

perimattic

Beyond Text: Building Multimodal AI Assistants For Enterprise Use

SEPT 2025



Table Of Contents

1. Introduction
2. Why Multimodal AI Is a Necessity for Modern Enterprises
3. Multimodal AI Agents: Game-changers for Enterprises
4. Benefits in the Practical World
5. The Power of Real-World Application
6. Factors to Consider Before Opting for Multimodal AI Agents
7. Roadmap to Building Multimodal AI Agents

Introduction



Enterprises today operate in an environment where data is generated in unprecedented volumes and formats. Beyond emails, documents, and databases, organizations now contend with voice recordings, customer call logs, medical images, surveillance video, IoT sensor feeds, charts, dashboards, and even subtle human cues such as tone of voice or facial expressions. In this setting, text-only artificial intelligence (AI) is no longer sufficient to build truly intelligent, knowledge-driven systems.

This is where **Multimodal AI assistants** emerge as the next frontier. Unlike traditional AI that relies primarily on text-based training, multimodal AI integrates and processes **diverse data inputs such as text, speech, images, video, audio, and numerical data, all into a unified understanding**. By doing so, these systems can deliver richer, more context-aware insights that mirror the complexity of real-world enterprise operations.

The market momentum reflects this shift. According to industry research, the **global multimodal AI market size was estimated at USD 1.73 billion in 2024 and is projected to reach USD 10.89 billion by 2030, growing at a CAGR of 36.8% from 2025 to 2030**. This remarkable growth underscores how organizations increasingly recognize that their competitive advantage depends not on the quantity of data collected, but on their ability to interpret, correlate, and act on it holistically.

The rise of Retrieval-Augmented Generation (RAG) pipelines has accelerated this transformation. By retrieving data from structured and unstructured enterprise knowledge bases, RAG systems provide AI assistants with the most relevant, up-to-date, and contextual information to generate precise responses. When coupled with multimodal AI, enterprises can move beyond static Q&A to build assistants capable of performing advanced analytics, interpreting dynamic dashboards, recognizing customer sentiment, and even interacting naturally in human conversations.

The purpose of this whitepaper is to look at the role of multimodal AI assistants in the enterprise landscape. Specifically, we will examine:

- **What makes multimodal AI a necessity for modern enterprises**
- **Why multimodal AI agents are game-changers and the benefits they bring**
- **Examples of current multimodal AI systems**
- **Key factors to consider before adopting multimodal AI solutions**

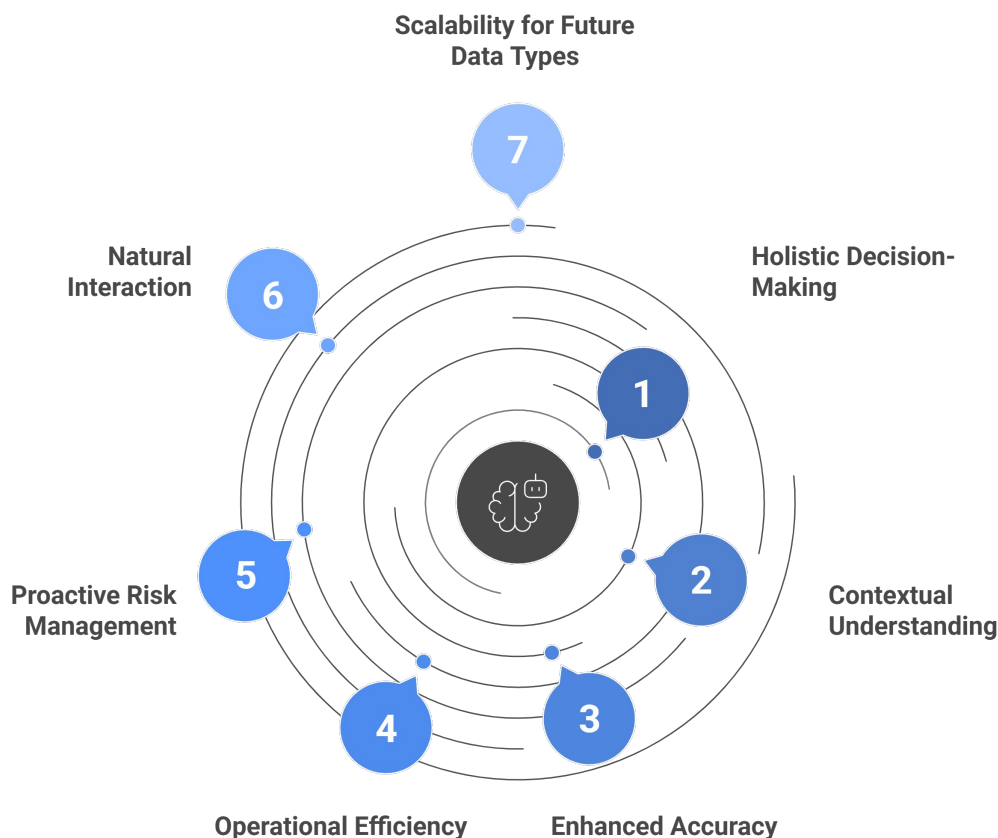
In doing so, we will highlight how enterprises can transition from fragmented, single-mode analysis to integrated, multimodal intelligence, unlocking new levels of decision-making, efficiency, and innovation.

Why Multimodal AI Is a Necessity for Modern Enterprises

The enterprise data landscape has shifted irreversibly. Consider a modern financial services firm: its operations involve transaction records (structured data), compliance documents (text), customer calls (voice), fraud detection footage (video), and market trends (charts, graphs, and time-series data). Limiting AI assistants to text-only inputs leaves massive blind spots, hindering their capacity to provide actionable insights.

Multimodal AI closes this gap. By combining inputs across diverse formats, enterprises can extract deeper patterns, identify correlations, and deliver more robust decisions. In other words, multimodal AI transforms fragmented information into an integrated intelligence layer.

What Makes Multimodal AI a Necessity for Modern Enterprises



Traditional enterprise AI was designed to solve narrow problems: chatbots for customer service, dashboards for analytics, or voice transcription for compliance. However, as business environments became more interconnected, this siloed approach revealed significant inefficiencies.

For example:

- A customer support assistant that processes chat transcripts but cannot analyse customer call recordings fails to capture emotion, intent, or urgency
- A fraud detection engine that reviews transactional data but not video footage from ATMs risks missing critical anomalies.
- A healthcare diagnostic assistant trained on patient notes but not imaging data is incomplete at best, dangerous at worst.

Enterprises cannot afford these blind spots. Multimodal AI represents not just an upgrade but a necessity.

Additionally, the rise of **Retrieval-Augmented Generation (RAG)** pipelines is amplifying this necessity. Enterprises can no longer rely on static, pretrained models that are quickly outdated. Instead, RAG enables assistants to pull the latest knowledge from enterprise repositories—documents, reports, databases, and APIs, before generating responses. When enhanced with multimodal analysis, these assistants deliver insights that are both current and context-rich.

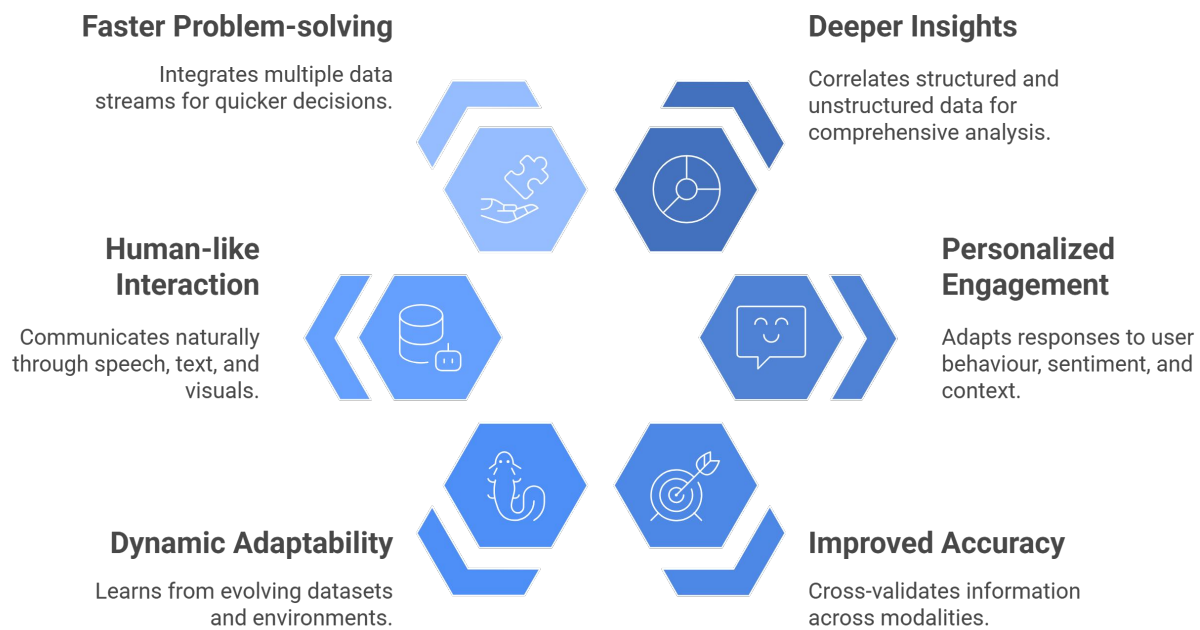
The necessity is further magnified in sectors where compliance, accuracy, and timeliness are mission-critical:

- **Healthcare** – AI that analyses medical imaging, physician notes, and patient vitals simultaneously.
- **Finance** – AI that integrates market sentiment analysis with structured financial data.
- **Manufacturing** – AI that combines IoT sensor data with video monitoring to predict and prevent failures.

Multimodal AI Agents: Game-changers for Enterprises

If multimodal AI represents necessity, then multimodal AI agents represent impact. These agents are not passive analytical systems; they are **interactive, adaptive, and autonomous entities** capable of assisting humans in real-time decision-making.

Benefits of Multimodal AI Agents



Unlike traditional AI systems, multimodal AI agents replicate the way humans process information. A human decision-maker doesn't rely only on what is said (text) but also considers tone of voice, gestures, visuals, and contextual data. Similarly, multimodal agents synthesize signals across modalities to deliver superior outcomes.

For enterprises, this translates to:

- **Smarter Knowledge Workflows** – Analysts can rely on AI agents to parse through thousands of documents, correlate them with data dashboards, and even interpret video or image evidence to provide a consolidated recommendation.
- **Augmented Human Decisions** – In healthcare, for example, IBM Watson Health has demonstrated the potential of analysing both textual medical records and diagnostic imaging to recommend treatment options—providing physicians with confidence-backed insights.
- **Reduced Errors** – Multimodal validation reduces reliance on single data points, significantly lowering the risk of misinterpretation.
- **Customer-Centric Interaction** – AI agents can move beyond rigid scripts to recognize emotions in a call, analyse historical behaviour, and deliver more empathetic, personalized support.
- **Scalable Innovation** – As enterprises expand into areas like augmented reality, connected devices, and hybrid workplaces, multimodal AI agents provide a scalable framework that can continuously integrate new modalities.

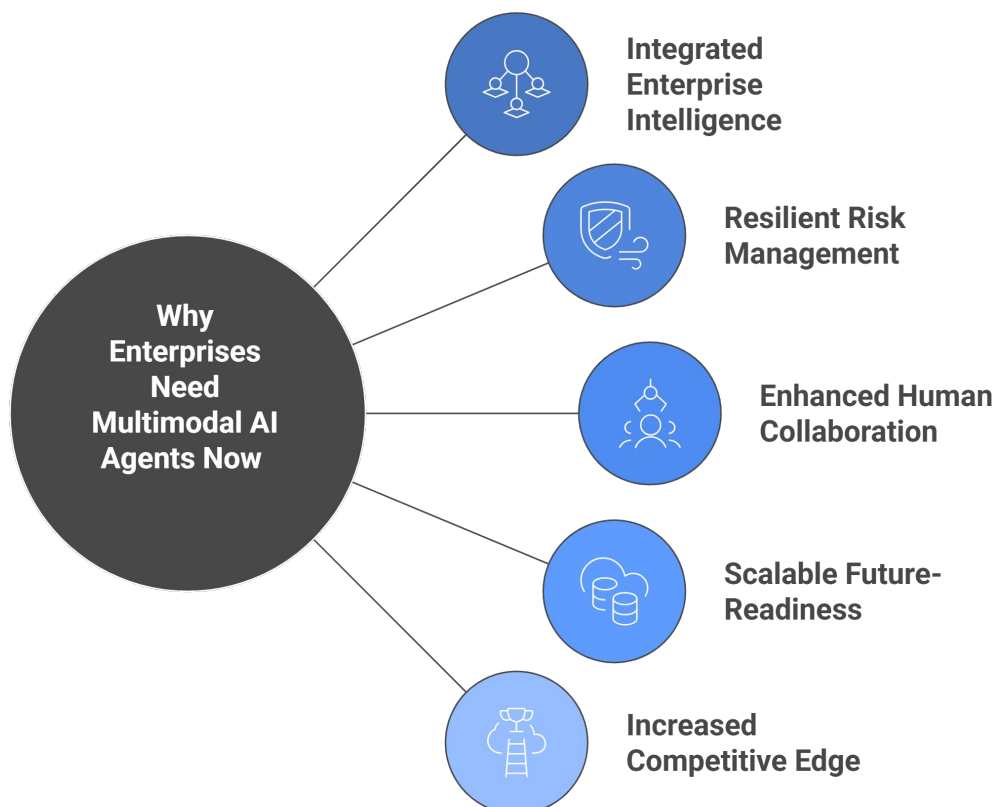
Case Example: IBM Watson Health

IBM Watson Health provides a strong contextual illustration of the power of multimodal AI agents. By combining structured data (patient demographics, lab results) with unstructured data (clinical notes, medical literature) and imaging data (radiology scans), Watson Health was able to support oncologists in identifying treatment pathways. The value here lies not in replacing medical expertise but in augmenting it with data-driven insights that no single modality could achieve in isolation.

This example underscores why multimodal AI agents are transformative: they elevate enterprise functions beyond the sum of their parts.

Benefits in the Practical World

1. **Cross-Functional Intelligence** – Enterprises rarely operate in silos. A multimodal agent can link HR analytics (employee sentiment), operational performance (IoT dashboards), and financial forecasts into a unified strategic view for decision-makers.
2. **Proactive Risk Anticipation** – Agents that analyse facial recognition data in security footage alongside transactional anomalies can flag fraud attempts in real-time.
3. **Adaptive Customer Experience** – Retail AI assistants that combine online browsing behaviour with in-store video analytics can tailor promotions dynamically.
4. **Accelerated Research and Development** – In pharmaceuticals, multimodal agents can align clinical trial text data with molecular imaging, dramatically shortening discovery cycles.



Strategic Value

Enterprises adopting multimodal AI agents move from reactive operations to proactive intelligence ecosystems. These systems don't just provide answers—they provide insights, foresight, and adaptability.

By combining multimodal data inputs with RAG-powered retrieval, enterprises enable AI assistants to:

- Always remain **up-to-date** with the latest organizational knowledge.
- Deliver insights contextualized to both **historical trends and real-time inputs**.
- Support **mission-critical decision-making** in healthcare, finance, manufacturing, and beyond.

Multimodal AI agents are not an optional enhancement but the next step in enterprise evolution. They represent the shift from information retrieval to knowledge orchestration, where enterprises can fully leverage the diversity of their data for competitive advantage.

The Power of Real-World Application



Multimodal AI ceases to be an academic exercise the moment it touches real-world enterprise workflows. Applications across healthcare, finance, legal, design, and customer service are already showing how these systems can bridge gaps between modalities to provide capabilities that were previously inconceivable. For instance, pairing natural language understanding with computer vision allows for diagnostics in healthcare that go far beyond text-based analysis. Similarly, when enterprise knowledge discovery systems combine textual research with visual evidence, they accelerate business intelligence exponentially.

The strength of multimodal AI lies in its ability to contextualize. It can examine a financial chart, interpret the spoken concerns of an analyst, cross-reference textual research reports, and surface the most relevant insights—an orchestration of modalities that mirrors human reasoning. This is why multimodal AI is rapidly becoming central to enterprise AI adoption.

Let us now look at some of the leading **multimodal AI systems that are helping businesses gain edge across various fronts today:**

GPT-4

OpenAI's GPT-4 was among the first widely deployed multimodal large language models (LLMs). Unlike its predecessor GPT-3, which was text-only, GPT-4 can process both text and images. This capability allows it to, for example, analyse a graph or a photograph and respond with contextually appropriate answers. Enterprises have adopted GPT-4 to automate tasks such as technical support, compliance review, and content analysis where both visual and textual cues matter.

From a technical standpoint, GPT-4's architecture integrates computer vision into its transformer-based model, enabling joint reasoning over language and images. While it does not yet natively process audio or video, integrations with external pipelines allow enterprises to extend its capabilities further. For organizations building AI assistants, GPT-4 represents a major step in making assistants more human-like by reducing dependence on text alone.

GPT-5

GPT-5 builds on this trajectory with even broader multimodal functionality. In addition to text and image comprehension, GPT-5 is designed to integrate audio inputs more natively, extending its usability for voice-driven applications such as enterprise call centers and digital assistants. With improved context handling and longer input windows, GPT-5 significantly reduces hallucination risks while enabling enterprises to query across larger, heterogeneous datasets. ...

For knowledge-intensive organizations, GPT-5's expanded multimodal reasoning allows assistants to respond not only to questions about documents but also about charts, audio transcripts, and visual material in tandem. This evolution reflects the enterprise need for assistants that mirror the diverse modalities of the data they are fed. GPT-5 is, therefore, not just an iteration but a step toward enterprise-ready multimodal orchestration.

Watson Discovery

IBM's Watson Discovery focuses specifically on enterprise search and knowledge discovery, with the ability to process unstructured text, documents, tables, and metadata. Unlike general-purpose LLMs, Watson Discovery is tuned for enterprises that need to retrieve precise, context-aware insights from vast repositories of data.

While Watson Discovery is not a fully multimodal model in the same sense as GPT-4 or GPT-5, it can be integrated into multimodal pipelines to power assistants that need to cross-reference structured and unstructured data. Its strength lies in enterprise-grade retrieval augmented generation (RAG)—providing factual, relevant information rather than speculative answers. In regulated industries like finance, legal, and healthcare, Watson Discovery demonstrates how enterprise adoption of multimodal AI often begins with retrieval and structured knowledge augmentation.

Flamingo

Developed by DeepMind, Flamingo is a multimodal model designed for vision-language tasks. It can process sequences of images, video frames, and text inputs together, enabling dynamic applications such as video question answering and multimodal reasoning. Unlike static image models, Flamingo is optimized for continuous streams of information, which makes it particularly relevant for enterprises monitoring live video feeds, security footage, or complex training simulations.

For instance, a manufacturing company could deploy Flamingo in monitoring assembly lines, where the AI simultaneously analyses sensor data (numerical) and video streams (visual) to identify potential defects. In research environments, Flamingo demonstrates the power of aligning visual and textual reasoning at scale, which enterprises can adapt into multimodal AI agents capable of both observation and explanation.

ImageBind

Meta's ImageBind is an ambitious research model designed to bind together six different modalities: images, text, audio, depth data, thermal imaging, and motion sensors. Unlike models trained on pairwise modalities (e.g., image + text), ImageBind is designed to learn joint embeddings across all six modalities without requiring explicit pairwise training for each combination.

For enterprises, this is revolutionary. Imagine a logistics assistant that correlates IoT sensor data (motion, thermal) with shipment records (text) and camera feeds (images). ImageBind's capability allows AI to create a shared semantic space across modalities, making data fusion seamless. While still primarily a research model, ImageBind illustrates how enterprise multimodal AI is evolving toward systems that can integrate any form of data into a single reasoning framework.

Claude

Anthropic's Claude is primarily a text-based conversational AI designed around constitutional AI principles for safer, aligned outputs. While not natively multimodal in the same way as GPT-4 or Flamingo, Claude has been integrated into pipelines that extend its reasoning with multimodal inputs, particularly in retrieval-augmented contexts. Its emphasis on alignment, reliability, and reduced hallucination makes it a strong candidate for enterprise deployments where safety and factual accuracy are paramount.

Enterprises often pair Claude with multimodal preprocessing pipelines, such as transcribing audio into text or converting images into descriptions, before feeding the information into Claude. This approach ensures that enterprises gain the benefit of Claude's safer, contextually aligned outputs without sacrificing multimodal capabilities. In practice, Claude becomes the reasoning engine in multimodal systems, particularly in regulated domains.

Firefly

Adobe's Firefly represents the creative frontier of multimodal AI. Designed for generative design, Firefly can take text prompts and transform them into high-quality images, design variations, and creative assets. For enterprises, this is transformative in domains such as marketing, product design, and advertising, where time-to-market is critical.

Firefly's integration with Adobe's ecosystem allows organizations to build workflows where text, brand assets, and visual data seamlessly combine into final creative outputs. Importantly, Firefly has been designed with commercial safety in mind, trained on licensed and open datasets—addressing enterprise concerns about intellectual property risk in generative AI.

Apache Airflow

Apache Airflow, while not a multimodal AI model itself, plays a critical role in orchestrating multimodal AI pipelines. As an open-source workflow management platform, it allows enterprises to coordinate data ingestion from multiple modalities—audio, video, structured datasets—into unified AI processing pipelines.

For example, an enterprise could use Airflow to orchestrate a workflow where call centre recordings (audio) are transcribed, customer support tickets (text) are processed, and performance dashboards (structured data) are analysed, before feeding everything into a multimodal AI assistant. Airflow thus enables scalable and reliable operationalization of multimodal AI systems at the enterprise level.

Crew AI

Crew AI is an emerging framework designed to coordinate multiple AI agents working collaboratively. In a multimodal context, this means one agent could specialize in image recognition, another in natural language reasoning, and another in structured data analysis, with Crew AI orchestrating their interactions.

For enterprises, this agentic approach mirrors how cross-functional teams work: specialized roles, coordinated toward a shared outcome. Crew AI offers enterprises the ability to deploy modular multimodal AI assistants that can be extended or adapted as new modalities and tasks emerge.

MS AutoGen

Microsoft's AutoGen framework is designed to enable multi-agent conversations and workflows. It allows developers to create systems where multiple AI agents collaborate to solve complex tasks—essential for multimodal AI pipelines that demand coordination across inputs and reasoning layers.

For instance, one agent might retrieve knowledge from documents, another might analyze charts, while a third evaluates sentiment from customer calls. AutoGen provides the framework for these agents to collaborate seamlessly, ensuring enterprises can build sophisticated multimodal AI assistants that are modular, adaptable, and transparent.

Factors to Consider Before Opting for Multimodal AI Agents

While the promise of multimodal AI agents is transformative, enterprises cannot adopt them without careful consideration of risks, operational dependencies, and compliance requirements. Unlike text-only AI systems, multimodal AI agents deal with multiple data types simultaneously—increasing complexity, surface area for vulnerabilities, and the stakes of deployment. Adoption, therefore, requires a strategic approach grounded in security, governance, scalability, and ethical design.

Key Factors to Consider

1. Data Quality and Availability



The performance of any AI system depends on the quality of the data it processes. Multimodal AI requires aligned, high-quality datasets across modalities—text, audio, video, and structured data. Missing or biased modalities can skew insights, leading to flawed decision-making. Enterprises must invest in robust data preparation, normalization, and governance processes before deployment.

2. Computational Infrastructure



Multimodal AI models are significantly more resource-intensive than single-modal models. Training and running these systems requires specialized hardware (GPUs, TPUs), distributed compute clusters, and high-throughput storage systems. Without the right infrastructure, organizations face delays, scalability issues, and spiraling costs. Cloud-based solutions can offset this, but they introduce their own considerations around dependency and vendor lock-in.

3. Security and Privacy Risks



Multimodal AI increases the exposure surface for sensitive data. For instance, analyzing facial expressions involves biometric information, while call-center audio contains personally identifiable information (PII). Enterprises must enforce end-to-end encryption, access controls, anonymization protocols, and strict retention policies. In sectors such as healthcare or finance, compliance with HIPAA, GDPR, or other regional data privacy frameworks becomes non-negotiable.

4. Threat Vectors



The addition of multiple modalities creates new attack surfaces:

- Adversarial inputs (e.g., manipulated images fooling vision models).
- Prompt injection and jailbreaks in language models.
- Data poisoning in multimodal training sets.

Enterprises must deploy robust adversarial testing, red-teaming, and monitoring systems to defend against these evolving threats.

5. Interpretability and Auditability



Decision-making in critical industries demands explainability. When multimodal AI agents synthesize across images, text, and audio, the reasoning path can become opaque. Without interpretable outputs, enterprises risk compliance violations and stakeholder mistrust. Deploying explainable AI (XAI) frameworks, coupled with audit trails for multimodal pipelines, is essential.

6. Cost of Ownership



Beyond infrastructure, enterprises must evaluate licensing costs, API dependencies, retraining needs, and compliance investments. Multimodal AI is not a plug-and-play solution; ongoing costs include maintenance, monitoring, and retraining to accommodate new modalities or data.

7. Workforce and Change Management

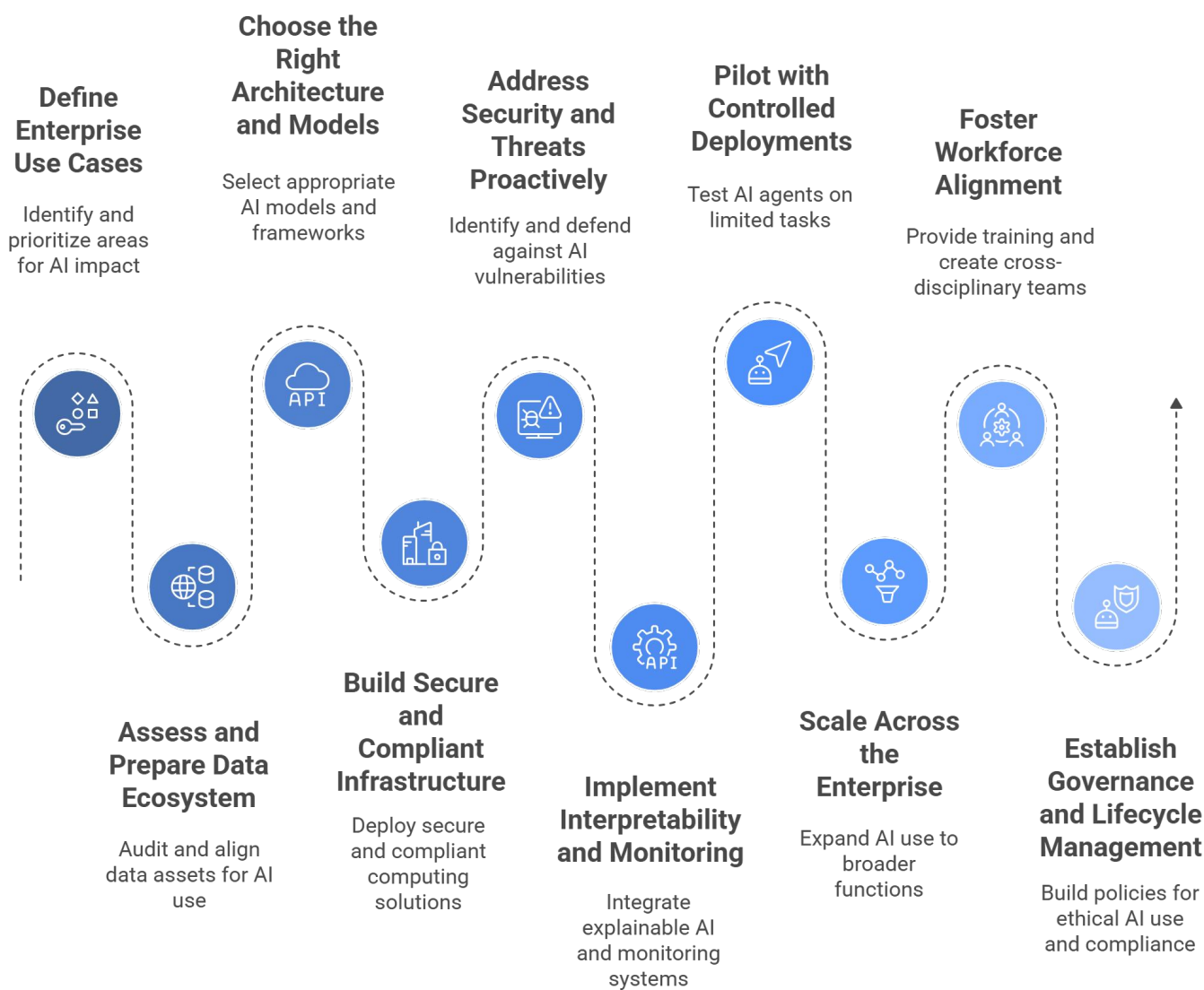


Introducing multimodal AI requires upskilling teams in data science, cybersecurity, compliance, and model management. Resistance to change or lack of understanding can derail adoption. Enterprises must implement training programs and foster a culture of collaboration between AI experts, domain specialists, and compliance officers.

Roadmap To Building Multimodal AI Agents

To move from concept to deployment, enterprises require a structured roadmap that balances innovation with operational discipline. A roadmap ensures that multimodal AI agents are built not as isolated experiments but as enterprise-ready systems aligned with business goals, regulatory requirements, and security standards.

Roadmap to Building Multimodal AI Agents



Multimodal AI will define the next decade of enterprise intelligence. But without a disciplined roadmap grounded in security, compliance, and scalability, enterprises risk turning promise into peril. The path to adoption is clear: align business strategy with data ecosystems, enforce security, build responsibly, and scale with governance. Only then can multimodal AI agents become game-changers rather than experiments.



Custom, Secure And Scalable Digital Solutions For Your Business

Perimattic builds digital solutions that change how your business works from automation to cybersecurity and everything in between.

[Contact Us Now](#)



US +1 (442) 319-7124

UK +44 (20) 3769 0051

IND +91 92142 66896



sales@perimattic.com



www.perimattic.com

Follow Us

