Google Cloud

Real-world generative AI adoption.

Insights from
Google Cloud Consulting.

Table of contents

Introduction

Generative AI has quickly become a major focus for businesses. Since its rapid rise in 2022, interest has only increased, with organizations across industries exploring how best to adopt and integrate the technology. Until now, the story of enterprise technology was all about infrastructure—servers, networks, and databases were the core focus. IT teams were constantly battling capacity limits, optimizing performance, and trying to keep everything running smoothly. Think late nights troubleshooting outages and the constant pressure to keep up with growing demands. It was a steep learning curve, but it also provided a deep understanding of the nuts and bolts that make modern computing possible.

That experience proves incredibly valuable now. While there's a huge buzz around generative AI, its true potential hinges on more than just clever algorithms. It requires a robust and scalable infrastructure to handle the massive computational demands and data flows. Generative AI is here, and your foundation is built to support it.

Traditional AI and machine learning models rely on precise statistical methods, extensive data training, and strict validation processes. These models excel in delivering high accuracy but often require longer development cycles and operate within established governance frameworks.

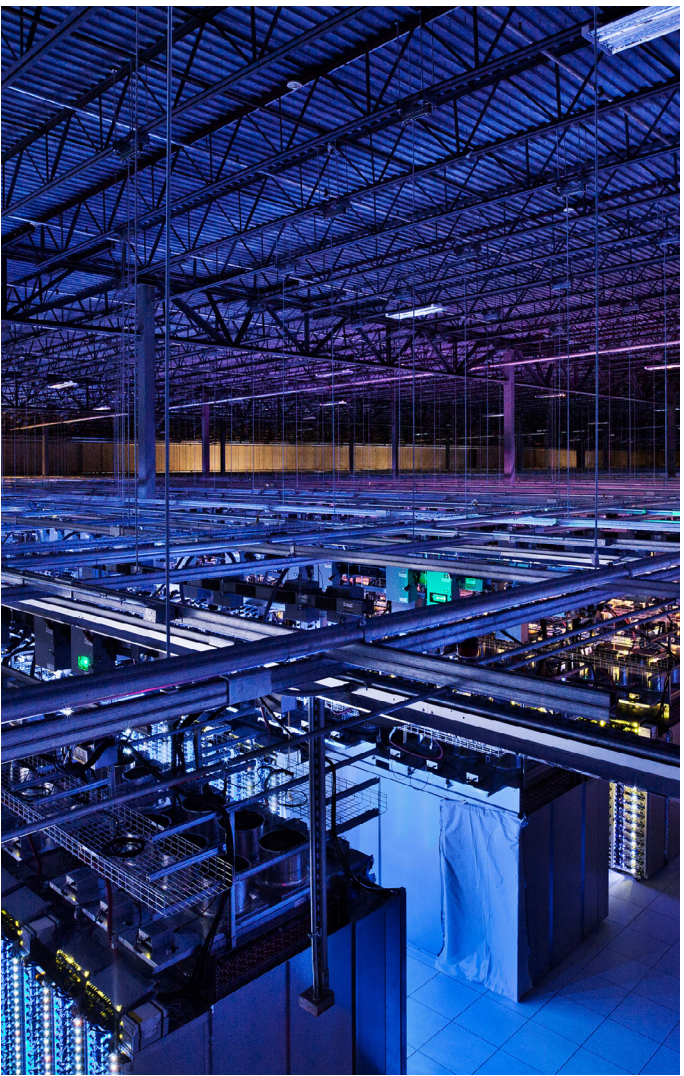
Generative AI, powered by large language models, takes a more dynamic and iterative approach to development. Out-of-the-box models like the Gemini and Imagen suites can generate new content, extract and summarize information, and engage in human-like conversations. These capabilities open the door for rapid experimentation and innovation in areas like customer service, sales automation, and employee productivity.

To help you navigate this evolving landscape and achieve value at scale, this ebook shares insights from recent customer engagements delivered by Google Cloud Consulting. These lessons cover the full journey, from demos and developing proof-of-concept projects, to building minimum viable products and deploying user-focused applications in production.

With the right approach, we can help you cut a clear path forward. Our extensive first-hand experiences in selling technology, building solutions, and witnessing the transformative potential of generative AI are here to help guide you. We'll provide you with a roadmap, highlighting key considerations for successfully bringing a generative AI-powered application to market, just as we've done for numerous customers.

Three phases to generative AI adoption at scale

Generative AI's enterprise adoption depends on proving repeatable value at scale and in a sustainable way. To accomplish that, this journey is best understood through a phased approach, consistently guided by your business strategy. Sound familiar? This is how you built your existing cloud infrastructure.



Phase 0:
Before thinking about launch, it's important to look at your foundations. Data readiness, infrastructure capabilities, and a clear understanding of ethical implications are paramount; otherwise, you're building a house on sand. To help establish a solid infrastructure foundation in the cloud, explore our Modern Infrastructure and App Development Resources Hub. Now, at this stage, you also want to decide on your first generative AI use case (more on this in the next section)

Phase 1:
Once a foundation is built and a viable use case identified, priorities shift to selecting the right models, establishing robust data pipelines, and defining KPIs—a crucial step often overlooked in the initial excitement.

Phase 2:
While developing and deploying an application for multiple use cases, continuous monitoring, iterative improvements, and addressing user feedback are critical to achieving widespread adoption.

Think of it as building a bridge—each phase requires careful attention to ensure stability. Skipping any step could result in a structure that looks good but ultimately can't support the weight of real-world use. Successfully navigating these three phases ensures you're not just dabbling in generative AI, but genuinely harnessing its transformative power.

Nine things we've learned about driving generative AI adoption

Where do I start?

01

Strategy: Play use-case chess, not checkers.

Embracing generative AI successfully, just like any other tech, requires a strategic, not haphazard, approach. Think of it as developing a blueprint to follow before building a structure.

First, understand your organization's capabilities—looking at resources, infrastructure, and existing skills—through pragmatic experimentation with a use case. This initial understanding is crucial. Prioritize investment based on this intel, focusing on areas ripe for AI-driven innovation. Most importantly, set clear, measurable objectives for generative AI adoption. Without a defined goal, you risk spreading resources too thin across multiple projects, leading to minimal returns —a costly and frustrating outcome.

Sound obvious? This iterative approach, focused on learning and adaptation, helps avoid costly mistakes and maximizes returns on investment. Balancing planning and experimentation isn't easy, but it's essential for leveraging generative AI effectively. In short, strategic planning and iterative learning are key to avoiding aimless experimentation.

Our experience shows that initial use cases are often the most challenging.

So, how do you choose your first use case? Look for a use case that will provide measurable business benefits, but is also feasible to implement. It needs to be relevant, with access to the right data and people willing to invest in optimizing the status quo. To simplify the process, start with data dumps and focus on a limited number of users, without stressing over system integration as a first concern. The use case should align with your core business and impact a broad group of users who are excited to see the benefits in action. Some appealing examples might be data analysis, customer service automation, internal process automation, code generation, and content creation as popular use cases. Automating routine tasks is often a good starting point. But above all, remember: simplicity is key.

Need some extra use case inspiration? Check out this blog: [601 real-world gen AI use cases from the world's leading organizations.]

02

Adoption: Champion AI for stakeholder buy-in.

Scaling generative AI needs strong support and leadership. Success depends on getting buy-in from key stakeholders. Leaders must champion AI adoption and cultivate a culture of change and innovation. Identify the champions within your organization—those excited about AI and equipped with the right mindset for managing change. Think who your DRIs (Directly Responsible Individuals) will be and who will lead the charge for adoption. Think about the talent and expertise required, as well as the team structures that will support this shift. Ask yourself: Which structures will align with the continuous experimentation and learning required? How will I foster a culture of relentless testing and fast learning within my organization?

COMMERZBANK

Customer story

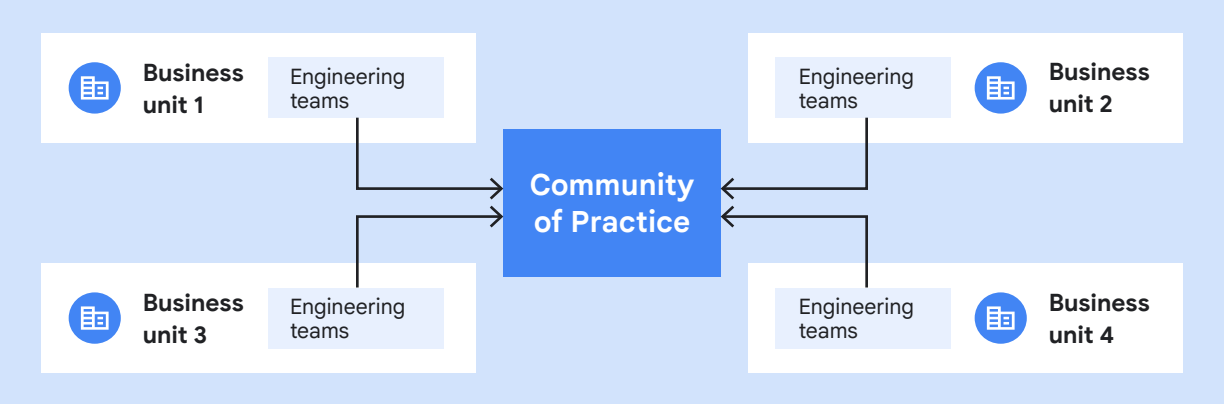
Company profile: Headquartered in Frankfurt, Commerzbank is a leading German bank with over 155 years of history, offering financial services to private and corporate clients, with a focus on small and medium-sized enterprises. It operates in over 40 countries.

Commerzbank established a strong team around individual use cases while preparing for generative AI at scale.

Commerzbank follows a two-fold approach, which ensures technical depth and execution power where needed. They have aligned dedicated teams and resources around use cases with strategic importance. These teams are composed of strong product owners who take full accountability. Dedicated Project Management Office (PMO) support ensures milestones are defined and reached, and that respective governance bodies are provided with the relevant information to make go or no-go decisions. This avoids decision bottlenecks, allowing teams to progress within sprints. It also ensures continuous alignment with business objectives and efficient use of resources while allowing room for experimentation.

Technical findings are shared in technical working sessions, progress is evaluated with the business in the same room as tech teams, and continuous communication with the Cloud and AI platform team ensures a balance between autonomy and standardization within the organization.

Lastly, to allow the entire organization to benefit from the findings of individual use cases, communities of practice facilitate cross-domain enablement, a culture of knowledge sharing, and above all, momentum to drive innovation. These communities are organized by dedicated people, and time is carved out to ensure these exchanges can happen.



03

Priorities: Choose speed, control, or both?

Enterprises adopting generative AI face a crucial choice: prioritize speed and innovation or maintain strict control—walking the tightrope between cost and transformation. This isn't a trivial decision; it requires careful consideration of your company's specific needs and use cases.

Tighter control often means better budget management, while embracing innovation boosts the potential for disruptive business improvements. However, too much control can stifle creativity. Prioritizing transparency and traceability encourages innovation while enabling quick adaptation to change.

Finding the right balance between speed and control is key to your success.

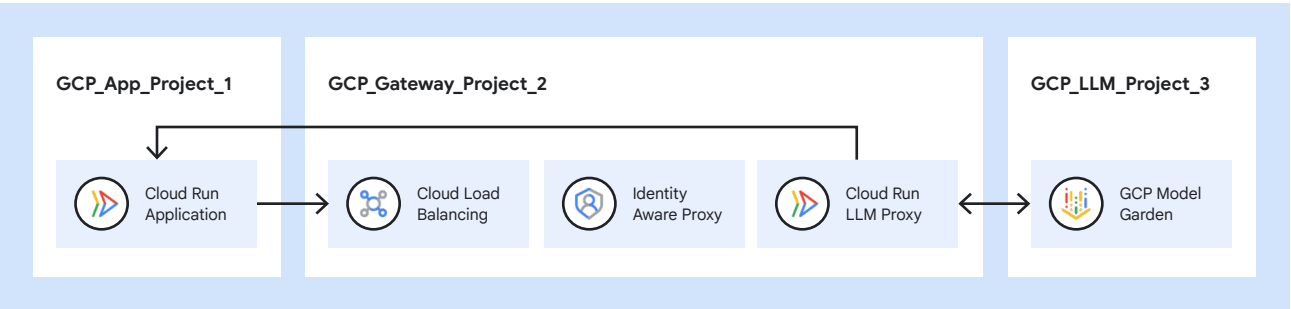
This challenge should feel familiar. Whether it was your first on-premises data center architecture or your first cloud deployment, You've navigated similar decisions before. Remember to apply those lessons here.

Ensuring balance between both financial prudence and groundbreaking advancements aren't one-size-fits-all. it requires careful evaluation and constant adaptation, and failing to assess your approach may leave you falling behind.

At this stage, you might wonder: is there a middle ground? At Google, we tend to favor speed, decentralization, and autonomy, but many of our customers lean towards control, standardization, and centralization for efficient resource allocation. The right approach depends on your specific setup, resources, and strategy. However, we predict most companies will land somewhere in between—a hybrid approach that balances both sides.

As a result, control mechanisms come in reactive or proactive shapes and forms.

Illustrative example: Balancing speed and control while implementing proxy and gateway solutions



For example, reactively monitoring the use of LLMs within your organization—via proxy and gateway solutions or a monitoring and logging concept—is extremely useful to facilitate both cost control and transparency into who is doing what within your organization. Having a proxy or gateway solution in place simplifies things for platform teams: you can only improve if you understand what's happening. Additionally, controlling which LLMs are used boosts efficiency in quota management and resource allocation, while enforcing key security measures.

Customer story

Company profile: A global fashion retailer based in Europe, offering trendy clothing, accessories, and cosmetics at affordable prices. With over 100,000 employees, it operates worldwide and has a strong online presence. They are committedto sustainability, using recycled materials and promoting ethical sourcing.

Global fashion retailer is building a generative AI Gateway to streamline LLM applications across the organization.

With the ever-increasing supply of models, it can become difficult for developer teams to know what's best for them. For this clothing company, this is where a central generative AI platform team comes into play. This team is not only aligned with common best practices provided by the Google Cloud Platform team (Infrastructure, deployment patterns, network design, CI/CD etc.), but has embarked on a mission to build a central Gateway, monitoring the use of LLMs across the organization.

The Gateway acts as a central onboarding layer, where developers wanting to build an LLM application can obtain an API key to a model of their choice. In return, they receive security as a service and mechanisms for cost control, which is central to the technology agenda. The Gateway allows developers to track performance across models, provides flexibility, and prevents vendor lock-in. This means a conscious decision has been made by the platform team to own a certain degree of maintenance. This could be in the form of updating SDKs, provisioning environments, and more.

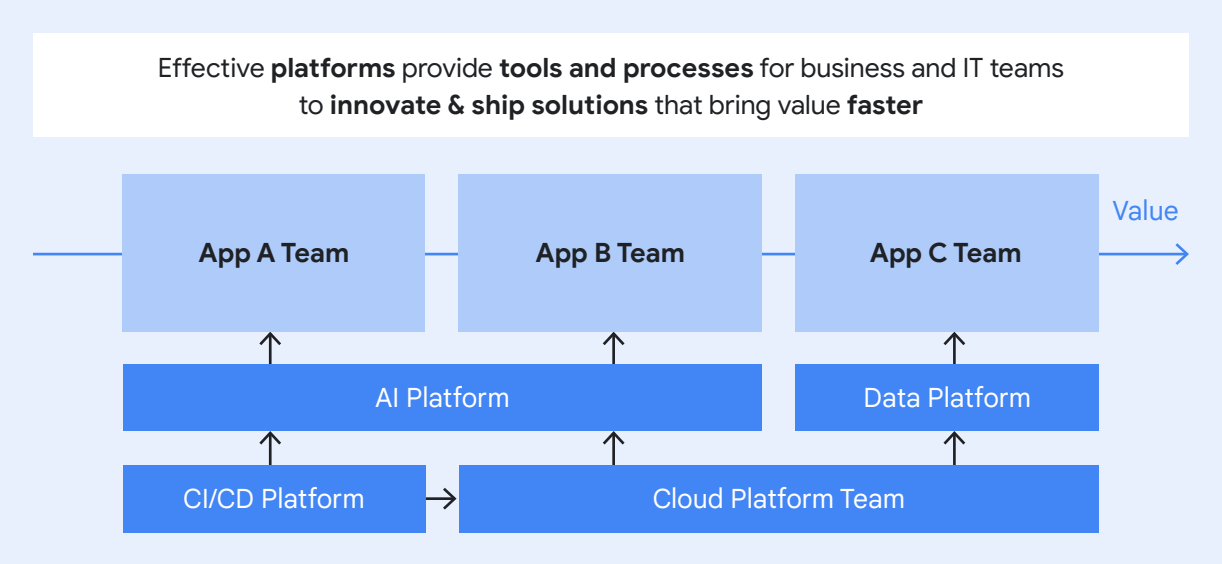
An added benefit of having a generative AI platform team is that it allows application teams to focus on developing the business logic required for their specific use case. This sets them up for scale, and as time evolves, provides an opportunity to standardize individual components of generative AI applications. This includes RAG as a Service, terraformed deployment with Cloud Run, embeddings as a service, or evaluation blueprints to ensure high-quality LLM applications across the organization.

04

Implementation: Decide on a platform.

Choosing the right platform is crucial, ideally open-source and allowing for experimentation. As LLMs become commoditized, managed services offer a compelling path to rapid deployment and scalability—a boon for business stakeholders. While managed services might initially limit developers' complete transparency, the accelerated experimentation enabled will likely win them over. This faster pace allows businesses to quickly test, learn, and adapt, ultimately providing a competitive edge. Having the courage to enable AI platforms to focus on what's important for your business at scale, will ultimately pay dividends in the race for AI value.

View your platforms as products that serve the needs of internal teams. Well-designed platforms, addressing foundational layers like security, data governance, infrastructure provisioning, and core API access, can dramatically accelerate the work and value of individual teams. This allows them to focus on their core value-added and end-user applications, rather than each team reinventing the wheel for common needs. Teams can focus their energy, and skill development, on the unique value they bring to the business and user challenges.



A dedicated platform team, focused on AI, can build reusable components, manage models, and abstract complexity of other Cloud infrastructure layers—including the application and validation of crucial security constraints.

Platforms also drive strategic acceleration in areas with transformative opportunities, like generative AI. Google uses this approach to focus investments, distribute tools with best practices and security, and centralize lessons learned, aligning talent with specialized skills.

I have a use case, what now?

05

Quality: Don’t blame hallucinations if your data is the problem.

Generative AI adoption hinges on data quality, not just sophisticated model selection. Blaming hallucinations solely on the model is like blaming a chef for a bad dish when the ingredients were rotten. High-quality data is paramount, and requires a proactive approach from the outset. This includes creating clear feedback loops to continuously improve data quality and, importantly, fostering a data-centric AI culture, moving away from the model-centric approach that often dominates.

You can think of a strong data governance program as the invisible scaffolding supporting the whole AI structure, it’s key to implementation. Successful enterprises understand data governance and reap the rewards from actively investing in such programs. In short, without clean data, even the most advanced generative AI model becomes a sophisticated parrot, squawking nonsense. Therefore, focusing on data quality is not just good practice; it’s the foundation upon which successful generative AI adoption is built.

Customer story

Company profile: Erste Group is a financial services provider headquartered in Vienna, Austria. The company operates across seven countries in Central and Eastern Europe. Erste Group is known for its focus on digital innovation and customer experience, and has a strong drive for innovation in FSI.

Erste Group addressed data accuracy from the start, to successfully deploy generative AI-driven SQL and RAG agents.

Erste Group wanted to add a conversational layer on top of structured data. The goal was to enable research analysts to talk to their data and query across hundreds of research reports. While the application of generative AI looked promising, an interesting discovery quickly emerged: data quality is key and correlates directly with minimizing hallucinations.

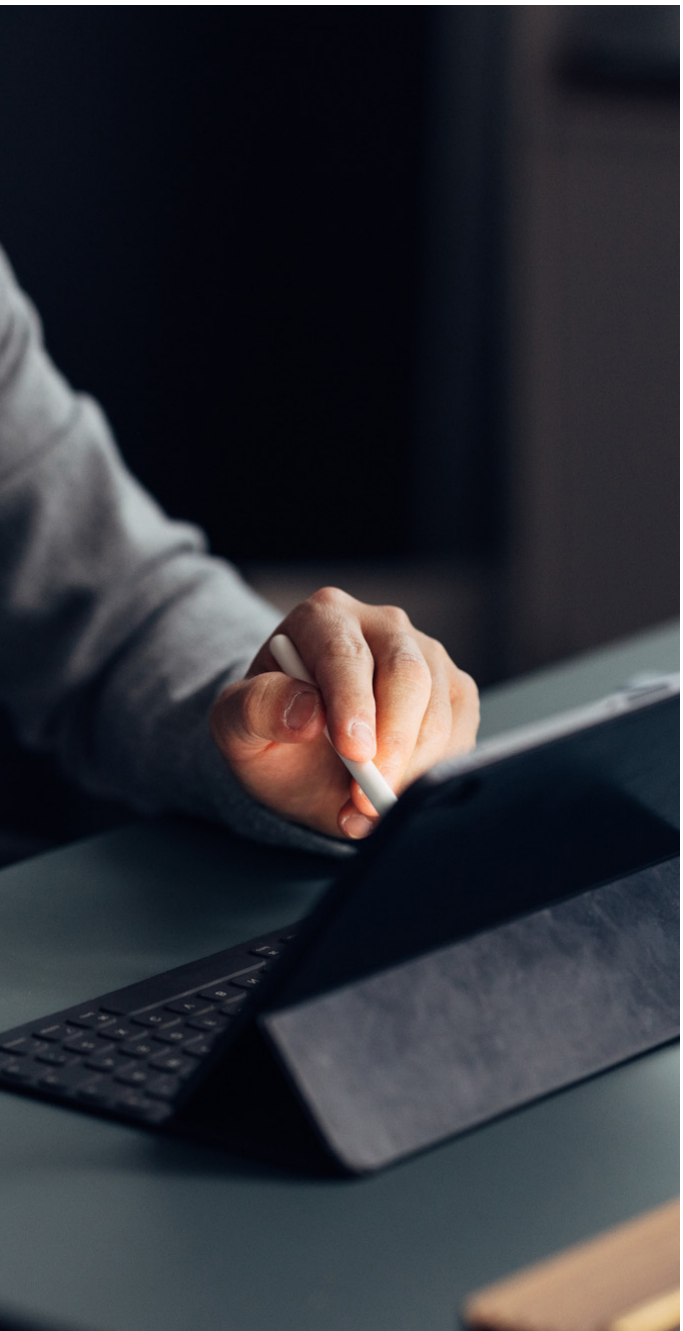
Evaluation was a key pillar in developing the agents from the start. When debugging question-answer pairs to determine answer quality, the first results suggested that the LLM solution deployed was hallucinating. After further evaluation, the team uncovered that some of the reports had referenced unwanted information. While this is not a significant issue, it did demonstrate the power of LLMs when it comes to large sets of unstructured information.

Even more importantly, RAG agents require a large corpus of information to be indexed and embedded in a vector database—something Erste Group had. Now, when an LLM queries that database, in this case using Vertex AI Search (VAIS), VAIS provides the LLM a corpus of relevant and consolidated information in the form of snippets, extracts, and a summary. The LLM essentially uses that pile of information to create a final answer. As you can imagine, the better that corpus of documents is prepared and tagged with metadata, the more powerful the retrieval service of RAG applications will be. The moral of the story? Data quality and preparation remain key to ensuring that LLMs provide the answers they are supposed to.

06

Reality: Focus on ground truth.

Integrating generative AI successfully requires a strategic, phased approach to avoid common pitfalls. Start small; identify a single, high-impact use case, establish a measurable evaluation framework, and gather user feedback to create a baseline dataset.



Then, develop a robust testing framework, refine your AI iteratively, and work with a small group of power users before scaling. Optimize your AI against this initial dataset, gradually expanding the scope of your training data to avoid being overwhelmed by unexpected data idiosyncrasies later on. Remember, focusing on the user experience is paramount; all other considerations flow from meeting their needs. By following this measured, iterative process, enterprises can successfully navigate the complexities of generative AI adoption and reap its substantial rewards, avoiding the pitfalls that sink many a well-intentioned project.

To limit hallucinations and ensure relevant insights, ground LLM responses in your own data or public knowledge bases. Generative AI tools now support grounding with native connectors to enterprise knowledge and external sources, allowing agents to provide accurate, unique information. Techniques like Retrieval-Augmented Generation (RAG) help LLMs access and synthesize multimodal data, enhancing response accuracy while maintaining security and integration within your organization.

SIGNAL IDUNA

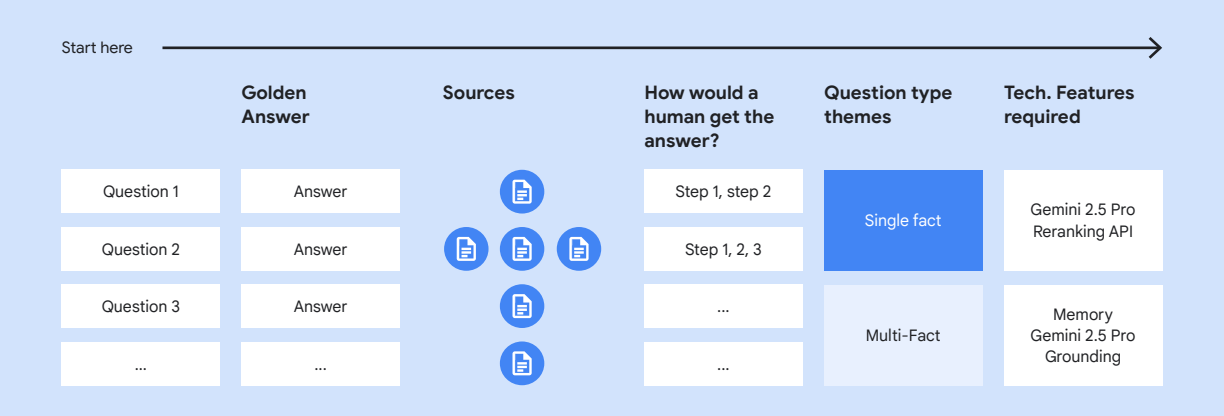
Customer story

Company profile: SIGNAL IDUNA is a German insurance and financial services group offering life, health, property insurance, and investments. Primarily operating in Germany with a presence in other European countries, it focuses on comprehensive financial security and has around 10,000 employees.

SIGNAL IDUNA established an extensive ground truth data set, continuously evolving, to ensure end-user satisfaction.

Companies that have embarked on the generative AI journey right from the start, have also been the first to experience the drawbacks of a new technology. In the case of generative AI, this is non-determinism, where prompting an LLM to perform a certain task may yield variations in outcomes, despite keeping the prompt and task constant. As the field progresses, both Google and users alike have worked on minimizing the cause of the symptom and dealing with it in hindsight. One of the best practices leveraged by SIGNAL IDUNA is a strong focus on a ground truth data set. It's important to state that the ground truth varies between use cases.

But what is the ground truth? In generative AI, ground truth refers to the actual, verifiable, and accurate data or information about a specific context that the AI model is trying to represent or learn from. It's the objective reality that serves as the basis for training, evaluating, and validating the model's outputs. Think of it as the gold standard against which the AI's performance is measured. When you want to see how well a model is performing, you compare its outputs with the ground truth. For example, if SIGNAL IDUNA's AI model generates an answer to an insurance policy-related question, they can compare that answer to an accurate, human-written description (the ground truth answer) to assess the quality of the generated answer. Another example may be, if you want the model to answer a question based on your knowledge base, you compare the generated answer with the optimal answer of a human.



In case of a RAG application: Develop a golden data set, which represents your user questions and work towards features required. This provides you with an indication on development complexity and acts as your evaluation dataset.

In the case of SIGNAL IDUNA, they're dealing with an RAG-based application. So the ground truth contains question-answer pairs, where questions represent insurance clerk queries with the end goal of addressing end customer questions. Answers, on the other hand, contain the optimal answer based on insurance policies. Therefore, the ground truth also contains information on source data, like the page numbers the information was extracted from. Every time the RAG application's features are optimized or additional features are implemented, the developer team triggers an evaluation pipeline containing certain metrics (for example, groundedness, answer quality, or metrics around retrieval quality). The results allow the team to trace application and LLM performance. Over time, it's important to reassess the accuracy of the question-answer pairs to ensure business logic is up to date and the wide array of answer scenarios are as close to the real world as possible.

So what comes first: An evaluation set or end-user testing? The answer is both. It's crucial to think about evaluation right from the start. It doesn't need to be fully automated or visualized with a Dashboard. Notebooks can help find the right balance between evaluation work and actually building the application.

07

Collaboration: Work together for a strong foundation.

Integrating generative AI into production demands compliance with fundamental software engineering principles—but you don't need to reinvent the wheel. Collaboration with platform and architecture teams who've navigated cloud infrastructures and application deployments before is essential. You should adhere to foundational design patterns on your code (such as modular design), ensure testability is prioritized, develop in provision staging and testing environments (not production), document your code, and ensure encapsulation and separation of concerns. The list goes on.

With the importance of collaboration, you need to have empowered product teams with a deep understanding of genuine user needs, who are also willing to work with other teams. Ignoring these best practices is like trying to build on a foundation of sand—a recipe for disaster. Effective generative AI implementation requires a robust, collaborative approach, grounded in established methodologies. Ultimately, the goal remains consistent: when delivering value to users, generative AI is a powerful tool in the journey, not a magical solution that overrides best practices.

Adoption at scale, what's the secret sauce?

08

Assurance: Trust is hard to win and easy to lose.

Enterprise adoption of generative AI can be hindered by a significant trust deficit, which requires careful bridging rather than a quick fix. Despite the appeal of increased efficiency, only 24% of developers express substantial trust in AI-generated code, according to 2024 DORA research.

This seed of doubt isn't easily overcome and demands a proactive, sustained effort. The DORA research (outlined in more detail at the end of this eBook) identified the five most effective strategies to address developer concerns (more on these strategies later).

Another concern is about long-term career impact. While not immediately critical, it does show a potential future challenge that COEs should monitor and proactively address. Ignoring these concerns risks the whole generative AI project becoming a colossal, expensive, and ultimately useless effort. Cultivating trust is paramount for successful enterprise-wide AI adoption. Without it, the journey is likely destined for failure.

[1]The 2024 DevOps Research & Assessment (DORA)

Customer story

Company profile: Genuine Parts Company (GPC), founded in 1928 and headquartered in Atlanta, is a leading global distributor of automotive and industrial parts. With a vast network across North America, Europe, and Australasia, GPC serves both individuals and large corporations.

Genuine Parts Company (GPC) established a special projects team to focus on high-value generative AI use cases.

Understanding the need for focused investment and a data-driven approach, GPC's Center of Excellence is spearheading the exploration of generative AI to streamline internal processes and enhance customer experiences. This strategic decision allows GPC to start small with promising use cases, quickly demonstrating value to the organization and securing buy-in for further investment.

This approach also enables a crucial feedback loop with users, allowing GPC to refine its operating model and tailor solutions to specific business needs.

By fostering a culture of experimentation and learning, GPC is building a strong foundation for long-term success with generative AI in a fast-paced environment. Their AI Innovation Center managed to have two products go live in the first year of operation, with multi-million dollar efficiencies generated for the organization.

By earning trust from business units and the board through early successes, GPC can validate its 'hybrid AI' approach—that leverages central capabilities while empowering local business and IT teams—and continuously improve their IT platforms and processes.

[1]The 2024 DevOps Research & Assessment (DORA)

09

Continuous adaptation: Set yourself up to evolve with AI.

The field of AI is constantly evolving. It requires mechanisms for continuous adaptation and a culture of learning. Just as we were scratching the surface of generative AI, we see new models and ideas emerging. From RAG, to Agents, to agentic AI capable of memory, reasoning, and the human-like ability to carry out complex tasks. Companies and leaders need to be prepared to adapt and learn continuously.

Allowing space to grow and adapt often gets put on the back burner—because there's usually a more pressing issue around the corner. The DORA research program demonstrates that organizations which are most successful at adopting generative AI encourage ample time for focused learning as well as education events on trends that move their industry. Experimentation with new technologies is important to rethink the present and always challenge the status quo.

What the research tells us?

The 2024 DORA research study identified five strategies that drive successful generative AI adoption by developers.

Be transparent about how you plan to use AI.

Providing developers with information about your goals and adoption plan—as well as addressing procedural concerns such as permitted code development and available tooling—can improve adoption by 10-12%. Transparency is also key to developing a healthy organizational culture.

Alleviate anxieties about AI-related job displacement.

Leaders who address concerns in clear terms enable developers to focus on learning how to best use AI, rather than worrying about its impact on their job security. The research estimates that this can boost adoption by 125% compared to organizations that ignore these concerns.

Allow ample time for developers to learn how to use AI.

Our data indicates that individual reliance on AI peaks at around 15 to 20 months—and that giving developers dedicated time during work hours to explore and experiment with AI tools leads to a 131% increase in team AI adoption.

Create policies that govern the adoption of AI.

This provides developers with a framework to confidently, responsibly, and effectively use these tools. Organizations that do this show a 451% increase in AI adoption compared to those that don't.

Encourage ongoing engagement with AI resources.

Simply providing resources is not enough; organizations must actively create opportunities for developers to experiment and integrate AI into their workflows. Without this continuous support, adoption rates will lag significantly.

About the DORA research.

The DevOps Research & Assessment (DORA) program is a long-running Google research program that seeks to understand the capabilities that drive software delivery and operations performance, which culminates in the publication of the annual State of DevOps Report.

Download the full 2024 DORA report here.

[1]This article from Google Cloud CISO provides additional guidance on how to craft an Acceptable Use Policy for generative AI.

Accelerate your AI journey with Google Cloud Consulting

The journey to generative AI at scale is complex. Google Cloud Consulting works alongside you to speed up time to value, reduce risk, and ensure your AI initiatives deliver transformative business growth. We combine deep product knowledge with an innovation mindset to help you succeed at every stage.

How we help you succeed on your AI journey

We align our expertise with the key phases of AI adoption discussed in this paper, providing tailored guidance and hands-on support where it matters most.

Phase 0: Build Your Foundation

Before you launch, we help you establish the strategic and technical groundwork for success.

AI Readiness & Strategy: Assessing your current state and building your AI blueprint to meet key business needs.

Data & Infrastructure Readiness: Designing robust data governance models and ensuring your infrastructure is ready for the demands of generative AI.

Operating Model Design: Defining the right team structures, governance, and processes to support your AI initiatives, while fostering a culture of innovation and identifying the expertise required for success.

Phase 1: Prove the Value

As you identify your first use case, we provide the deep expertise to build impactful solutions and get it right the first time.

Use-Case & ROI Analysis:
Prioritizing high-impact use cases and defining clear, measurable objectives for adoption.

RAG & Grounding Expertise:
Implementing cutting-edge grounding techniques to connect models with your enterprise data for high-quality, relevant responses.

PoC & MVP Implementation:
Building curated architectures and accelerating the development of your first AI-powered applications.

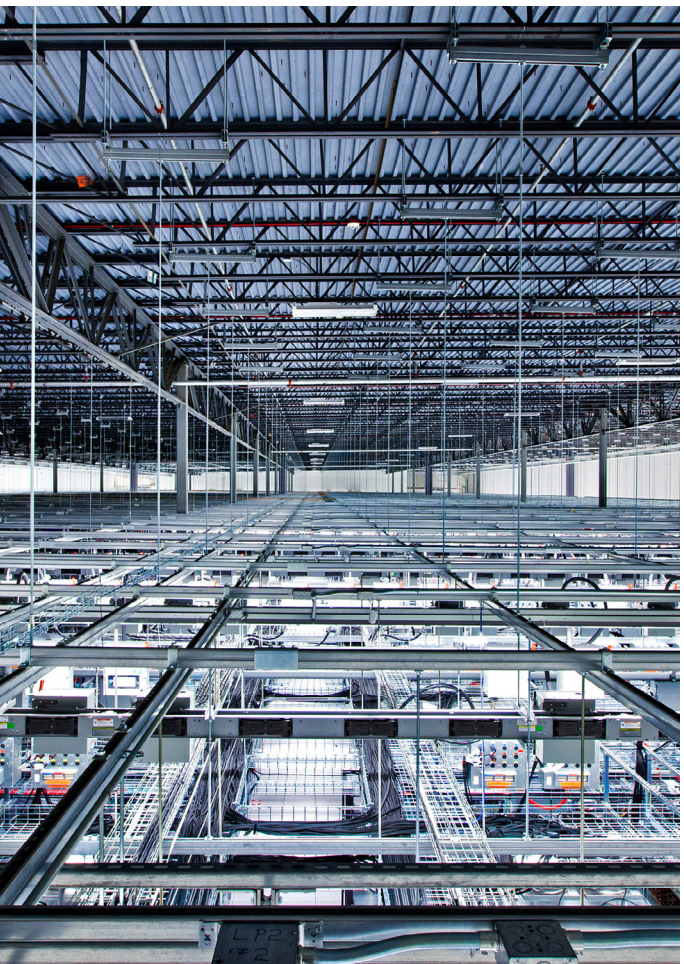
Phase 2: Scale with Confidence

As you move from a successful pilot to enterprise-wide deployment, we help you optimize, scale, and solve for the future.

MLOps & Platform Engineering:
Guiding you through best practices for deploying and managing models at scale to ensure operational health and stability.

Responsible AI & Security:
Implementing the necessary guardrails and security measures to build and maintain trust in your AI systems.

Enterprise Enablement:
Empowering your organization with customized training programs so your teams can leverage AI effectively and autonomously.



Let's build your AI future, together

Learn more about Google Cloud Consulting services at cloud.google.com/consulting or continue the conversation with your sales representative.

What's next?

No matter where you are on your generative AI adoption journey, Google Cloud can support you with products, solutions, and services. Here are some ways to get the AI ball rolling:

For solution **recommendations for your use case** use AI to generate a solution guide, including reference architecture and available pre-built solutions.

To quickly understand your **strengths and opportunities for AI adoption** take the AI Readiness Quick Check. It will provide you with best practices against the six dimensions in our AI adoption framework. For in-depth support, the AIR Program is a three-week engagement designed to accelerate value realization from your AI efforts.

Ready to accelerate your AI journey? Connect with Google Cloud Consulting experts today.

For help **migrating infrastructure** to help you accelerate your efforts, check out this migration guide and checklist or sign up for a complimentary migration and modernization assessment.

Explore an executive's guide to delivering value from data and AI with 5 steps for maximizing AI success in your organization.

Identify the best AI-ready platform for your developers and see how the right platform strategy will help your teams succeed with AI.

Ready to connect?

Let's chat

Google Cloud