# **Google Cloud:** Integrating GenAI with Infrastructure

## A leader's guide to Generative AI decision making, accelerated value realization, and risk management

Google Cloud thoughtworks

# Contents

# Executive Summary

Generative AI (GenAI) is a technology experiencing rapid change in an ever-evolving business landscape that can also lead to revolutionary impact. The regular release of next stage AI products, ongoing identification of new use cases, and rapidly growing number of supporting solutions challenge even the most tech-savvy navigating the GenAI landscape.

This guide provides a path for decision makers and technologists to streamline their evaluation of Google Cloud's GenAI with Infrastructure options, align business challenges to the most fit for purpose solutions, and accelerate realization of value without sacrificing legitimate risk management and other concerns. As you consider adopting GenAI, it's important to understand its base concepts, how to get organizational buy-in, how to evaluate which model and application to use, how to overcome known adoption challenges, and how to start building a strategy for execution.

Aligning your adoption and use of AI applications to an optimized platform engineering model is key to success. Coupling Google Cloud's GenAI offerings, such as Vertex AI, with existing or new infrastructure platforms based in Google Kubernetes Engine (GKE) or Cloud Run provides an accelerated path to value realization, cost management, and differentiated business impact.

To achieve these outcomes, executives and developers must delineate true business opportunities from hype, nurture innovation while establishing governance, manage risk without stifling experimentation, and implement business models to support rapid realization of business value.

When building GenAI applications, innovation happens at the intersection of where the model lives and where the app runs. Until recently, application workloads and generative AI workloads have largely remained separate, but this is changing. Businesses have begun integrating these two in new, innovative ways, and Google has built an ecosystem to support these efforts. This is how GKE, Vertex AI, and Cloud Run can position you for success.

As with all revolutionary technologies, companies integrating GenAI and related infrastructure platforms

- Migrating a proof of concept to production
- Insufficient or low quality data
- Lack of expertise and talent
- Maintaining observability of GenAI production systems
- Understanding the need for GenAI

The technical challenges are solvable using Google Cloud's GenAI tooling and cloud infrastructure. Google Cloud provides everything you need to get started — from models like Gemini to cloud infrastructure for training and operating models — to then realize business value faster. Your company can also save cost and benefit from better performance, helping you rapidly test ideas without hurting your budget.

With new GenAI technologies released every day, figuring out which tool to choose becomes an easier decision when Google provides several that integrate seamlessly together. Google Cloud is also the "only major cloud provider offering both first-party AI models and third-party AI models on equal footing, which is an incredible differentiator" according to Thomas Kurian, Google Cloud CEO.

As you consider integrating GenAI into your business, let's first discuss how you achieve organizational buy-in.

# Achieving Organizational Buy-in

GenAI is a class of artificial intelligence with the ability to generate new content after learning from prior observations. While traditional AI gives predefined outputs based on specific inputs, GenAI's learned behavior and natural language understanding supports creation of rich text, images, voice and video — and can aid decision making.

Since the explosion of GenAI in 2023, everyone has heard about the possibilities of incorporating GenAI into your business — but not everyone is convinced of its transformative potential. As you consider how to incorporate GenAI into your business, one main challenge is persuading others to incorporate it into the overall strategy.

## Identifying Business Value

Research shows that GenAI does in fact deliver more than just hype — offering strategic value across your entire business. According to McKinsey's research in 2023, four business areas made up 75% of GenAI use cases: customer operations, marketing and sales, R&D, and software engineering. These areas are a great starting point for figuring out where GenAI fits into your business.

**Integrating the transformative technology of GenAI into these prioritized areas of your business can transform your business through:**

**Competitive edge**

AI enables businesses to anticipate and adapt to market shifts by analyzing data and trends. For example, it can predict customer preferences and personalize products, helping you stay on top of changes.

**Cost reduction**

AI can streamline logistics, optimize resource allocation, and predict equipment failures, leading to significant potential cost savings.

**Innovation and agility**

GenAI tools can help generate new ideas and make those ideas a reality faster, fostering a culture of continuous improvement and rapid iteration.

**Productivity**

Repetitive tasks like code generation and testing can be automated, freeing up time for strategic problem-solving and innovation.

# Evolving Business Capabilities

As an emerging technology, GenAI's value proposition continues to evolve — with use cases across industries.

### Adaptation and learning

Automatically adapt content generation based on changing inputs like customer behavior, inventory, or market conditions. The model's learned behavior can also be adapted through a regular process of tuning. While this ability to adapt can be limited by the model's design and the complexity of the environment, this adaptability is now feasible at large scale.

### Automation

The adaptability and scalability of GenAI enable automated tasks and processes by processing and analyzing data, generating content, interacting with humans, and streamlining administrative tasks. While challenging for varied inputs, many routine tasks can be automated effectively to minimize manual effort.

### Enhanced communication

Transform the way your business connects with customers by using cutting-edge language processing. This helps tailor interactions, clarify messages, predict needs, provide instant support through chatbots, and connect better with diverse audiences on various platforms.

### Knowledge discovery

Sift through vast datasets to uncover hidden insights, identify correlations and facilitate knowledge discovery efforts. While limited in new scenarios, it can optimize algorithms for specific domains.

### Optimization

Analyze large datasets to refine systems and processes, making resource allocation more efficient, cutting waste, and boosting operational effectiveness with personalized strategies and predictive analytics.

### Creativity augmentation

Enhance creativity with GenAI by uncovering overlooked ideas and improving people's own creative processes. While it draws from existing knowledge and adaptability, AI works alongside human creativity rather than replacing it.

### Decision support

Produce insights from data to guide informed decision-making. Even with data vulnerabilities, thorough preprocessing and validation guarantee dependable decision support.

### Personalization

Analyze preferences, generating personalized content for enhanced engagement. While accuracy may be limited, iterative learning refines models over time to improve effectiveness.

### Prediction and forecasting

Interpret and generate natural language to analyze datasets for accurate predictions. While hindered in new scenarios, careful preprocessing and validation mitigate this, guaranteeing reliable predictions.

### Risk management

Enhance risk mitigation by analyzing patterns and anomalies in data to predict potential issues, enabling proactive measures and improving decision-making processes across various operational and strategic contexts. Robust preprocessing and validation techniques ensure reliable risk management outcomes.

## Building an Adoption Strategy

A business can only benefit from the capabilities described above if GenAI is implemented in an effective way, starting with developing a GenAI strategy. To build a strategy, you must consider how to communicate the value to the rest of your organization, navigate risk concerns, focus on the use cases that matter, and determine the right tools for your business.

**To effectively integrate GenAI into your organization, Thoughtworks recommends five crucial stages:**

1. **Build a case for change**
   Craft a compelling narrative that highlights GenAI's business value to stakeholders.
2. **Explore use cases**
   Identify and prioritize GenAI use cases based on analytics and potential impact.
3. **Illustrate the art of the possible**
   Demonstrate value to stakeholders with a proof of concept built with a tool like Vertex AI to show new features in action.
4. **Plan for execution**
   Formulate a strategy for effective GenAI implementation by considering governance, costs, and roadmaps.
5. **Manage your AI portfolio**
   Track and refine a portfolio of initiatives for continuous improvement as your business learns from each project.

Shaping a compelling strategy will set you on the fast track for adopting GenAI and adapting your business. Thoughtworks can help you craft a comprehensive GenAI strategy for transformative growth, starting with stakeholder insights and GenAI concepts, exploring cross-domain use cases, and compiling an AI experiment backlog.

# Defining the Landscape

To effectively utilize GenAI, organizations must have realistic expectations about what's possible — including its capabilities and limitations.

## Capabilities

Recent advances in neural network architectures, access to specialized hardware, the evolution of domain-specific software, and increased data availability have redefined what's possible with generative technologies.

### Core capabilities of GenAI:

**Contextual fluency**
Effectively use contextually-relevant information.

**Content generation**
Produce rich text, images and video.

**Extensive knowledge and comprehension**
Create encoded information on a larger scale.

**Natural language**
Directly interpret and generate natural language.

**Scalability**
Run larger workloads with consistent performance.

## Limitations

Despite the hype, GenAI is not a panacea and has inherent limitations that businesses should know so they can build responsible products and avoid risks:

**Data reliance**
Exposes data bias and quality issues.

**Derivative output**
Limits creativity due to learned patterns.

**Inability to generalize**
Leads to poor output in new scenarios.

**Resource intensive**
Demands costly hardware for advanced models.

**Shallow comprehension**
Results in nonsensical output.

Because these limitations are inherent, it's important to choose tools that can mitigate them as much as possible — this is one place where Google Cloud's offerings are differentiated.

# Google's Differentiated AI Offerings

Google has established itself as a leader in AI infrastructure according to the 2024 Forrester Wave: AI Infrastructure Solutions report. Access to massive amounts of specialized hardware, broad software support, flexible consumption models, and wide selection of AI models set them apart. Beyond being a great reference, Google is a great partner to your business for several reasons.

## Hardware

Google stands out as a leader in AI thanks to its cutting-edge hardware solutions designed for optimal performance and efficiency. This includes a range of computing resources such as GPUs, TPUs, and CPUs, as well as advanced storage and networking capabilities like block, file, and object storage, along with high-speed networking technologies like OCS and Jupiter.

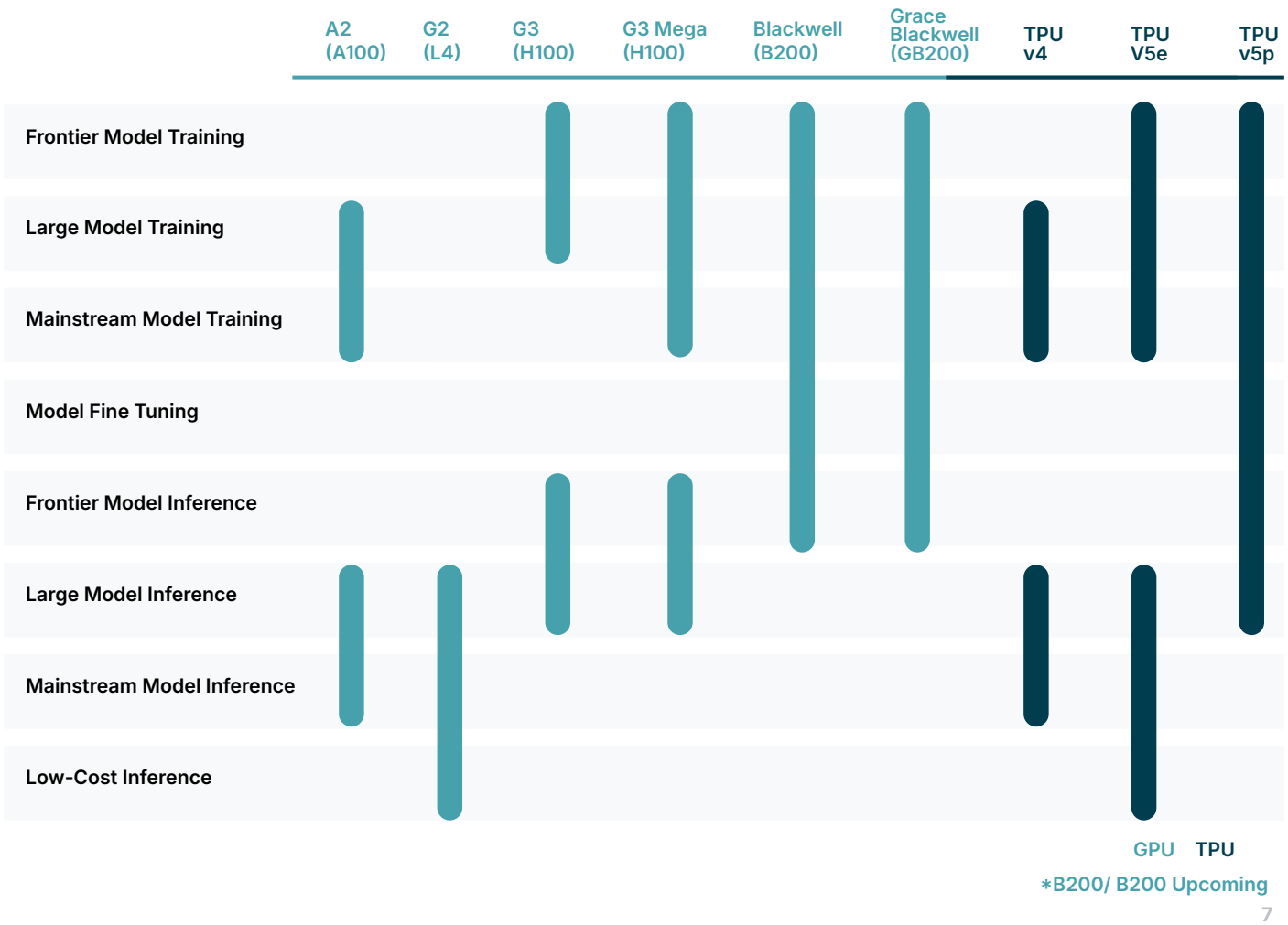
One of Google's key strengths is its pioneering role in launching hardware tailored specifically for machine learning (ML) tasks and AI applications. For instance, Google was the first to introduce L4, A100, P4, and T4 accelerators for accelerated ML inference training and GenAI development.

Noteworthy offerings include the G2 VM equipped with L4 GPU to speed up GenAI inference, A3 Mega VMs featuring H100 GPUs for large-scale AI training, and the availability of DGX Cloud on A3 VMs.

Google continues to push the boundaries with innovations like the new cloud TPU v5p, which delivers 2.8x faster LLM training and 1.9x faster embeddings training compared to its predecessor, TPU v4. Additionally, Google Cloud will soon welcome the next-generation NVIDIA Blackwell GPUs, slated for early 2025.

**The matrix below provides a comprehensive overview of Google's purpose-built AI acceleration offerings:**

## Comprehensive portfolio of purpose-built AI acceleration



## Software

Google's tools support open-source software such as libraries like JetStream, MaxText, and MaxDiffusion and frameworks like JAX, TensorFlow, PyTorch, and XLA. Together with GKE and Compute Engine this software ensures high flexibility, usability, and optionality.

Google is a leading contributor of OSS development tools such as JAX and OpenXLA. Google offers the GKE Enterprise that improves team productivity by 45%. This service is built on the Jupiter network, which uses 40% less energy, and is adaptive and flexible.

## Consumption

Google supports a flexible consumption model. For example, you can use Dynamic Workload Scheduler (DWS), On Demand, CUD, or Spot instances to intelligently optimize your resource consumption.

DWS specifically is designed to match GPU availability to the demand for AI workloads.

It works across all Google Cloud services such as managed instance groups on Google CloudE, Batch for Google CloudE, GKE, and Vertex AI. It supports two distinct modes:

It supports two distinct modes:

**Calendar mode**: Assures job start times with future reservations. Use cases: (re)training, recurring fine-tuning.

**Flex start mode**: Provides optimized economics and obtainability for on-demand resources. Use cases: time flexible experiments, fine tuning, batch inference.

## Models

In order to create a GenAI application, the first step is selecting a model to determine the core functionality of your app. Fortunately, Google Cloud offers several cutting-edge models and developer platforms like Vertex AI, which includes services such as Model Garden, ML Ops, AI Solutions, and AI Tools.

**The table below represents some native large language models (LLMs) provided by Google with a short description of their supported use cases.**

| Name | Description | Use Cases |
|---|---|---|
| **Gemini** | Multi-modal LLM | Wide variety of tasks from natural text, image, audio, and video understanding to mathematical reasoning. |
| **PaLM 2** | LLM with multi-language capabilities | Understanding, generating, and translating nuanced multilingual text. Applying logic, common sense reasoning, and mathematics. Coding. |
| **Imagen 2** | Foundational model for AI generated Images | Generation of high-quality images based on input text. |
| **MedLM** | Foundation model fine-tuned for Healthcare industry | Generation of medical documentation, improving pre-clinical research and development, and improving healthcare experiences and outcomes. |
| **Codey** | Foundation model specialized for code | Answering code-related questions, debugging, creating documentation from code, and learning code. |

In addition to first-party models, the Vertex AI Model Garden provides access to 130+ open-source and third-party foundation models such as Gemma and CodeGemma, Meta's Llama 2, TII's Falcon, Bert, T-5 FLAN, ViT, EfficentNet, and Anthropic's Claude 3. The number of models will only continue to grow.

## Hugging Face collaboration

While using Google Cloud, you also have access to Hugging Face, the world's largest community for AI development that includes over 500,000 models to experiment with and utilize. This strategic partnership with Google Cloud allows you to browse, deploy, and run Hugging Face models with Vertex AI and GKE. The training and serving of their models is also now optimized for both GPUs and TPUs.

## Open-source vs. commercial LLMs

To decide between open source and commercial LLMs and stay aligned with organizational needs and goals, you must consider several key factors:

**Cost and scalability**

Use open source LLMs when starting small or experimenting with AI capabilities on a budget. Commercial LLMs often offer better scalability and managed services for handling large volumes of queries or enterprise-level deployment.

**Customization and control**

Choose open source LLMs if you need deep customization or integration into proprietary systems without vendor lock-in. Commercial LLMs typically provide less flexibility but include robust support and maintenance.

**Compliance and security**

Opt for commercial LLMs when dealing with sensitive data or requiring strict compliance with data protection regulations, as these models often provide enhanced security features and compliance certifications not always guaranteed by open source solutions.

# Implementing GenAI with Google

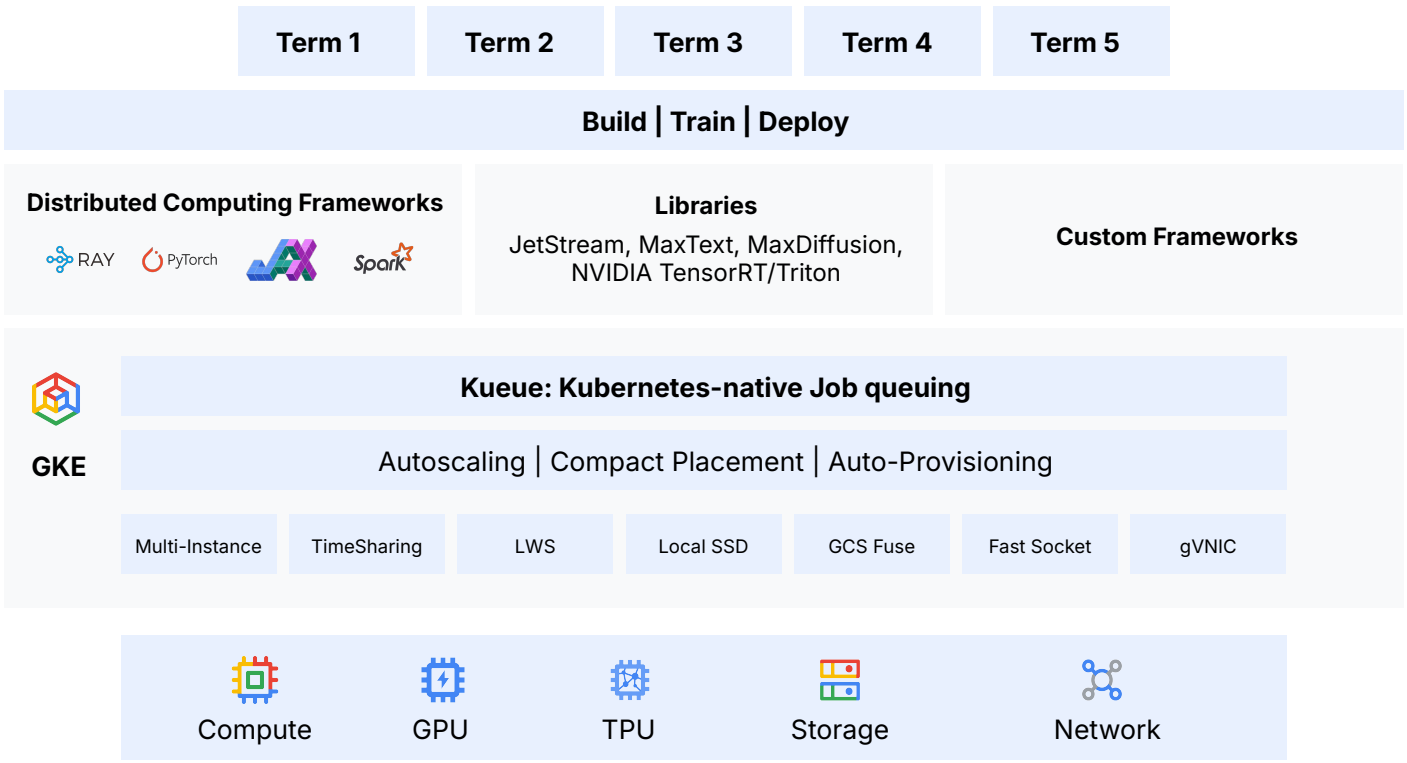
Once you decide to utilize GenAI in your business, you need to consider what tools to use. Google Cloud provides multiple options for building, training, and deploying GenAI applications.

## Accelerate Your AI Journey with GKE

Developers and data scientists are increasingly turning to GKE to optimize and run demanding AI/ML workloads. With GKE, you can implement a robust, production-ready AI/ML platform with all the benefits of managed Kubernetes, such as:

- Infrastructure orchestration that supports GPUs and TPUs for training and serving workloads at scale.
- Increased workload performance with multi-instance GPUs, time sharing, local SSD, GCS Fuse, Fast Socket, and gVNIC.
- Built-in support for auto-scaling, compact placement, and auto-provisioning.
- Kubernetes-native job queueing with Kueue.
- Flexible integration with distributed computing and data processing frameworks.
- Multi-team support on the same infrastructure to maximize utilization of resources.

**Below is a graphical representation of the AI/ML platform built on top of GKE.**



# Portability & Customizability

GKE supports the use of open source software, enabling you to build applications with tools and frameworks you are most comfortable with, while taking advantage of the price-performance benefits of the AI Hypercomputer architecture.

To enable platform customization, GKE also supports a variety of portable frameworks and ecosystem tools. There are several open source tools to help build your applications:

**JAX:** A compiler-oriented and modular framework tightly integrated with the XLA compiler. It is built for performance, scales easily, and supports fast iteration and innovation, which is critical for the rapidly evolving GenAI field. JAX only requires Python knowledge to perform well with very little code, leading to greater developer productivity and accessibility.

**JetStream**: A cost-efficient LLM serving engine for JAX and PyTorch/XLA delivering high performance. It has high throughput and memory optimizations designed for large scale models.

**MaxDiffusion**: Performant and efficient text-to-image tool for model training and serving. As a reference implementation for latent diffusion models on JAX, it's optimized for large scale inference with multiple TPU pods. It supports Hugging Face JAX models.

**PyTorch/XLA**: A Python package that enables performance and portability across XLA devices. PyTorch's flexibility and dynamic nature make it a popular choice for deep learning researchers and practitioners. Developed by Google, XLA is a specialized compiler designed to optimize linear algebra computations — the foundation of deep learning models. PyTorch/XLA offers the best of both worlds: the user experience and ecosystem advantages of PyTorch alongside the compiler performance of XLA to accelerate PyTorch code for Cloud TPUs/GPUs and use JAX optimizations.

# Performance & Scalability

Orchestrating large-scale AI workloads with TPUs and GPUs has historically required manual effort to handle failures, logging, monitoring, and other foundational operations. GKE is the most scalable and fully managed Kubernetes service, considerably simplifying the work required to operate TPUs and GPUs. Leveraging GKE to manage supercomputer-scale training and inference workloads improves AI development productivity.

# Cost-Efficiency

When building a production-ready, highly scalable, and fault-tolerant managed application, GKE brings additional value by reducing your total cost of ownership for inference on TPUs.
**With GKE, you can:**

- Reduce platform efforts by using a Kubernetes standard platform.
- Minimize cost with autoscaling.
- Auto-provision the necessary compute resources.
- Ensure high availability with built-in health monitoring for TPU VM node pools.
- Minimize disruption from updates and hardware failures with proactive handling of maintenance events and gracefully terminating workloads.
- Gain full visibility into your TPU applications with GKE's mature and reliable metrics and logging capabilities.

GKE helps to increase utilization of valuable resources while reducing operational overhead through several innovations:

Fast workload startup by preloading the container image and/or model weights on a **secondary boot disk**. This supports near-constant latency even at massive scale. For example, this technique results in up to 29x faster time to mount a 16GB container into a Running status.

Faster data loading using **Cloud Storage FUSE with caching**. By caching data onto local SSD from Cloud Storage, ML jobs can fetch data much faster. Portability across clouds/on-premises is an additional benefit because no code changes are needed.

Faster data loading for training with **ParallelStore**. This service improves the read throughput performance up to 6.3x compared to Lustre Scratch offerings achieving ~200MB/sec per TB. It is optimized for demanding AI/ML workloads by providing key capabilities such as distributed metadata management, extreme IOPS, and key-value architecture.

Efficient resource sharing using **Kueue**. Open Source multi-tenant Kueue service allows multiple teams to share scarce GPUs and TPUs and enable fair sharing, bursting, job-level priority, and preemption. It provides diverse ecosystem support, including k8s Job, Kubeflow MPIJob, RayJob, and more.

# Choosing the Right Tools

**There are two primary components of a GenAI system:** the application workload and the model inference workload.

## Application workload

The application workload handles requests from users, applies business logic, and likely connects to various data sources.

The application runtime environment you choose depends on what you're building. **The most common considerations for choosing this environment are:**

**Workload characteristics**: Application type (web, microservice, etc.), statefulness, resource needs (CPU, memory, network), and scaling patterns (predictable vs. spiky).

**Operational factors**: Level of control desired (Kubernetes vs. serverless), team's DevOps experience, portability needs, and essential monitoring/observability tools.

**Cost**: Pricing models (pay-per-use vs. provisioned), overall cost structure (fixed vs. variable), and opportunities for cost optimization.

**Ecosystem**: Compatibility with existing tools (CI/CD, monitoring), language support, and any necessary integration with specific cloud services.

**Security and compliance**: Encryption, access controls, and alignment with industry-specific compliance standards (e.g., HIPAA, PCI DSS).

## Model inference workload

This new kind of workload focuses on running an AI model for generating results (inference). Runtime decisions revolve around the model's unique needs and complexities, requiring careful optimization. Performance and specialized hardware are crucial due to the workload's computational demands. Additionally, seamless MLOps integration is vital for system maintenance in production.

**Considerations when choosing an inference runtime environment include:**

### Model-specific factors

Model complexity, framework, data types, and hardware needs determine platform compatibility and resource requirements.

### Performance & scalability

Ensure the platform delivers the required latency and throughput, and can scale to handle prediction demand.

### Operational considerations

Consider deployment ease, how it fits into your workflow, and MLOps tools for monitoring/retraining.

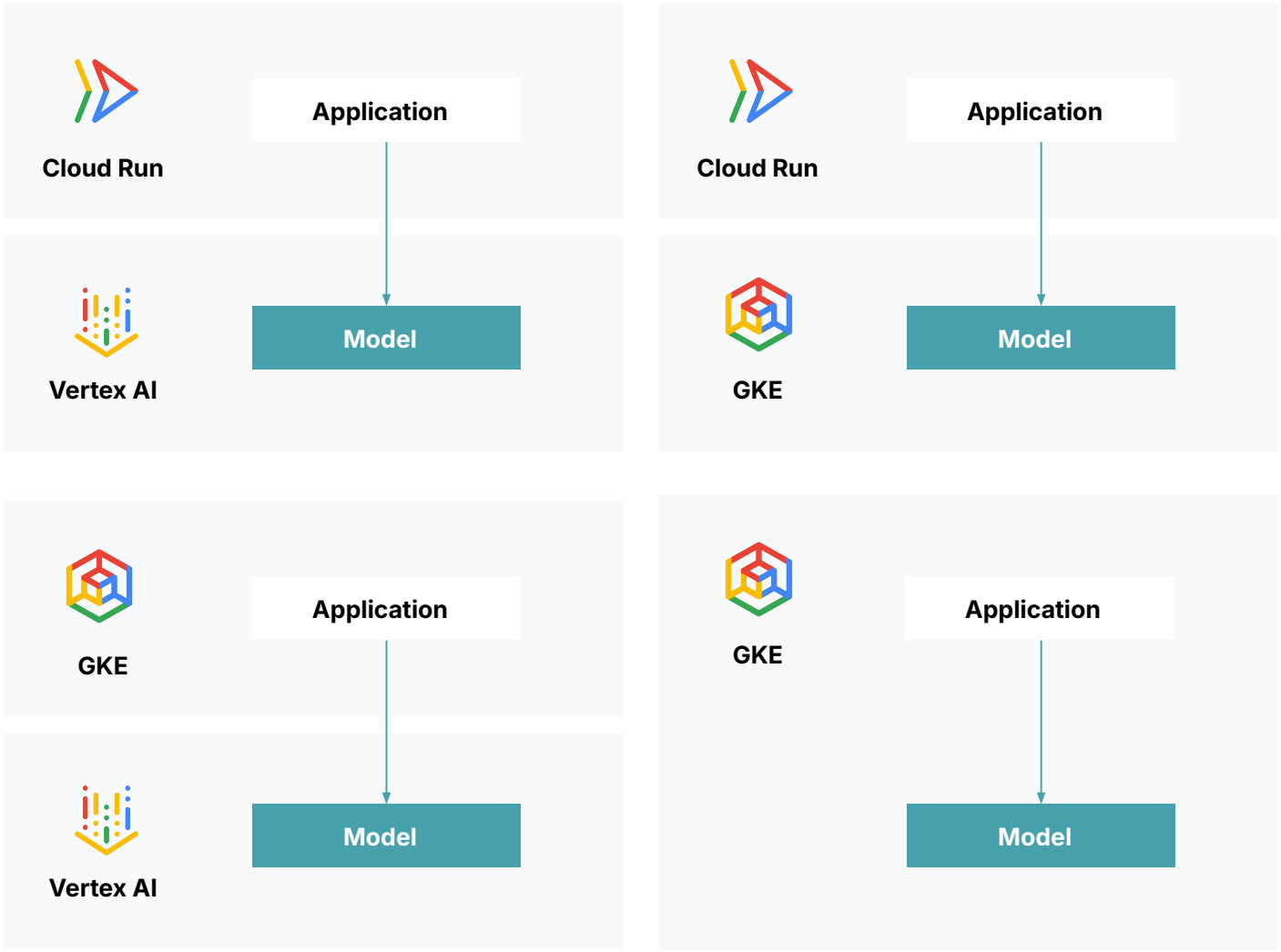
### Cost optimization

Balance the cost of hardware (especially GPUs/TPUs) vs. benefits and the tradeoffs of managed vs. self-managed solutions.

# Google Cloud's Offerings

With Google Cloud, you're free to mix-and-match your runtime environments as you see fit. Start simple with the application in Cloud Run and the model in Vertex AI. Move towards production by shifting the application workload into your existing GKE cluster. Need the most flexibility? Have an existing system? Build in GKE to leverage the best accelerators, great support for popular AI libraries and frameworks, resource allocation control as you see fit, and excellent integration with other cloud services.

**These are four different runtime combinations for running your GenAI application utilizing Google Cloud:**



## Cloud Run for your Application Workloads

Cloud Run is a fully-managed platform for serverless workloads based on containers. It can process requests from users, batch operations, or perform event-based processing.

Cloud Run is easy-to-use, code-centric, pay-as-you-go, and allows fast iteration. Leverage it for stateless web applications and APIs, event-driven workloads, and highly-variable traffic patterns.

## Vertex AI for your Model Workloads

Vertex AI is a managed platform to train and deploy models and applications or customize LLMs for use in your AI-powered applications. Begin using Google's premier LLM, Gemini, with just a few lines of code; choose from a variety of premier, open source, or specialized models in Model Garden; tune models in Vertex AI Studio; build search engines or AI agents in Vertex AI Agent Builder; provide data scientists with the scientific notebooks they need; and keep models up-to-date with MLOps.

## GKE for your Model and Application Workloads

GKE is a managed Kubernetes service, offering top-notch TPU and GPU support, cost optimization, and seamless integration with open-source tools. You can deploy both application and model workloads, making GKE ideal for demanding AI tasks, experienced Kubernetes users, or diverse environments. Use GKE Autopilot to attach TPU or GPU accelerators to your workloads, create dedicated node pools of TPUs and GPUs to maximize utilization, and take advantage of Google's Multislice for near-linear scaling up to tens of thousands of TPUs.

## Observability for GenAI Applications

Observability for GenAI applications functions just the same as observability for standard applications. Using Google Cloud Logging, real-time metrics can be recorded, monitoring interactions with models and creating the data pipelines that help validate your models' quality. Ensuring that responses to prompts maintain the context of the conversation is necessary to reduce the risk of presenting false or misleading information. Working in conjunction with Google Cloud Monitoring, alerts can be raised when responses from your models begin to deviate from the prime context and present a less accurate interaction with the customer.

In the same regard, monitoring customer prompts will ensure that you can store interactions and later determine if and when the models need to improve. This filters out imperfect conversations and focuses only on edge cases that improve the context, not deter from it.

# Balancing Business Considerations

As you integrate GenAI into your business, there are several things for you to consider, such as ROI, risk management, building upon your current infrastructure, and being ready for future trends. Let's dig into each.

## Return on Investment

When implementing GenAI technologies, it's crucial to measure the Return on Investment (ROI) to understand value creation and AI pr

**Cost reduction**

GenAI can significantly enhance operational efficiencies and reduce costs by automating repetitive tasks, optimizing workflows, and minimizing manual errors. This results in reductions in labor costs, lower error rates, and decreased requirements for rework or quality control. These cost savings can be quantified by tracking operational expenses before and after the deployment of GenAI technologies, such as Google Cloud's Vertex AI, AutoML, and Gemini Code Assist.

**Speed to market**

GenAI can accelerate research and development, significantly reducing the time required to bring new products and services to market. By leveraging Google Cloud's Vertex AI Studio, Imagen, or Codey, you can speed up your company's R&D processes. Measure the effectiveness of these tools by tracking the shortened product lifecycle timelines and comparing them against industry benchmarks.

**Revenue growth**

GenAI can significantly enhance product offerings, marketing strategies, and customer experiences, thereby increasing revenue streams. To evaluate impact, measure changes in sales volume, conversion rates, and customer acquisition and retention rates. Also make sure to analyze the revenue generated from GenAI-enabled products and services — such as those run on Google Cloud's Dialogflow, Contact Center AI, Translation AI, and Vision AI — to establish a clear link between the use of these technologies and improved financial performance.

**Employee productivity**

GenAI's impact on enhancing employee productivity is seen by the increase in output per employee and the reduction in time spent on low-value tasks. By implementing Google's AI Readiness Program and Cloud Identity, organizations can effectively measure these improvements using productivity metrics and gathering employee feedback to assess changes post-GenAI implementation.

## Risk Management and Compliance

As always, new technology delivers new and significant challenges. Stakeholders must act quickly to address both the opportunities and the risks. Multiple concerns based on risks have already surfaced:

**Transparency**

Transparency in AI refers to the ability to understand and explain the decision-making process of AI systems. It's important because it builds trust between the AI and its users. Furthermore, transparency can help identify and correct biases in AI systems, leading to more fair and equitable outcomes. GenAI relies on neural networks with billions of parameters, challenging our ability to explain how any given answer is produced.

**Fairness**

Models may expose algorithmic bias due to imperfect training data or decisions made by the engineers developing the models. Fair AI systems can potentially help promote equality and reduce discrimination in various sectors such as law enforcement.

**Intellectual property (IP)**

GenAI models can generate content that is similar to or copies existing copyrighted works. This creates significant IP risks, including infringing on copyrighted, trademarked, patented, or otherwise legally protected materials. Organizations must understand what data went into training and how it's used in outputs even when using a cloud provider's GenAI tool. Google has stated that they will assume responsibility for the potential legal risks involved if you are challenged on copyright grounds.

**Privacy**

Since GenAI models are often trained on large datasets, this can pose privacy risks with sensitive or personal information. In addition, privacy concerns can arise if users input information that later shows up in model outputs that are personally identifiable. Google was one of the first in the industry to publish an AI/ML privacy commitment, which states that customers should have the highest level of security and control over their data stored in the cloud. That commitment extends to Google Cloud GenAI products. Google ensures its teams are following these commitments through robust data governance practices, including reviews of Google Cloud's input data to its products. You can find more details about how Google processes data in the Customer Data Processing Addendum or the data processing agreement for your Google Cloud service.

**Security**

GenAI can create realistic but false content, leading to misinformation and enabling impersonation. GenAI may be used by bad actors to accelerate the sophistication and speed of cyberattacks. It also can be manipulated to provide malicious outputs. For example, through a technique called prompt injection, a third party gives a model new instructions that trick the model into delivering an output unintended by the model producer and end user.

**Accountability**

Accountability in AI means that AI systems and their creators should be held responsible for the systems' actions and outcomes. If an AI system causes harm, there should be mechanisms in place to hold the appropriate parties accountable and ensure that AI systems are used responsibly and ethically.

## Gemma and Gemini Safeguards and Compliance

To ensure trustworthy AI solutions, you can use Google's Gemma models, which represent a suite of lightweight, open AI models that prioritize safety and compliance. These models undergo rigorous training on curated datasets and are fine-tuned using techniques such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), enhancing their effectiveness and safety. Gemma also includes automated safeguards to filter sensitive information from training sets, ensuring data privacy. Additionally, Google's Responsible Generative AI Toolkit provides resources for applying best practices for responsible use to open models.

In addition to Gemma, Google's Gemini AI technology is designed with robust security standards to safeguard your data. Compliant with industry regulations such as ISO/IEC 27001 (Information Security Management) and ISO/IEC 27701 (Privacy Information Management), Gemini also adheres to specific regulations like GDPR, CCPA, HIPAA, and PCI DSS. This comprehensive approach ensures that your AI solutions are not only effective but also meet stringent security and compliance requirements.

# Building Upon Your Current Infrastructure

While it's critical to have the right talent to deliver GenAI successfully, this doesn't require an overhaul of your current teams. Given the standard platform team, delivering a GenAI platform with Google Cloud only requires changes to the infrastructure to easily self-serve and deliver work to production.

A modern delivery infrastructure platform handles the operations that enable the secure deployment of applications to production environments and ensure a high level of quality is maintained.

T**his platform includes capabilities such as:**

- **Application security**
- **Build and deployment tasks**
- **Database migration and upgrades**
- **Fault tolerance**
- **Pipeline orchestration**
- **Resilience testing**

This is all built upon the cloud infrastructure of choice to enable developers to self-serve production deployments and rapidly deliver value for the organization.

# Adding GenAI to your infrastructure

To evolve your delivery infrastructure to support GenAI, you need to include:

- **Access to standard GenAI tools (Google Cloud's Vertex AI, Cloud Run, etc)**
- **Data collection for tracing customer actions throughout the application lifecycle**
- **Enabling TPU/GPU access**
- **Management of your Vector Databases**
- **Versioning strategy for your models**

If we look at our delivery infrastructure platform through the lens of a modern GenAI  enabled company, we start to see the platform evolve with the above capabilities, enabling you to treat GenAI applications like any other application type.

# Future Trends

GenAI is developing at a breakneck pace, so it's important to be aware of future trends to make informed decisions and set up your organization for success. Here are some emerging trends highlighted in Thoughtworks' "AI everywhere" Looking Glass report:

**Adversarial machine learning**: Attacks designed to exploit vulnerabilities in machine learning systems, or attacks carried out by misusing these systems.

**Affective (emotional) computing**: Systems capable of understanding, processing, and responding to human emotions.

**AI agents**: Agents built around models demonstrating more sophisticated behaviors, guided by advanced techniques for task breakdown, analysis, and reflection, allowing for more complex goal-oriented actions.

**AI safety and regulation**: Government guidelines ensure responsible AI use, including monitoring and accountability measures.

**Causal inference for ML**: Techniques helping machine learning models understand cause-and-effect relationships, making them better at generalizing and learning from less data.

**Cross-modal AI**: AI models capable of remarkable cross-modal understanding, enabling advanced image-based Q&A, multi-modal search, and more creative content generation.

**Decision science**: A field combining AI techniques with behavioral and management sciences to improve complex decision-making.

**Digital humans**: AI-powered virtual beings designed to simulate human-like interactions within the metaverse.

**Easing access to GenAI**: New tools and interfaces streamline the use of GenAI, empowering users without specialized technical knowledge to harness these powerful models.

**Fine-grained data access controls**: Precise access controls that consider context when determining who can access specific information.

**Intelligent machine to machine collaboration**: Technologies enabling devices to share information and make decisions autonomously.

**More and better specialization**: GenAI is finding exciting new applications in diverse fields, leading to a proliferation of specialized models and tailored solutions for various industries.

**Production immune systems**: Systems that monitor complex systems, automatically detecting problems and taking corrective actions.

**Scalable models**: Innovative techniques allow for the creation of smaller AI models that retain the power of their larger counterparts, making them more accessible and easier to deploy.

**Trustworthy data**: Techniques for ensuring reliable data origins and responsible data use, especially helpful in tracking sustainability goals.

**Understandable consent**: Making it easier for users to understand terms of service and how their data is used.

# Conclusion

As the rapid development and implementation of GenAI continues, it's clear that GenAI's relevance to businesses is not just a trend but a transformative movement reshaping industries.

The seamless integration of Google Cloud's offerings — including Vertex AI, Cloud Run, GKE, and Gemini models — empowers businesses to navigate and leverage this new technology effectively.

Ultimately, adopting GenAI with Google Cloud's comprehensive tools and frameworks allows businesses to realize significant value, driving competitive advantage and sustainable growth in an ever-evolving digital landscape.

When considering whether to adopt or implement a GenAI strategy for your business, we're here to help. Google Cloud offers products and solutions to effectively build with GenAI, and Thoughtworks has 30+ years of experience transforming businesses through strategy, design, and engineering.

# Additional Resources

Check out the following resources for further learning and exploration.

From a Google research fellow: Generally Faster - The Economic Impact of Generative AI - Googleapis.com (with intro from Google)

McKinsey's take on the economic impact: Economic potential of generative AI | McKinsey

Lenovo's reference architecture

Our own Generative AI and the software development lifecycle: much more than coding assistance | Thoughtworks

Thoughtworks' yearly trend report, the Looking Glass, AI everywhere | Thoughtworks

IBM's paper, Beyond the hype: Creating business value from generative AI

# Appendix

Key Terms

| Term | Definition |
|---|---|
| **Agent** | A component of a GenAI system designed to interact with users, perform tasks, answer questions, or generate content based on input. |
| **Fine-tuning** | A machine learning process in which a pre-trained model is further trained on a smaller, specific dataset to adapt its general capabilities to a particular task. This technique enhances the model's performance on specialized tasks with less data and computational effort than training from scratch. |
| **Foundational model** | A versatile, large-scale machine learning framework trained on extensive datasets, capable of performing a broad range of tasks. Its adaptability allows developers to build specialized applications on top of it without requiring vast amounts of data or computational resources themselves. |
| **Foundational model: Large language model (LLM)** | A type of artificial intelligence that processes and generates human-like text by learning from a vast amount of textual data. LLMs are capable of performing a variety of language-based tasks, such as answering questions, writing content, and engaging in conversations. |
| **Pre-trained model** | Any model previously trained on a dataset before being applied or adapted to a specific task. |
| **Prompt engineering** | The practice of carefully crafting inputs, known as prompts, to guide an LLM towards generating specific output. Examples include providing step-by-step instructions, asking the LLM to adopt a particular persona, or including examples of the expected output. |
| **Retrieval Augmented Generation (RAG)** | A method that combines the retrieval of relevant text from a large corpus with creating new content. Based on a query, this approach fetches documents and then generates accurate and contextually-enriched responses. This process makes the model better at giving helpful and specific results. |

## Vertical-Specific Use Cases

### Automotive and Manufacturing

**Automation**: Implement robotic assembly processes to increase precision and reduce human error.

Keep in mind: This requires significant hardware investments

**Insight generation**: Identify inefficiencies in the supply chain and manufacturing processes, using extensive data analysis despite potential data biases.

**Optimization**: Enhance production line efficiency by optimizing machinery settings, which balances computational demands with the need for constant performance.

**Prediction and forecasting**: Reduce downtime by predicting machinery maintenance needs and part failures.

Keep in mind: It must be calibrated to reliably handle unexpected scenarios.

**Risk management**: Assess safety and compliance risks in real-time, mitigating potential inaccuracies with detailed data analysis.

## Banking, Financial Services, and Insurance

**Automation**: Enhance efficiency and reduce errors by automating routine transactions and data processing.

Keep in mind: Must adapt to handle varied customer interactions.

**Enhanced communication**: Improve customer service interactions through personalized communication utilizing language processing and checking for context relevance.

**Insight generation**: Optimize service offerings by uncovering financial trends and consumer behavior insights.

**Personalization**: Provide tailored financial advice and product recommendations based on user behavior, refining models over time to increase accuracy.

**Risk management**: Assess credit risk and detect fraudulent transactions using detailed data analysis.

Keep in mind: Potential inaccuracies due to AI limitations.

## Energy

**Enhanced communication**: Improve customer interaction by providing personalized energy usage reports and savings tips.

**Insight generation**: Propose sustainability improvements with consumption patterns analysis.

Keep in mind: This requires large, quality datasets.

**Optimization** :Enhance efficiency, reduce waste and balance high resource demand through optimized grid management and energy distribution.

**Prediction and forecasting**: Predict energy demand peaks and optimize renewable energy output.

Keep in mind: This requires accurate historical data to ensure precision.

**Risk management**: Assess infrastructure risks and predict potential system failures, ensuring robust data analysis.

Keep in mind: AI's generalization limitations.

## Government and Public Sector

**Automation**: Increase efficiency for administrative tasks such as document processing and public record management through automation.

Keep in mind: Flexibility is required for diverse inputs.

**Enhanced communication**: Facilitate better public engagement through AI-driven information dissemination and query handling.

**Insight generation**: Analyze large-scale public data to improve urban planning and resource allocation.

Keep in mind: This is constrained by AI's current capabilities in handling complex urban data.

**Prediction and forecasting**: Forecast economic and demographic trends to inform policy making, with validations to ensure the predictions are reliable.

**Risk management**: Assess risks related to public safety and security, leveraging comprehensive data analysis to manage potential inaccuracies.

## Healthcare and Life Sciences

**Enhanced communication**:Facilitate clear patient communication through natural language processing, improving healthcare advice.

Keep in mind: Make sure to monitor these tools for accuracy.

**Insight generation**: Identify potential new drug interactions and treatment effects through deep data analysis.

Keep in mind: AI's current understanding limits.

**Personalization**: Tailor treatment plans and medication dosages to individual patient genetics and histories.

Keep in mind: It's important to address potential data bias.

**Prediction and forecasting**: Improve healthcare resources management by predicting disease outbreaks and patient admissions.

Keep in mind: This requires precise and validated models.

**Risk management**: Enhance patient outcomes through monitoring patient health and predicting complications with advanced analytics.

Keep in mind: Data limitations.

## Media and Entertainment

**Content generation**: Enhance viewers' engagement with personalized content streams and recommendations.

Keep in mind: Must manage the derivative nature of AI-generated content.

**Creativity augmentation**: Assist in developing new concepts for shows and movies, complementing human creativity.

**Enhanced communication**: Utilize advanced language models to generate interactive and engaging marketing materials.

Keep in mind: Be cautious of generating nonsensical outputs.

**Insight generation**: Analyze viewer data to predict trends and preferences to develop targeted content while still relying on existing data patterns.

**Personalization**: Refine recommendations iteratively to improve user experience by adjusting to user preferences and viewing habits.

## Retail and Consumer Goods

**Automation**: Adapt to and handle peak times while reducing labor costs to streamline your inventory management and customer service interactions.

**Enhanced communication**: Create personalized marketing campaigns trained on consumers' behaviors to improve overall customer engagement.

Keep in mind: Avoid generating nonsensical content due to AI's comprehension limits.

**Insight generation**: Optimize stock levels and identify trends by analyzing sales data based on new, unexpected market conditions.

**Personalization**: Tailor product recommendations to improve the customer experience with consumer data.

Keep in mind: Data privacy concerns.

**Risk management**: Improve your monitoring and predictions of market changes and supply chain disruptions utilizing robust data validation.

## Supply Chain and Logistics

**Adaptation and learning**: Enhance system responsiveness to global supply chain dynamics by continuously updating models based on new shipping data.

**Automation**: Optimize for efficiency through the automation of routing and scheduling of deliveries.

Keep in mind: Adaptability to real-time traffic and weather conditions.

**Prediction and forecasting**: Forecast supply and demand needs to adjust inventory levels.

Keep in mind: Carefully validate to ensure reliability.

**Optimization**: Optimize logistics networks by analyzing travel times and costs.

Keep in mind: Consider AI's limitations in handling new routes.

**Risk management**:  Ensure accurate risk assessment by monitoring potential supply chain disruptions and applying robust validation techniques.

## Telecommunications

**Automation**:: Minimize manual effort and ensure adaptability by automating customer service interactions for routine inquiries and complaints.

**Enhanced communication**:  Enhance user interaction with AI-driven customer support, leveraging natural language processing.

Keep in mind: Potential inaccuracies.

**Optimization**: Optimize network traffic and resource allocation to improve service quality, considering the high resource demands of AI.

**Prediction and forecasting**: Ensure customers get the right level of service by predicting network load and scaling resources accordingly.

Keep in mind: This requires continuous adaptation to handle unexpected usage spikes.

**Risk management**: Identify and mitigate potential service disruptions and security breaches, applying robust preprocessing to enhance reliability.