unvired
Digital Delivered

The Potential of Generative AI for Business Innovation

Image generated using Adobe Firefly

Table of Contents

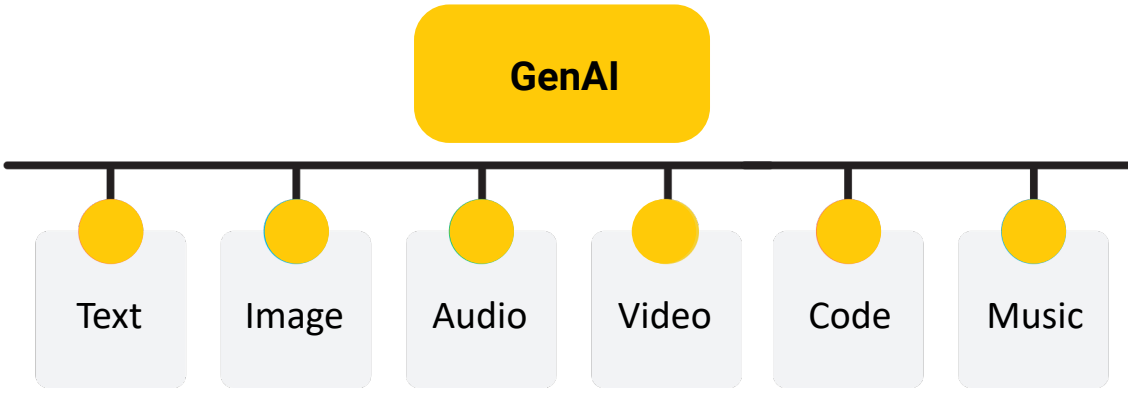
Introduction: What is GenerativeAI?

In today's rapidly evolving digital era, the possibilities for innovation seem limitless, thanks to the emergence of Generative Artificial Intelligence (GenAI). GenAI is a transformative branch of artificial intelligence that can generate new content across various mediums, including text, images, audio, code, and videos. It achieves this by leveraging cutting-edge technologies such as Large Language Models, Vector Databases, Prompt Engineering, Fine-tuning, and more, all of which enable computers to perform tasks with a level of human-like sophistication that was once unimaginable.

GenAI is not merely a technological advancement; it is a paradigm shift reshaping the landscape of business operations across industries. This transformative power has the potential to elevate productivity, stimulate innovation, and fuel the growth of businesses in ways previously thought unattainable.

Unvired, a GenAI solutions provider, invites you to co-innovate and explore with us. Let's redefine how businesses operate and revolutionize digital experiences with GenAI. Discover its possibilities for your organization's future in this eBook.

GenAI is different from traditional AI, usually referred to as Predictive AI. One of the main differences is that GenAI has creative and generative capabilities. While Predictive AI makes predictions and decisions based on predefined rules and historical data, GenAI does more than that - it creates entirely new content across a broad range of media, including crafting text, code, and audio and developing images, music, and videos.



Key Components of GenAI Building Blocks

1. Large Language Models (LLMs):
LLMs are a class of artificial intelligence models designed to process and generate human language. They can capture complex language patterns and structures.

2. Vector Databases:
A vector database contains numerical representation of the of textual data that enables semantic searching. It facilitate efficient retrieval and manipulation of data for AI applications.

3. Prompt Engineering for Contextually Relevant Results:
Prompt engineering results in crafting well-designed prompts or queries and fine-tuning them to elicit the desired information or behavior from LLMs.

4. Fine-Tuning for Task-Specific Performance:
Enhances task-specific performance through the customization of pre-trained LLMs and the refinement of model performance and adaptability for specialized use cases.

These building blocks combine to create robust GenAI systems capable of understanding, generating, and manipulating text data for various applications.

Popular GenAI Models and Platforms

GenAI has seen widespread adoption due to the availability of various platforms and models that empower businesses to harness its capabilities effectively. Here are some of the popular platforms and models that have gained significant traction:

OpenAI: GPT, DALL·E

OpenAI is a leading organization in the field of GenAI, responsible for groundbreaking advancements for GPT (Generative Pre-trained Transformer).

Google: Vertex AI Platform

Vertex AI is a machine learning (ML) platform that lets you train and deploy ML models, AI applications, and LLMs for using in your AI-powered applications. Google offers a Gen AI App Builder that allows developers to build apps using a combination of text and other modalities such as images and videos for improved customer interactions.

Amazon: SageMaker & Bedrock

Amazon offers a variety of Generative AI tools to create a wide range of generative AI apps:

- **Amazon SageMaker** is a ML platform offered by AWS. It provides a comprehensive set of tools for building, training, and deploying machine learning models.
- **Bedrock** is a fully managed service that makes foundation models (FM) available via an API.

Microsoft: Azure Open AI

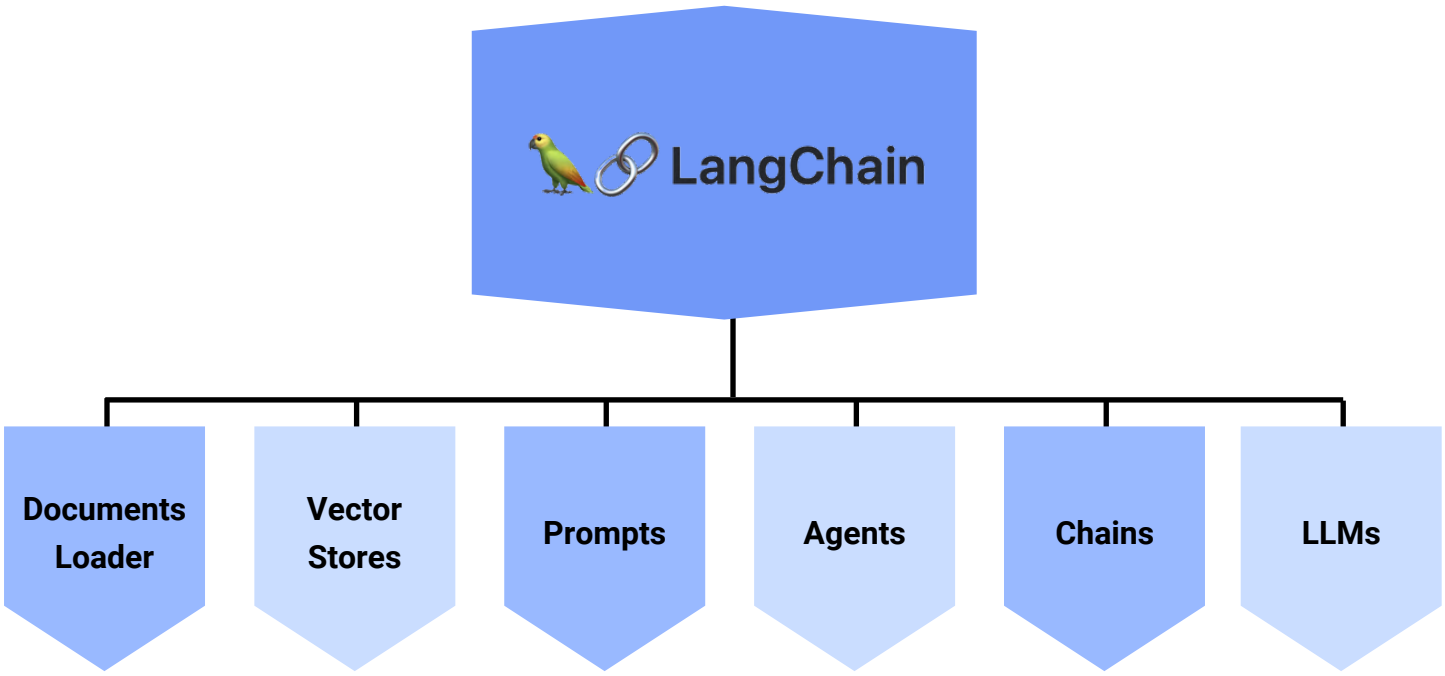
Microsoft Azure OpenAI Service is a cloud-based platform that provides access to powerful language models, including GPT-4, GPT-3, Codex, and DALL-E, enabling the seamless creation of various generative AI applications.

Hugging Face

Hugging Face is a large open-source community that builds tools for users to create, train, and deploy machine learning models. The company focuses on natural language processing (NLP) and offers a platform and libraries for building and working with generative models.

LangChain

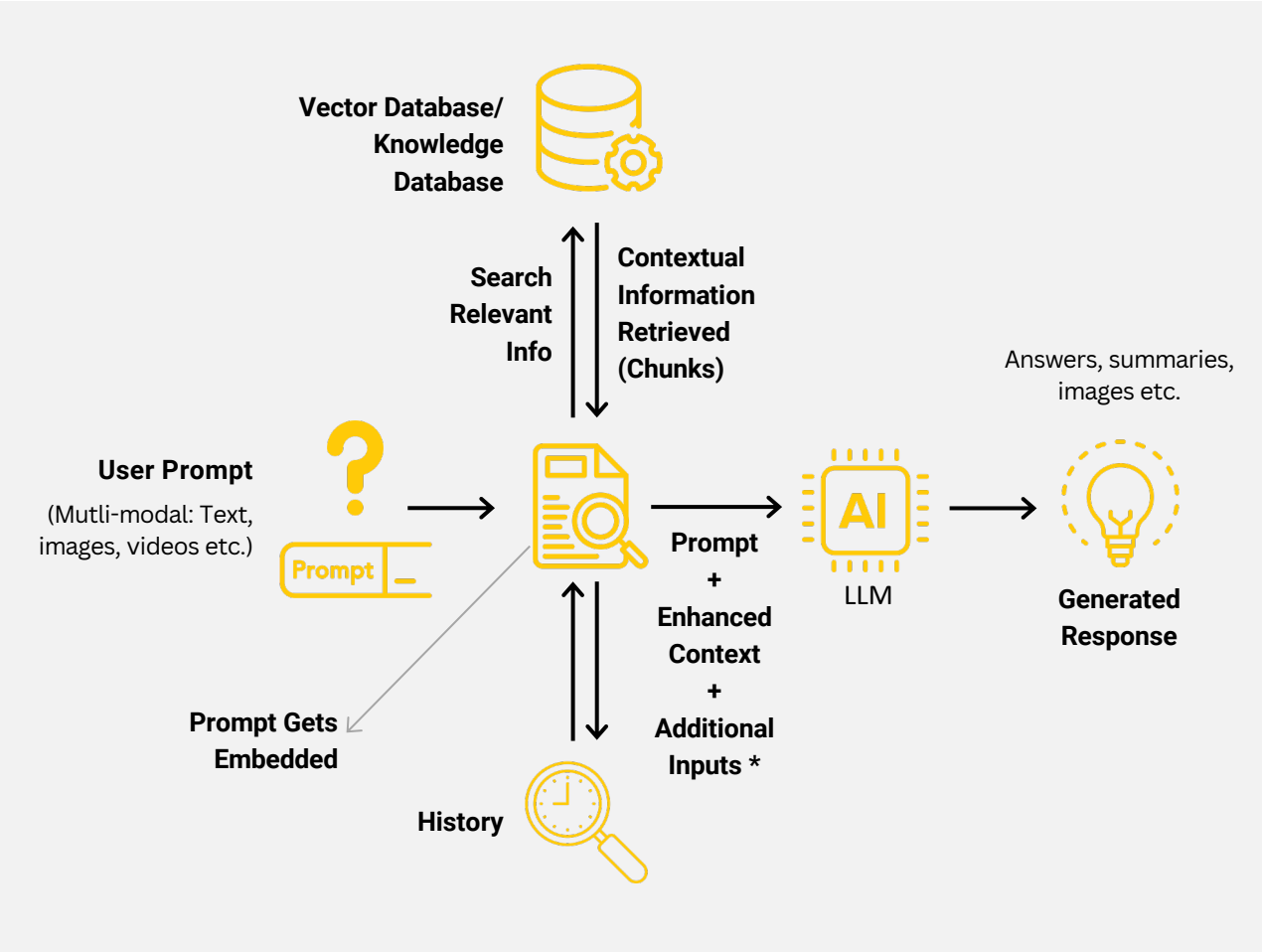
LangChain is a framework that can be used to create generative AI applications using large language models (LLMs). It provides a set of tools and components that make it easy to build and deploy applications that can generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.



Note: *It's important to note that categorizations may evolve over the time as the Generative AI landscape continues to evolve and new vendors and models emerge.*

Reference Technical Architecture

GenAI operates by leveraging complex algorithms that enable it to autonomously generate new solutions, designs, or simulations based on patterns it learns from extensive datasets and predefined parameters.



Additional Inputs *

Guardrails: Ensures the Quality of Outputs

Guardrails are essential for maintaining the quality and safety of LLM-generated content. These mechanisms filter out inappropriate, harmful, or inaccurate outputs. Guardrails can include content filters, sentiment analysis, and other checks to ensure the generated content aligns with ethical and quality standards. This pattern is crucial to prevent the model from producing harmful or biased content, a significant concern in AI and NLP.

Personally Identifiable Information (PII): is protected by using data anonymization techniques to hide or replace sensitive information that can be traced back to individual identities, safeguarding user privacy.

GenAI- Challenges

Data Challenges:

- **Data Governance:** Ensuring data is managed, stored, and used in compliance with regulations and internal policies.
- **Unstructured Data:** Handling and extracting insights from non-tabular, messy data formats like text, images, and audio.
- **Integration:** Seamlessly incorporating GenAI solutions into existing systems and workflows.

Technology Challenges:

- **Choice of Tools/Models:** Selecting the right AI tools and models for specific business needs and goals.
- **LLMs not company-specific:** Large Language Models (LLMs) may need more company-specific knowledge, requiring customization.
- **Rapid Advancements:** Keeping up with the rapidly evolving GenAI technology landscape.

Risks/Concerns:

- **Data Security:** Safeguarding sensitive data from unauthorized access or breaches.
- **Hallucinations:** Addressing the risk of AI generating inaccurate or misleading content.
- **Model Bias:** Mitigating biases in AI models that can lead to unfair or discriminatory outcomes.

Readiness Challenges:

- **Skills/Knowledge Gap:** Building a team with the necessary GenAI development and deployment expertise.
- **Costs/Scalability:** Managing the financial aspects of GenAI implementation and scaling.
- **Deciding Where to Begin:** Identifying the most suitable use cases and starting points for GenAI integration.
- **Pilot to Production:** Transitioning from initial pilot projects to full-scale production while maintaining efficiency and effectiveness.

Unvired Accelerators

1. Eureka: Unvired GenAI Knowledge Assistant

Eureka is a GenAI Knowledge Assistant app that leverages advanced artificial intelligence to generate human-like text responses to queries to streamline enterprise asset management. It analyzes and generates contextually relevant information for users to make informed decisions, optimize asset utilization, and enhance operational efficiency.

Document Processing

- Text Extraction
- Analysis
- Document Summarization
- Entity Search- PO Nos, Customer address, etc.

Enterprise Search

- Q&A
- Conversational Interface/Chat
- Structured Data
- Unstructured Data- PDFs
- Websites

Analytics

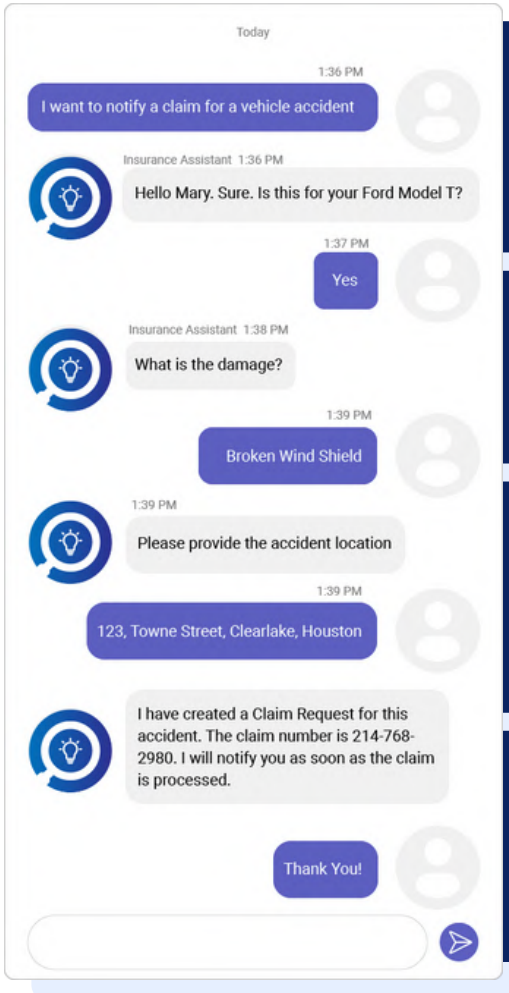
- Natural Language queries
- Query CSVs
- Visualize Charts/Dashboards

Scenarios:

Maintenance Manuals, Legal docs, Contracts, Healthcare Policy docs.

2. Unvired Eureka Agents: Streamlining Workflows and Tasks

Unvired Eureka Agents represent a significant leap forward in the realm of AI-driven automation in industrial organizations. They are designed to seamlessly integrate into various business processes, performing tasks and workflows with precision and intelligence. Whether you're in enterprise asset management, IT service management, asset maintenance, or Insurance management, Eureka Agents have something to offer. Some of their critical applications:



Eureka for SAP simplifies interactions with SAP systems, allowing users to execute tasks and access critical information through natural language commands.

Eureka for ServiceNow/ITSM streamlines IT service management by efficiently handling service requests, incident management, and more.

Eureka for Asset Maintenance ensures optimal equipment and infrastructure performance. It schedules tasks, tracks asset performance, and predicts maintenance needs based on data analysis.

Eureka for Claims Management automates insurance claim filing, verification, and processing while assisting with fraud detection and analytics to improve efficiency and customer satisfaction.

Accessibility:

Eureka Agents are highly accessible from different channels such as Microsoft Teams, websites, and messaging apps.

Deployment Options: Public SaaS, Public Cloud, and Private Hosting

Organizations evaluating the deployment of GenAI applications must weigh multiple alternatives, ranging from on-premise setups to public and virtual private clouds, where the interplay between costs and security plays a pivotal role in shaping the deployment decision.

Deployment Options with some examples:

- **Public SaaS:** OpenAI
- **Private Cloud:** Azure OpenAI, Google Vertex AI, AWS Bedrock
- **Private Hosting:** Hugging face

Safeguarding GenAI: Addressing Risks and Mitigation

While Generative AI opens up new avenues for creativity and innovation, it also presents unique risks/ data security challenges. This section explores the potential risks and outlines strategies for safeguarding Generative AI applications to mitigate those risks:

The **Open Worldwide Application Security Project (OWASP)** is a nonprofit foundation that works to improve the security of software. They produces freely-available articles, methodologies, documentation, tools, and technologies in the field of web application security.

OWASP has listed down the top 10 important vulnerability types for Artificial Intelligence (AI) applications built on Large Language Models (LLMs)

Prompt Injection	LLMs are at risk of prompt injection, which can lead to hidden manipulations and unauthorized actions that benefit attackers.

Insecure Output Handling	When plugins or apps blindly accept LLM outputs without adequately monitoring them, they lead to harmful actions that ultimately enable agent hijacking attacks.
Training Data Poisoning	LLMs learn from a wide variety of text, but there's a risk of bad data influencing their learning, leading to incorrect information. Overreliance on AI has its drawbacks.
Denial of Service	An attacker uses an LLM to demand a lot of resources, which can lead to the LLM's performance dropping for them and others, or it might result in significant resource expenses.
Supply Chain	LLM supply chains can face reliability issues from biases, security concerns, or system failures caused by pre-existing models, crowdsourced data, or plugin extensions.
Permission Issues	Improper plugin monitoring can result in prompt injections and unsafe usage, leading to confidentiality loss, privilege escalation, and remote code execution.
Data Leakage	Data leaks in LLMs can compromise privacy and security. Careful data cleaning and clear model usage guidelines are crucial to avoid this.
Excessive Agency	LLMs' APIs need safeguards to prevent unintended actions when connected to other systems, like web apps.
Overreliance	LLMs can generate incorrect information due to "hallucinations" and cause legal issues if not appropriately monitored.
Insecure Plugins	Plugins connecting LLMs to outside sources may have vulnerabilities if designed to accept any text input, allowing harmful requests to bypass security measures and potentially leading to unintended actions or remote code execution.

Source: [OWASP's Top 10 Risks or Vulnerability Concerns for LLM]

Cost of Generative AI

Understanding the cost dynamics of Generative AI is essential for businesses considering its adoption. The pricing of Gen AI is typically based on the number of tokens used, with tokens representing fragments of words. For instance, OpenAI charges per 1,000 tokens, roughly equivalent to 750 words. Using the GPT-4 model with 8k context as an example, the cost is approximately $0.03 per 1,000 tokens for Input and $0.06 per 1,000 tokens for Output.

Proof-of-Concept (POC) costs are usually affordable, but optimizing cost-effectiveness requires assessing various models and platforms. Select the right solution that balances performance and budget, aligning with your specific use case and needs. Thorough evaluation ensures businesses maximize value while controlling expenses.

The cost of productizing GenAI apps needs to be evaluated by considering various factors that would impact the usage of compute and storage services required.

Quick Start with POC: Initiating Your Generative AI Journey

01

Select the Use Case: When beginning your GenAI journey, it's essential to identify the specific use case that aligns with your business objectives. Knowledge Management, Customer Experience, and Analytics are common use cases.

02

Get Ready: Dedicate time and allocate the necessary talent and resources for your GenAI initiatives. Evaluate whether you have the datasets available that can be leveraged and determine the budget you can allocate to these initiatives.

03

Define the solution: We recommend starting with an existing Foundation LLM (open source or closed), and leveraging prompt engineering/ Retrieval Augmented Generation (RAG) for specific use cases. There are various frameworks like LangChain to build the Gen AI app and solution approaches.

04

Start Small: It is best to start with a low-risk business scenario and start small. As you learn, you can then grow the POC to a pilot. Evaluate Multiple LLMs. Implement various approaches: Retrieval Augmented Generation (RAG) or Fine Tuning.

05

POC Timeline: POC timeline for Generative AI applications typically spans 4 weeks.

Unvired Services for GenAI

Generative AI Consulting Workshop: Tailored strategies for industrial organizations leveraging GenAI to upskill and develop innovative approaches to solve complex challenges.

GenAI Ops/Productization: Unvired Accelerators simplifies the process from concept to deployment, aligning with your needs. Assemble various components quickly with the GenAI framework.

GenAI App Development: Unvired provides custom GenAI apps to meet your specific requirements while ensuring security and scalability.

Accelerators: Streamline GenAI integration with pre-built frameworks, algorithms, and expert guidance for continuous improvement and cost optimization.

About Unvired

Unvired is a Certified SAP partner company that offers Digital Solutions with a focus on AI applications. We have created a Generative AI Center of Innovation to enable businesses to benefit from this technology. Unvired has developed Gen AI apps for Knowledge Management, Analytics, and has also developed apps for Predictive AI.

Case Studies

Here are a few customer case studies showcasing successful implementations with significant gains:

U.S Steel

Steel Production

United States Steel Corporation (U.S. Steel) and Google Cloud have joined forces to leverage Google Cloud's generative artificial intelligence (gen AI) technology in enhancing operational efficiencies and employee experiences at the largest iron ore mine in North America. Their first gen AI-driven application, MineMind™, will streamline equipment maintenance by offering optimal solutions for mechanical issues, leading to time and cost savings, and ultimately boosting productivity. Powered by Google Cloud's AI technology, the application is set to significantly reduce work order completion times by an estimated 20%, offering guidance for repairs, parts ordering, and access to comprehensive references.

TIME

Storied Media Company

TIME, in collaboration with Google Cloud, is leveraging generative AI to transform its relationship with readers and build a stronger sense of community. Unlike the traditional one-way content delivery model, TIME aims to create two-way interactions with its audience by using generative AI prompts and chatbots. Burhan Hamid, the Senior Vice President at TIME, views generative AI as a tool for fostering community engagement. By using AI-generated content, such as summaries of articles and transcripts, TIME envisions enhancing its editorial efficiency and delivering personalized experiences. The magazine intends to harness the trust inherent in its journalism to provide valuable and reliable AI-generated content, helping create a more interactive and engaging platform.

Unvired Solutions

- Generative AI App Development Services
- Data Analytics
- SAP Mobile Asset Management
- Digital Forms
- SAP BTP Services

Contact Us

Unvired.Inc

UNVIRED.COM

+1-713-560-2760

SALES@UNVIRED.COM

THE ION 4201 MAIN STREET Houston, TX 77002