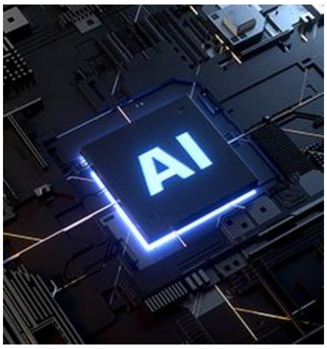
PL
POWERING SELF SERVICE
THE EVOLUTION OF DIGITAL CHANNELS
ENHANCING CX WITH EMBEDDED PAYMENTS WITH TransUnion
UNLOCKING ENTERPRISE ROI WITH AI WITH IBM
SECURING NETWORKS IN THE AI AGE
MOVING THROUGH DIMENSIONAL DATA
PROTECTING CRITICAL INFRASTRUCTURE
ENTERPRISE PRIVATE WIRELESS NETWORKS
AI TAX COLLECTION
THE FUTURE OF GENAI
THE HUMAN-CAPITAL GAP
TOP TRENDS

www.pipelinepub.com
Volume 21, Issue 2

# Unlocking ROI with GenAI: A Guide to Leveraging Enterprise Data and Ensuring Scalability

By: Volkmar Uhlig



Generative artificial intelligence (GenAI) is reshaping industries worldwide, driven by significant advancements in models such as OpenAI's ChatGPT. This transformation offers businesses the potential for increased productivity, reduced operational costs, and streamlined workflows. However, with these benefits come complex challenges, especially concerning data privacy, security, and compliance with industry regulations. As companies adopt GenAI, they must proactively address these risks by establishing robust policies designed to securely integrate AI technologies into existing systems and processes.

### Leveraging Proprietary Data for Maximum ROI

For organizations aiming to maximize GenAI's return on investment (ROI), effectively utilizing proprietary data is essential. With the right approach, companies can tap into their unique data assets to improve model performance and produce results aligned with specific business goals. Two primary methods, Retrieval Augmented Generation (RAG) and fine-tuning models, are particularly effective for this purpose.

RAG works by supplementing the outputs of a large language model with business-specific datasets, creating responses that reflect specialized knowledge and industry nuances. This technique enables the GenAI model to reference proprietary information, allowing it to produce more informed and accurate results in response to user queries. While fine-tuning involves adjusting a pre-existing model to align more closely with proprietary insights and organizational needs. This method can be especially useful for organizations with highly specialized terminology or industry-specific requirements. By customizing a model with company data, businesses can ensure more tailored and precise outputs that resonate with their brand voice and meet specific functional demands.

Both approaches hinge on having well-curated, high-quality data resources. Without reliable data, the effectiveness of GenAI decreases, as it becomes more prone to generating inaccurate or irrelevant responses. For companies investing in GenAI, building and maintaining strong data management practices is paramount.

### Key Phases of Enterprise AI Implementation

Implementing GenAI within an organization typically follows a five-phase approach. Each phase reflects a progressive commitment to GenAI, with corresponding shifts in complexity and customization. The five stages are as follows:

*Initial Experimentation*: Companies often start by experimenting with existing GenAI models to understand their capabilities and limitations. This phase is exploratory, allowing organizations to evaluate GenAI's potential value with minimal risk and investment.

*Employing RAG for Enhanced Outputs*: Once familiar with GenAI, companies can implement RAG techniques to enrich responses by incorporating industry-specific datasets. This step improves the model's relevance for business applications, providing a bridge between generic AI capabilities and domain-specific expertise.

*Transitioning to Leaner Models*: To meet performance demands, organizations may opt for leaner models optimized for quicker response times. This phase is essential for businesses with high transaction volumes or customer interactions, as it ensures rapid processing without sacrificing accuracy.

*Refining Models for Firm-Specific Competencies*: As GenAI use becomes more embedded, organizations can start refining models with firm-centric competencies. This involves tailoring the model to understand unique business processes, language, and objectives, resulting in outputs that are not only relevant but also actionable.

Customizing Models with Institutional Knowledge: The final phase of implementation involves extensive customization, embedding the organization's institutional knowledge within the model. At this stage, GenAI becomes deeply integrated into the organization, enabling it to serve as a reliable tool for decision-making, customer interactions, and internal operations.

Each phase requires meticulous planning and a keen focus on data handling and integration. The journey from experimentation to full-scale deployment is marked by iterative adjustments, ensuring that GenAI is optimized for business goals and compliant with regulatory standards.

**Choosing Between RAG and Fine-Tuned Models**

Deciding whether to employ RAG or fine-tuned models depends on various operational and compliance factors. The decision should be informed by considerations such as:

*Frequency of Data Changes*: For organizations with rapidly evolving data, RAG may be more suitable, as it allows models to incorporate updates without retraining. Fine-tuning is more effective for stable datasets.

*Access Permissions and Security*: RAG systems enable more controlled access to sensitive data, an important feature when dealing with high-security information.

*Complexity and Novelty of Information*: Fine-tuned models are often preferable for handling complex tasks that require in-depth comprehension of proprietary knowledge. RAG is generally better for simpler or more repetitive tasks.

*Compliance and Licensing*: Fine-tuned models are often subject to stricter compliance and licensing requirements, especially regarding copyright and intellectual property.

While both approaches have advantages, RAG can be more flexible for organizations with stringent compliance needs or fast-changing information. However, fine-tuned models offer deeper customization and are ideal for organizations looking to create a proprietary model that closely mirrors their brand and values.

**Advantages of RAG Systems**

RAG systems have unique advantages in terms of flexibility and scalability. They work by integrating diverse inputs — emails, chat logs, filesystem contents, coding archives, etc. — in formats optimized for machine processing. This data preparation requires various steps, such as transcription services for audio/video, parsing mechanisms for code, and indexing solutions for fast lookups. By converting raw data into machine-readable formats, RAG systems can deliver highly relevant information at scale.

Additionally, enhanced search functionalities within RAG frameworks make it easier to scale with expanding data volumes. These systems allow organizations to maintain privacy safeguards linked to original data sources, which is critical for companies handling sensitive customer or proprietary information.

### Benefits of Fine-Tuned Models

Fine-tuning offers companies more control over model outputs by aligning responses with industry-specific terminology and corporate standards. Despite token limitations in current language models, fine-tuning enables more contextually relevant outputs that capture the unique language, tone, and intent of the organization.

For instance, Red Hat InstructLab on Cloud allows businesses to craft responses that reflect both technical accuracy and brand integrity. Fine-tuning also supports nuanced language use, ensuring that outputs resonate with audiences and adhere to organizational standards. This approach is particularly valuable for customer-facing applications, where accuracy and tone are paramount.

### Overcoming Scalability Challenges

To scale AI effectively, companies must manage resources strategically, addressing both technical and operational challenges. GenAI’s resource-intensive nature requires organizations to optimize deployments, balancing cost, speed, and scalability. This involves segmenting information assets and sub-sampling training data to mitigate the risk of triggering catastrophic forgetting during fine tuning, where the model may forget prior knowledge in favor of the new training data. Scalability is further enhanced by segmenting data for optimized use. This approach allows models to retain critical information while maintaining performance. By strategically managing data, organizations can extend the model’s lifespan and ensure consistent quality across applications.

### Legal and Governance Considerations

One critical consideration in GenAI adoption is choosing a reliable base model. With thousands of open-source options available, many come with licensing restrictions and lack indemnification. ’s Granite model suite addresses these concerns by offering an exclusive model where retains full oversight of training data. This assurance reduces legal risk, as companies can trust that the base model aligns with compliance standards.

Operating without such safeguards can expose businesses to potential legal issues, especially if model outputs infringe on intellectual property rights. For companies focused on legal compliance, selecting a model provider that offers indemnification can streamline GenAI implementation and reduce associated risks.

### Model Development, Testing, and Scaling for Production

Effective model development involves a disciplined approach to data curation, training, and testing. Key metrics, such as RAG chunking accuracy, lookup quality, and inference performance, should be defined and monitored throughout the model’s lifecycle. Automated testing protocols similar to traditional software testing help identify and resolve issues early, ensuring that the model functions as intended.

Large-scale deployments require seamless integration between software and hardware, with redundancies in place for fault tolerance. Companies should assess performance metrics like dataset volume and refresh rates, as calculating embeddings and indexing requires substantial resources. Beyond core language models, additional models, such as those for ranking and query refinement, support improved functionality and compliance.

Determining the right hosting approach whether on-prem, in the cloud, or hybrid, depends on a fit for purpose approach that evaluates an organization’s capacity, security requirements, and strategic goals for data location and processing. Choosing the optimal setup ensures scalability, performance, and compliance, aligning the AI infrastructure with business objectives.

### Conclusion

GenAI for the enterprise is indeed rapidly advancing and shows no signs of slowing down. To stay ahead, organizations should begin developing their AI models immediately. Embarking on this journey presents fresh obstacles, yet many of these hurdles mirror conventional IT challenges. Effectively integrating various elements—including hardware, software, databases, models, and data—is demanding and introduces unfamiliar processes. Fortunately, assistance is readily accessible to navigate these intricate landscapes successfully.