as a stabilizer, ensuring that accuracy and consistency are maintained throughout the solution.

USER EXPERIENCE

Finally, the focus on User Experience (UX) is critical. While it may be a cliché to emphasize UX, it plays a pivotal role in successful user adoption. The choices made in this area will be reflected not just in the user interface but in the overall user interaction with the system. Key considerations include:

1. Integrating the proposed solution within existing enterprise workflows, applications, dashboards, and tools
2. Tailoring interface features and functionalities to the selected use case, including but not limited to suggested prompts, complex analyses, and specific visualizations
3. Providing role-based personalization to enhance relevance and usability
4. Offering the ability to customize the solution to meet user specific needs, depending on the use case

These elements need to be carefully designed to meet the unique needs of an enterprise. This lightweight, modular design can be easily adapted to a variety of enterprise use cases with limited additional effort. Consistent application of these principles will also allow enterprises to mature and sustain these solutions systematically.

ENTENTIA

GEN AI REVOLUTION OR EVOLUTION?

Our approach to designing Generative AI solutions draws heavily on established best practices from traditional software and application development. While Generative AI technology offers unprecedented new capabilities, successfully integrating it at scale within an enterprise environment requires us to build upon and evolve the proven methodologies of software development.

Or conversely, what is the cost of NOT taking a systematic and thoughtful approach? It could manifest in various ways: inflated costs due to overly complex implementation, subpar performance that fails to meet business objectives, and a lack of sustainability that burdens future operations with technical debt. In some cases, neglecting a careful approach could even result in legal repercussions, particularly if data privacy, security, or compliance issues arise from poorly designed solutions.

The proposed approach not only facilitates the seamless adoption of Generative AI but also allows enterprises to leverage their existing investments in data and technology platforms, including traditional AI systems. In our view, these investments will remain crucial, often serving as the foundational infrastructure upon which Generative AI solutions can be effectively developed and deployed.

WHAT'S NEXT?

We hope this whitepaper has provided valuable insights as you explore how to harness the power of Generative AI for your enterprise. 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.

ENTENTIA

At Ententia, our mission is to help enterprises harness the power of Generative AI. Our value-driven, focused approach to products and services help enterprises accelerate their Generative AI journey.

Get in Touch

https://ententia.ai

info@ententia.ai

Ententia-ai