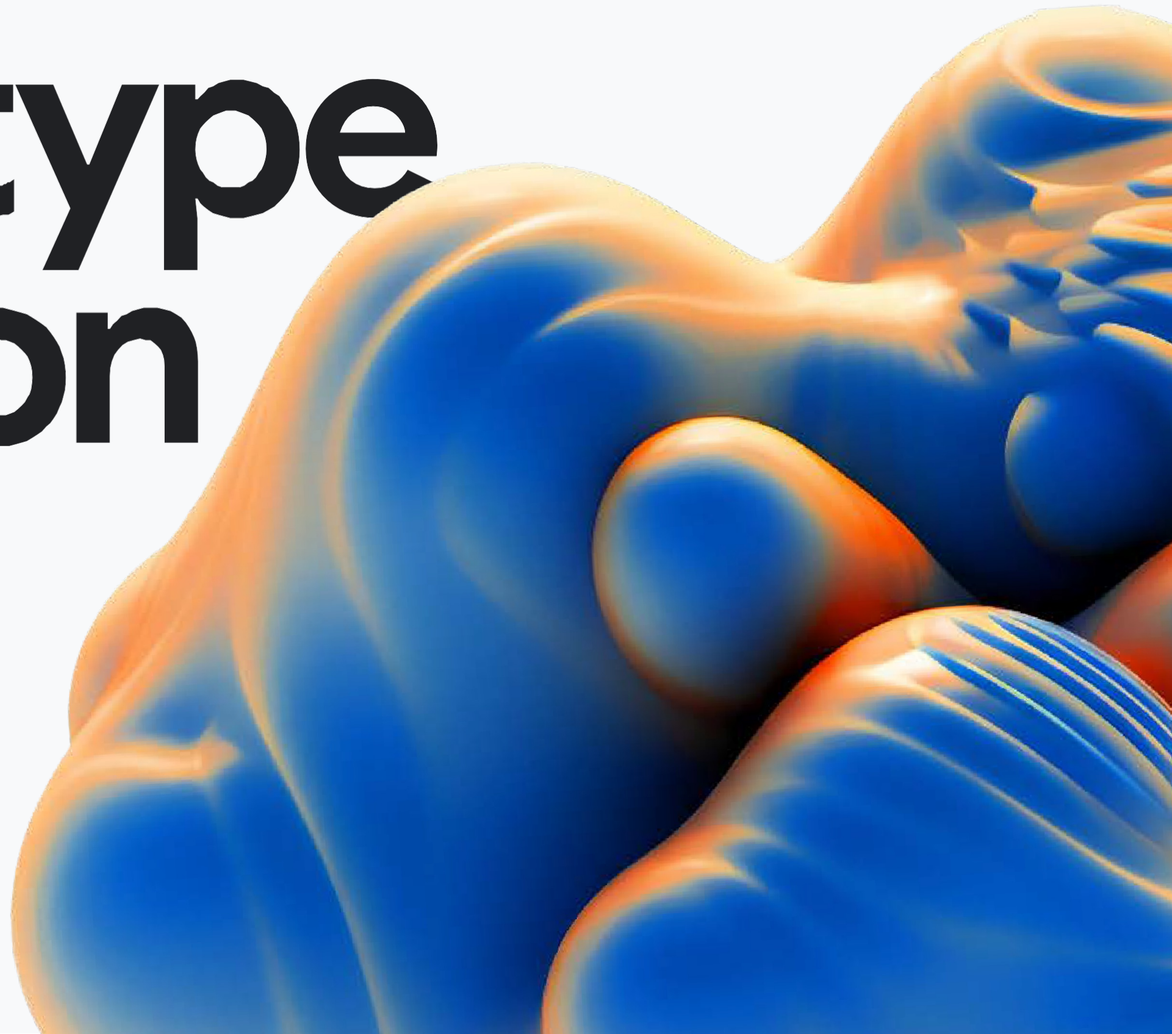
Google Cloud endava

From prototype to production

Your step-by-step guide to scaling generative AI



The emergence of generative AI is revolutionizing the field of artificial intelligence.

Foundation models—such as the large language models (LLMs) used to generate content and power AI agents—are powerful tools that enable enterprises to drive efficiency gains and launch innovative offerings.

Many businesses have prioritized pilot deployments of gen AI, using models to create new content, translate languages, produce different text formats, and answer your customer and employee questions.

We're inspired by what enterprises like yours have been building. We've seen a staggering 36x increase in Gemini API usage and nearly 5x increase of Imagen API usage on Vertex AI this year alone—demonstrating that enterprises are moving from gen AI experimentation to real-world production.

But extracting value from gen AI for your enterprise isn't as simple as typing a query into a model and getting a response. Taking full advantage of gen AI's capabilities requires a comprehensive strategy, including model selection, prompt management, evaluation, integrating retrieval augmented generation (RAG), and more.

It can feel overwhelming. But it doesn't have to be. This guide shares critical learnings from how our customers have moved from AI experimentation to production, and will help you get started.

-Saurabh Tiwary
VP & General Manager, Google Cloud AI

Google Cloud endava

We're at a turning point.

Revving your gen AI pilots into full production can be complex. But there's never been a better time to do it. Gen AI is no longer a blue-sky fantasy or a 'nice-to-have'— we've reached the turning point where gen AI is essential for remaining competitive. In fact, 61% of enterprises are running gen AI use cases. And once an enterprise has successfully implemented initial AI use cases, they can focus on efficiently scaling to additional applications, establishing proper governance, reusable components, and a comprehensive AI platform.

In this ebook, we'll share a step by step guide to overcoming the challenges of building with gen AI. Using learnings from our two decades of operationalizing AI at scale, we'll show you how to leverage proprietary or open models, rather than constructing new foundation models.

00

Clarify your AI objectives

Your business objectives define your starting point.

01

Choose the right model

Your model makes all the difference.

02

Start with evaluation

You can't improve what you don't measure.

03

Improve model behavior

Continuous model improvement delivers results.

04

Release, validate, and deploy

Go for launch.

05

Continue to monitor, effectively

Monitoring and maintaining your AI for production.

Google Cloud endava

Clarify your AI objectives

Your business objectives define your starting point.

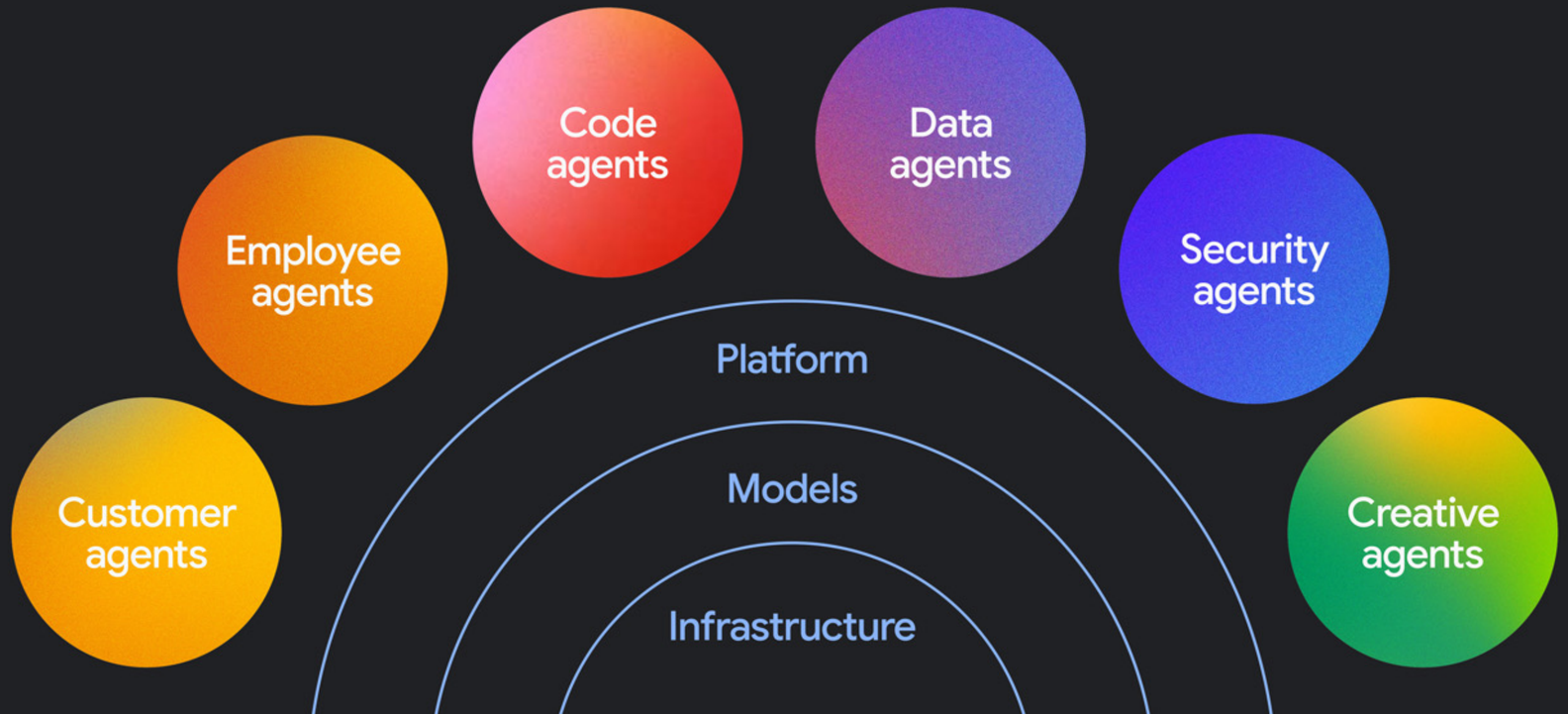
Google Cloud endava

When implementing a new technology, the first step is to identify the business problems you would like to solve—and gen AI is no different.

Take the time to delve into the specific challenges impacting your enterprise, and then target the areas and use cases where gen AI is best positioned to make an impact.

This strategic and considered approach will pay dividends. After all, not all problems are AI problems and more critically, not all AI problems are gen AI problems. In some cases, you might find that you need to reexamine your AI technologies to best accomplish your end goals.

Which use cases fit your org best?



Your organization will best benefit from identifying ways where AI can automate and streamline daily workflows, such as:

- Streamlining workflows with unstructured data
- Synthesizing data into more digestible formats
- Creating content for creative-based tasks

AI agents—intelligent systems that go beyond simple chat and predictions—can proactively take actions to help achieve specific goals, whether that's guiding a shopper to the perfect pair of shoes, helping an employee look for the right health benefits, or supporting nursing staff with smoother patient hand-offs during shift changes. We see AI agents center around six use cases: customer agents, employee agents, code agents, data agents, security agents, creative agents.

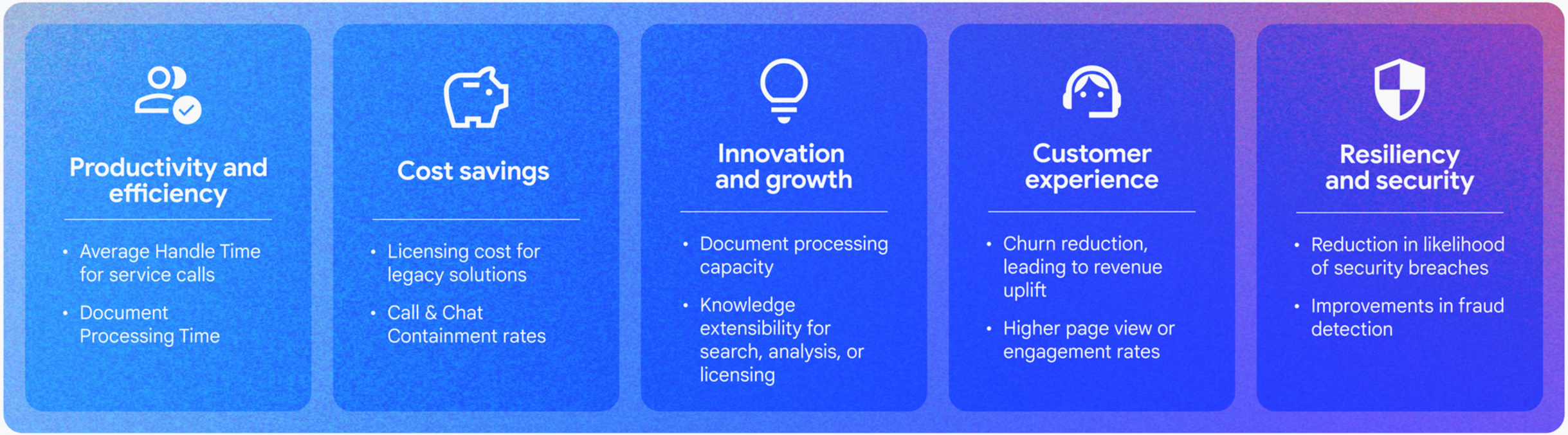
In fact, we've seen 185 real-world gen AI use cases from the world's leading organizations based on these six use cases.

To help you identify your own use cases, use our interactive Gen AI Navigator for personalized recommendations. Simply answer a few quick questions under each of our key areas: strategy, infrastructure, and skills.

Google Cloud endava

How will you measure outcomes?

Business value KPIs measure the overall impact of your AI initiatives on your organization.
Here are a few examples of key metrics to demonstrate the value of AI and justify continued investment.



For a complete ROI of gen AI, factor in the cost of building and maintaining these gen AI applications.

Google Cloud endava

Major cost drivers

	Potential impact on overall cost	Cost based on
Usage volume	★★★★	Number of users and number of uses
Data size and complexity	★★★	Amount of data you need to organize, index, and store
Gen AI model	★★★	Size and number of your gen AI models, plus model costs
Development and maintenance	★★	Ongoing resource allocations

Considerations for cost optimization:

- Acknowledge that different use cases will have different cost profiles, and plan accordingly. For example internal use cases such as knowledge search will likely incur much lower cost than external customer-facing applications.
- Optimization is crucial. Like any tool, the most effective gen AI applications are optimized with your organization's knowledge and data.
- Leverage the model(s) that fits your performance, latency, and financial needs. Some models are smaller, faster, and more cost effective; while others are larger and more costly, but can do more sophisticated tasks.

You may also like

Measuring the success of your generative AI: a deep dive into KPIs

Read more

Step 1: Choose the right models

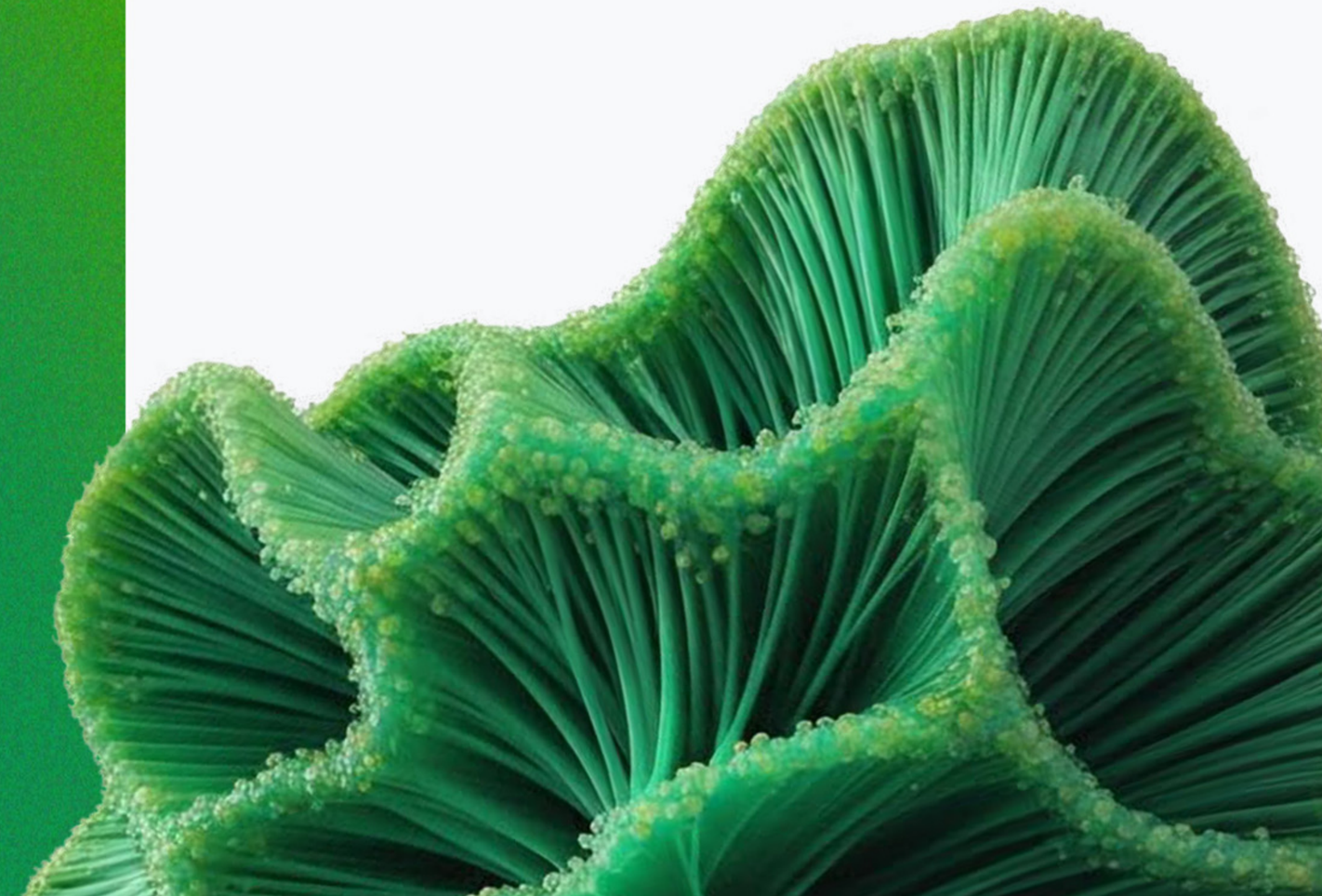
Your model makes all the difference.

Google Cloud endava

Once you've identified the business outcome use cases that are top priorities, you'll need to select a model provider.

> At Google Cloud, our objective is to bring you the best models suited for enterprise use cases, by pushing the boundaries across performance, latency, and costs.
>
> **-Saurabh Tiwary**
> **VP & General Manager,**
> **Google Cloud AI**

Your use case is unique, and so are gen AI models—you'll need to weigh the strengths and limitations of different models against your needs.



How do you decide which model is right for you?

Here are some key considerations to help you hone in on specific models for testing and evaluation.

Governance

Industry-specific constraints will impact the type of model you need. Healthcare, finance, and government often have stringent requirements for data privacy, security, and explainability. This might necessitate using models that have the right type of certification, open models that allow for higher levels of transparency and customization, or even models that can be run on isolated networks and infrastructure.

Use case

What tasks must the model perform for your use case? How complex are these tasks? Does the desired output need to be in a particular format or style?

Performance

What factors are most important—latency, cost, or customizability? Different models have different strengths, and you can use your enterprise and use case priorities as a filtering mechanism.

Model capabilities

To effectively achieve your use case and performance goals, consider the following capabilities to select the right model for your specific needs:

Context window: The amount of information the model can process at once.

Number of parameters: A measure of the model's complexity and learning capacity.

Training dataset: The data the model has learned from, which influences its knowledge and capabilities.

Multimodality: The ability to process and generate different types of data (e.g., text, images, audio).

Selecting a model isn’t a one-and-done exercise.

Once you enter the evaluation step, you may find that you need to update your model. Our recommendation is to start with a large model first. Doing so makes it easier to build a quality and safe application—versus smaller or open models which might require more guardrails to get started. As your business needs evolve—perhaps requiring lower latency, reduced costs, or enhanced quality in a specific domain—you can then strategically transition to smaller models to improve behavior along the criteria that you need. When you start with a large model, you can see those business needs arise and you will know what to do next.

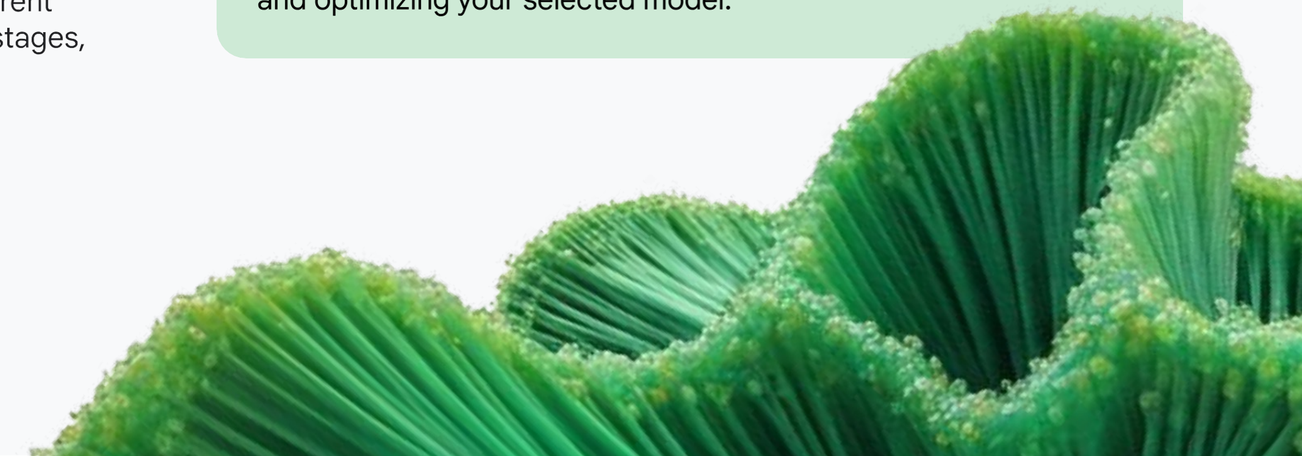
Models update often, offering performance and cost benefits. And so you may need to update your model version or choice of models accordingly, potentially on a regular basis.

However, migrating to a new model isn't easy; it requires restarting the process of prompt engineering and evaluation.

Some generative AI use cases employ multiple models to improve performance and cost. For instance, you might want to utilize a smaller, faster model for simple queries but direct to a domain specific or large model for more specific or complex queries. Additionally, many real-world applications involve intricate workflows with multiple stages. Different models can be assigned to specific stages, optimizing each step of the process.

How Vertex AI can help: Vertex AI offers a curated set of 160+ foundation models to fit different enterprise needs and budgets, including Google, open source, and third-party models. These models cover multiple modalities and sizes, with specialized options for specific industries.

To familiarize yourself with a model’s capabilities and understand its inputs and outputs, you can experiment using Vertex AI Studio. This experimentation will help you select “good enough” models as you approach the next step; building your evaluation set, testing and optimizing your selected model.



Step 2: Start with evaluation

You can’t improve what you don’t measure.

Google Cloud endava

After selecting the initial model for your use case, establish your criteria for evaluating model behavior to get a clear sense of how your model is performing off-the-shelf.

AI Studios are great for playing around and testing ideas, but when it comes to making progress toward production, you need a more tailored approach.

Your evaluation criteria will determine how you'll check the quality and safety of your models, as well as their performance.

Establishing evaluation criteria requires figuring out your evaluation rules and gathering example datasets. These datasets serve as your reference point, and will help you test how well your models are doing.

We'll show you how to get started creating your evaluation criteria.

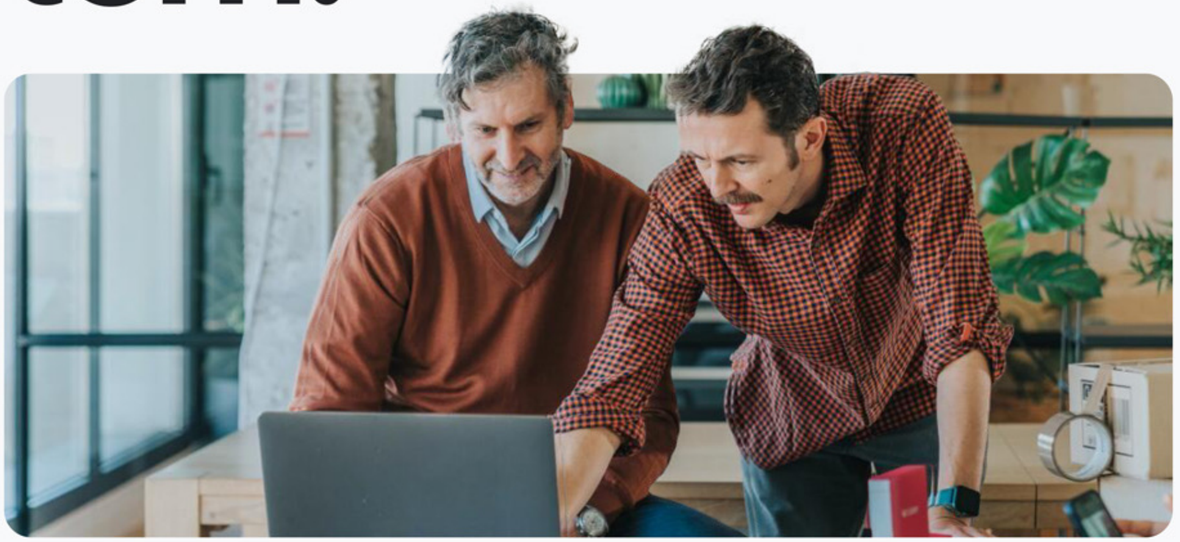
Google Cloud endava

Evaluation is the heart of your system.

With AI, we suggest moving from “Test driven development” to “Metrics driven development.”

Foundation models are non-deterministic systems, meaning their outputs can vary significantly even when given the same input. Whereas a deterministic system always gives the same answer to a query (such as a calculator giving the same answer to an addition problem), a non-deterministic foundation model might produce different responses to the same text-based prompt on different occasions. If you can’t measure improvement, then how will you know if you’ve improved?

Before you introduce change into your system, identify the metric and how you can measure it. And repeat for each step of the development process.



Model evaluation was critical to our production workflow, where manual checks and refinements weren’t feasible. Leveraging AutoSxS in Vertex AI, we established a generative AI evaluation framework that is simple, easy to set up, and scalable. This out-of-the-box service significantly improved our evaluation capabilities, leading to faster time-to-market and increased overall efficiency.

–Stefano Frigerio,
Head of Analytics Technical Leads,
Generali Italia

Google Cloud endava

Four tips for approaching evaluation

1. Prioritize creating a strong test set.

A **test set** is a collection of input prompts or questions with baseline responses where possible (also called labeled or ground truth data) used to assess the performance of a generative model. Test sets are important because they provide a standardized way to measure how well a generative model performs on different tasks and types of input.

A good test set should be:

- Representative: Cover a wide range of possible inputs and scenarios that the model might encounter in real-world use
- Unbiased: Not favor any particular type of response or style
- Large enough: Contain enough examples to provide a reliable estimate of the model's performance

Test sets should be high quality, up-to-date, and accurately reflect your specific business context. It is crucial that this data is created by individuals with in-depth business knowledge.

If you want to scale the creation of your test set, gen AI can help, too. Use your favorite tool to rewrite your evaluation data set with different vocabulary to get the quantity you need from a very small set of high quality and novel prompts.

2. Be task specific with evaluation criteria.

Summarization, Q&A, and content generation use cases will each require different evaluation metrics. On top of that, the criteria for your task will not be the same across use cases and companies. Adjust accordingly for best results.

How Vertex AI can help:
Vertex AI's evaluation service provides a template for you to define your own evaluation metrics. Also, it offers a library of pre-built templates for tasks such as multi-turn chat quality, summarization quality, or instruction following.

3. Use different methods for evaluation.

Computational: This is where you compare the model's generation to your ground truth, statistically. This method can help you understand if the model's output is similar to your ideal result.

> However, research shows that optimizing for a higher score in a computational metric, does not increase alignment with human preference. Generated results may differ from the ideal result (scoring low in this metric), but may still be quite good overall. That's because humans are better at detecting ambiguity in natural language and also that humans assign high scores to two results that are quite different.

Autoraters: Autoraters are AI models designed to automatically evaluate the quality of outputs of gen AI, assessing text, code, images, or even music. To ensure evaluation is aligned with business goals, consider adding custom metrics into AI autoraters.

Humans: Keeping humans in the loop, and ensuring evaluations are understandable and explainable, will remain important, even as AI improves. This includes escalation to a human when there is low confidence in a result output.

4. Start simple, add complexity as needed.

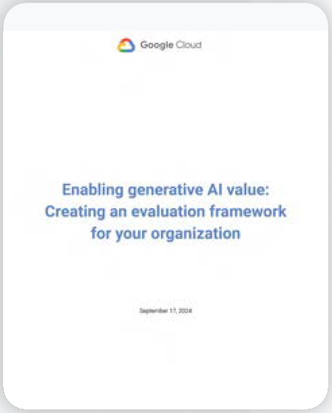
Due to their non-deterministic nature, the unpredictability of foundation models is amplified when multiple models or generations are chained together as agents, particularly if temperature parameters are introduced across various models. Temperature controls the randomness of the model's output—a higher temperature makes the output more creative and unpredictable, while a lower temperature makes it more focused and deterministic. And so if temperature parameters are introduced across various chained models, results will be more unpredictable. Therefore, it is crucial to maintain simplicity and clarity in your system design, ensuring that any changes made result in measurable and predictable progress.

Some systems have multiple predictive or gen AI models to working together on a task, it's crucial to evaluate both individual model performance and the overall effectiveness of the combined system.

You may also like

Enabling generative AI value: Creating an evaluation framework for your organization

Read more



Step 3: Improve model behavior

Continuous model improvement delivers results.

Google Cloud endava

Use your evaluation results to see where your AI model needs improvement in quality, latency, and cost.

There are two primary ways to improve model behavior: **customization and augmentation.** Improving latency and cost requires changing your model and revisiting model selection.

Customization

Customization focuses on adapting the model to better suit a specific task, domain, or style—modifying the model's internal workings to make it more specialized.

Some types of customization are easier than you might realize. For example, you only need 100 to 500 examples to use supervised fine-tuning with Gemini. If you can build an evaluation set, you can build a tuning set, but keep it simple—only tune as a final step.

How Vertex AI can help

Vertex AI offers multiple ways to customize a model, including:

Vertex AI Prompt Optimizer helps you avoid some of the tedious trial-and-error of prompt engineering. Based on Google Research on automatic prompt optimization methods, it helps you find the best prompt using any preferred model on Vertex AI.

Supervised fine-tuning (SFT) adapts model behavior with a labeled dataset, adjusting the model's weights to minimize the difference between its predictions and the actual labels. SFT enables you to tune the model to be more precise for your enterprise task. It's particularly effective for domain-specific applications where the language or content significantly differs from the data the large model was originally trained on.

Distillation for Gemini enables you to train smaller, specialized models that inherit the knowledge of the larger Gemini model. Deploying foundation models can be a resource-intensive challenge. With distillation techniques in Vertex AI, you can leverage the power of those large models while keeping your deployments lean and efficient, often achieving comparable performance with the flexibility of self-hosting your custom model on Vertex AI.

Once you build up more evaluation and tuning data, you will be able to distill smaller models—creating a beneficial feedback loop. Start with a large model with less data, and eventually you can shift to more data on a smaller model that's either lower latency, lower cost, or better trained for your evaluation or tuning set.

Reinforcement learning with human feedback (RLHF) on Vertex AI is a technique used to fine-tune foundation models by incorporating human feedback into the training process. It's a way to make these models more helpful, accurate, and aligned with human preferences.

Intelligence Engine enables you to optimize performance and reduce costs by directing prompts to different models based on predefined business rules and user input.

Augmentation

Augmentation focuses on providing the model with data and tools when generating its output.

Methods include:

- **Chaining:** Breaking down complex prompts into smaller tasks and routing them to the most suitable models.
- **Reasoning loops:** Implementing iterative processes like “react and reflect” to enhance the model’s reasoning capabilities.
- **RAG and grounding:** Connecting the model to your knowledge base for more accurate and context-aware responses.
- **Tools:** Granting the model access to external tools like RAG or Search, your APIs, and the internet to enhance its capabilities.
- **Memory or session history:** When the AI system can retain and recall information from previous interactions within the same session, it can lead to more coherent and relevant responses.

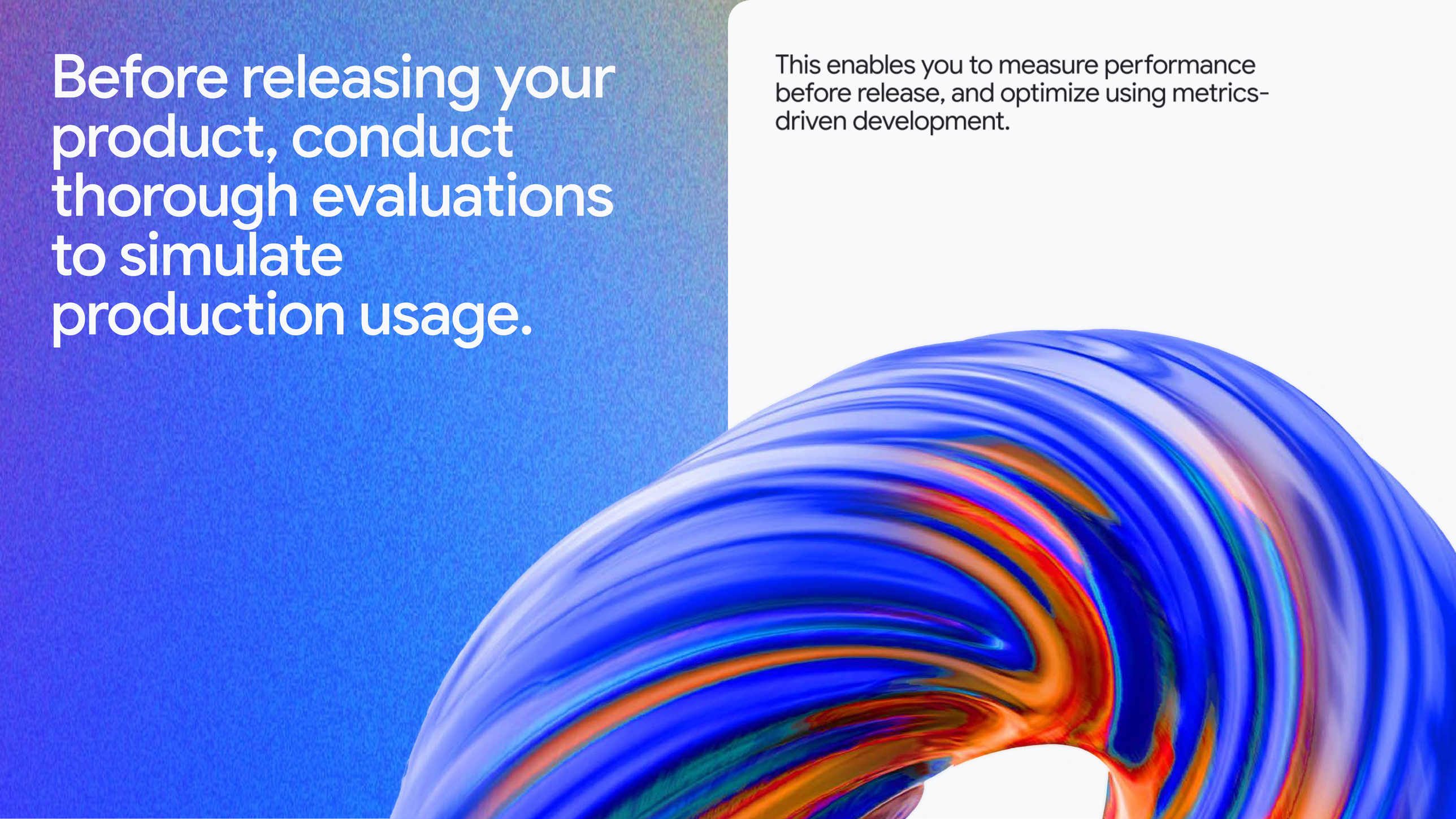
Step 4: Release, validate, and deploy

Go for launch.

Google Cloud endava

Before releasing your product, conduct thorough evaluations to simulate production usage.

This enables you to measure performance before release, and optimize using metrics-driven development.



The amount of validation you'll need to do before release depends on the complexity of your application or the customization you've implemented in your model.

If you're using a **model-as-a-service** to build a gen AI application (such as RAG) or an agent you can avoid packaging the model (below), but will still need to track versions of all the dependencies, version and release your configuration, prompts and code, and adhere to the metrics-driven development criteria we've already covered.

If you've **customized your model**, then releasing your gen AI system involves packaging the model and application, implementing version control, registering artifacts, pinning dependencies, and storing it in a model registry. Validation is crucial and includes evaluating performance, testing robustness, assessing bias and fairness, and conducting safety and security checks.

When it comes to deployment, this step requires selecting appropriate infrastructure, choosing a deployment strategy, setting up monitoring and logging, and implementing CI/CD pipelines for automation. It's also important to establish a feedback loop for continuous improvement.

As your gen AI application grows in popularity and usage, maintaining a consistent and responsive user experience becomes crucial. However, gen AI applications can sometimes face performance issues due to high demand of models running on shared computational resources. This can lead to delays and service disruptions, impacting the user experience.

Throughout theis process of releasing, validating, and deploying your generative AI application, ethical considerations, legal compliance, and clear documentation are essential. Releasing a generative AI system is an iterative journey that demands ongoing monitoring, evaluation, and refinement to ensure responsible and successful deployment.

Mitigating key risks

It's imperative to ensure your overall system safety and mitigate potential risks that are unique to gen AI. Here are several model vulnerabilities you should be aware of:

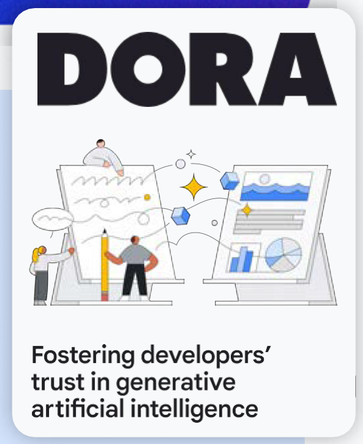
1. Recitation: Addressing potential copyright issues arising from the model's training data and generated content
2. Hallucination: Managing and reducing instances where the model generates outputs not grounded in factual reality
3. Jailbreaking and prompt injection: Protecting against unauthorized manipulation of the system's behavior through overrides or malicious prompts
4. Training data poisoning: Safeguarding the model's training data from contamination that could lead to biased or harmful outputs

Collaborating with a major vendor who has already addressed these considerations can provide a valuable foundation to build upon.

You may also like

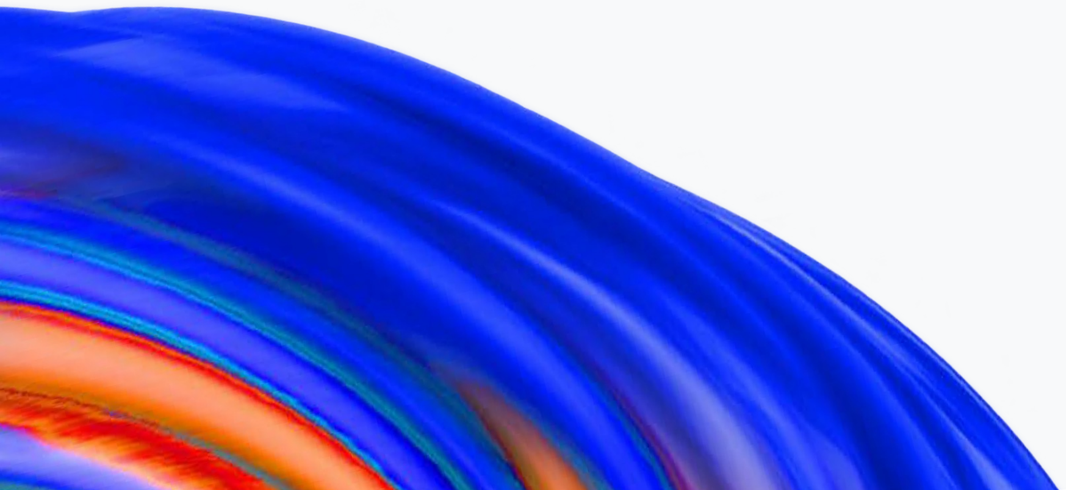
Fostering developers' trust in generative artificial intelligence

Read more



How Vertex AI can help

Vertex AI Provisioned Throughput provides reserved capacity for your gen AI models through monthly or weekly terms.



This ensures your applications receive prioritized access to processing power, guaranteeing a smooth and responsive experience, even during peak usage.

Consider the following factors when deciding if Provisioned Throughput is right for you:

Application type

Are you building a real-time application where immediate responses are critical, such as a chatbot or an AI-powered interactive assistant?

Traffic volume

Do you anticipate high volumes of requests or user interactions with your application?

Performance requirements

Does your application have strict performance targets or service level agreements (SLAs) that need to be met consistently?

Budget

Provisioned Throughput is a premium feature with a fixed cost. Evaluate if the benefits of guaranteed performance outweigh the cost for your specific needs.

Use the Provision Throughput Calculator in the Vertex AI console to estimate your requirements.

Google Cloud endava

Step 5: Continue to monitor, effectively

Monitoring and maintaining your AI for production.

Google Cloud endava

Conventional observability methods like logs, traces, and metrics remain crucial for gen AI.

Large models necessitate close monitoring of latency, block generations, error rates, and cost. Integrating these metrics into your system is essential for effective management.

In addition to traditional observability, we recommend two solutions for monitoring task-specific generations.

1. Supervised monitoring along dimensions of concern

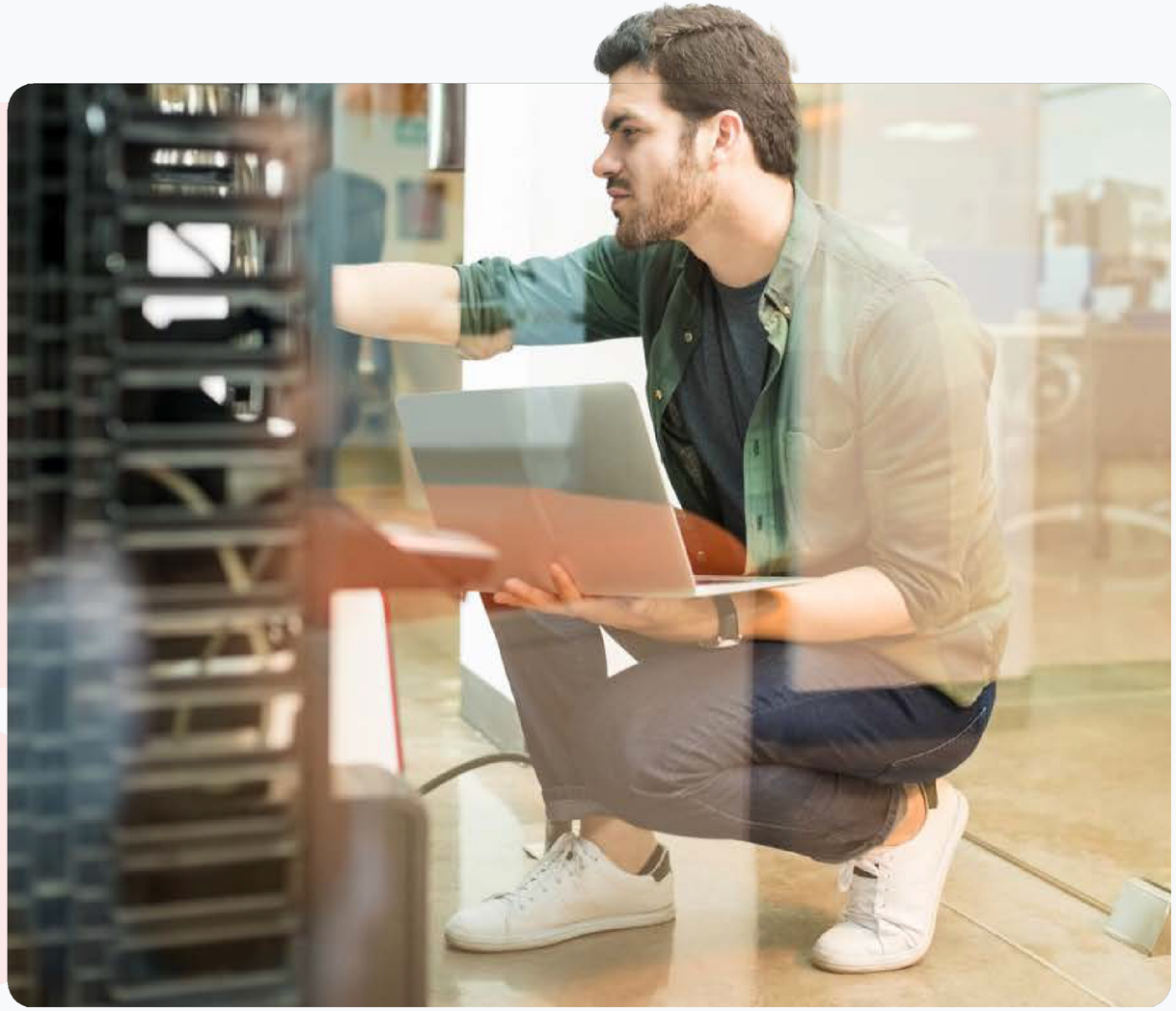
Identify the dimensions you care about (e.g. quality and safety scoring) and evaluate them continuously. However, it's not always feasible to rely on metrics and eliminate human involvement. In such scenarios, incorporating human-in-the-loop or LLM evaluators becomes necessary.

Tip: Classifying prompts by task and topic can help you resolve any areas of poor performance. While large models will not regress from your tuning corpus, your model may perform poorly on a specific task if not all of your data is represented in your evaluation corpus.

An evaluation service can help identify trends and highlight how patterns or prompts might evolve, which may necessitate reevaluating performance.

2. Unsupervised techniques for dimensionality reduction

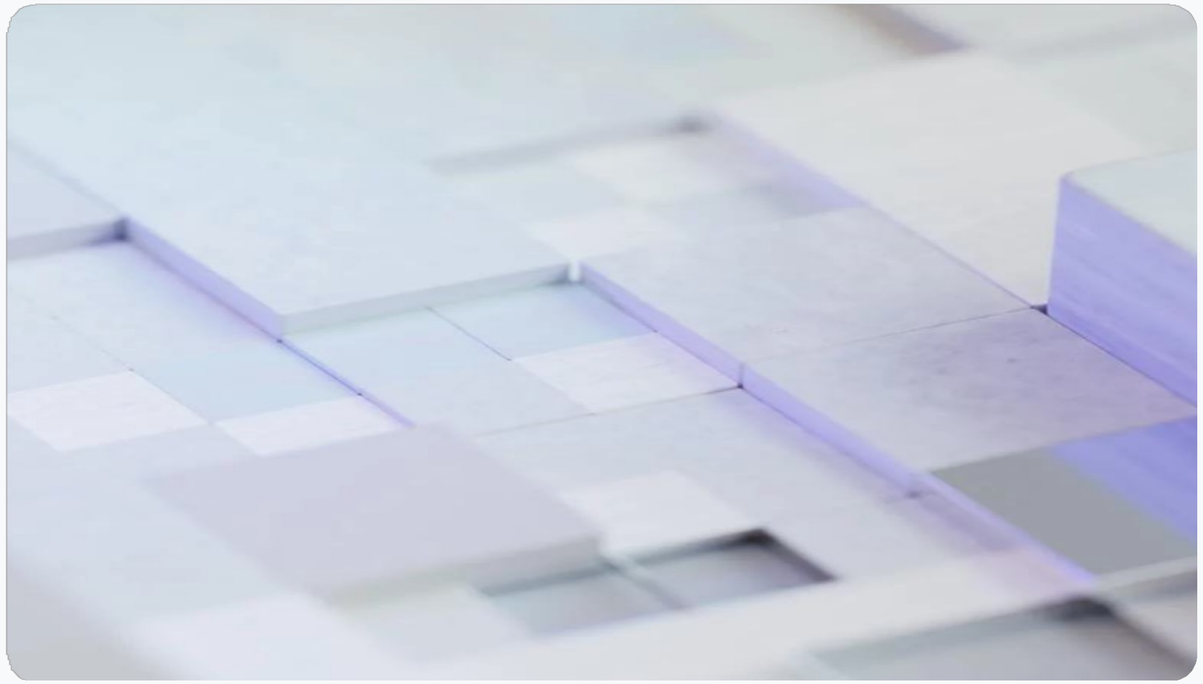
By converting generated outputs into an embedding space, you can cluster embeddings to identify outlier prompts—clusters should represent common issues that enable you to identify key areas of concern, such as latency, safety, or quality.



Double click: Governance, safety, and responsible AI.

AI's governance, safety, fairness, and effect on equitable economic opportunities are not a step along the path of prototype to production—but rather practices that need to be constantly upheld across model providers and your organization.

Google Cloud endava



Governance

To establish effective governance, prioritize the implementation of infrastructure-as-code and CI/CD deployment pipelines. These pipelines should simplify the creation of sandboxes and streamline deployment to testing and production environments. By automating the validation process, you empower Data Scientists and App Developers to focus on their core tasks, facilitating a continuous and efficient path to production.

Frameworks for inspiration include ISO 42001, NIST AI-RMF, MAS Veritas.

The legal landscape is a crucial consideration—legislation such as GDPR/CCPA and the EU AI Act need to be carefully considered while prototyping and moving to production.

Securing AI systems

There are new and unique new security risks that come with AI systems. Take care to reduce the risks of adversarial examples and prompt injections. These risks need to be considered during the development and use of AI systems.

To succeed, security teams should build on the foundation of traditional application security, data security, and system security, and add to the mix their new knowledge of AI use cases, AI threat, and AI-specific safeguards.

- Check out Securing AI: Similar or Different to learn more about the similarities and differences between securing traditional enterprise software systems and AI systems.
- Learn how to implement the Secure AI Framework (SAIF) with this quick guide.

Google Cloud endava

Responsible AI (RAI)

At Google, we incorporate responsible practices for fairness, safety, privacy, and transparency throughout the product development lifecycle. We offer AI products and capabilities with a deep sense of responsibility and the highest standards of information integrity with our AI principles.

In addition, we offer resources, tooling, and support to empower enterprises to build and use gen AI responsibly.

Product and use case reviews are designed to identify, assess, and mitigate potential impacts before they are generally available.

Education, research, and best practices help enterprises navigate responsible AI including Model Cards for our models, widely used by developers, journalists, and industry analysts to explain complex technology to general audiences.

RAI tooling, enablement, and support help enterprises identify, assess, and mitigate potential impacts within their use case and application(s).

- **Data prep and exploration:** Tools like Data Cards, Know Your Data, and Fairness Indicators are Google open sources techniques to help you better understand and curate your data for AI tasks.
- **Content moderation and safety:** The Moderate Text API enables you to detect and scan potentially offensive and harmful content, and works with open source and third-party models.
- **Model safety and system instructions:** Google's generative AI models, like Gemini 1.5 Flash and Gemini 1.5 Pro, are designed to prioritize safety by default, and enable you to define the right level of safety for your use cases.
- **Citation filtering:** Google's citation filtering will either cite or block known sources of copyright and IP to minimize the risk of infringement, so you can tackle copyrighted materials.
- **Explainability and bias tooling:** Tools such as feature attribution, example-based explanations, and advanced model probing allow you to explain and understand how and why predictive AI models came to a decision.
- **Image-based tooling for productionizing responsible AI:** Use built-in safety precautions to ensure generated images align with RAI principles with digital watermarking across image pixels, and identify AI-generated images via SynthID.

Take your project from prototype to production with Vertex AI.

What used to be long, drawn-out processes—building, tuning, and training models – are now significantly faster thanks to Vertex AI and Gemini. We're able to iterate and deploy at scale much more quickly.

–Niraj Nagrani
VP and GM for Consumer and Supplier Technologies, Wayfair

Google Cloud's gen AI capabilities empower you to manage the entire AI lifecycle, from initial concept to full-scale production, regardless of your specific use case, level of expertise, or operating environment.

Our comprehensive suite of tools includes:

- **AI infrastructure:** Our cutting-edge AI Hypercomputer provides the foundation for your AI endeavors
- **Vertex AI Model Garden:** Access a vast library of over 150 models from Google, partners, and the open-source community
- **Vertex AI Model Builder:** Build and customize your own models with a full range of tools to support your journey from prototype to production
- **Vertex AI Agent Builder:** Create and tailor AI agents for any skill level, featuring grounding capabilities essential for enterprise applications

These capabilities are all built with enterprise-grade security, compliance, and responsible AI practices in mind.

Google Cloud endava

AI for your enterprise

An end-to-end platform that unlocks your data for every use case, expertise, or environment.

Vertex AI

Google Cloud endava

Gen AI evaluation-driven development on Vertex AI

Discover | Ideate & Implement | Improve

1 Choose the right model

Choose the right model

Access industry leading models with Vertex AI Model Garden, which provides **160+ curated models** including Google, third party, and open models.

For **thousands of additional models**, Hugging Face integration provides one-click model deployment to Vertex AI.

Create your test set, the reference point for evaluating models.

2 Start with evaluation

Curate data

Manage and improve prompts with Vertex AI Studio, Prompt Optimizer and the prompt management SDK.

Organize and serve features with Vertex AI Feature Store and BigQuery.

Create a robust and scalable solution for data curation with Google Cloud Storage (GCS) and Dataplex. GCS provides a foundation for your data lake while Dataplex helps you discover, manage, and secure that data.

Predict: Generate a large and diverse set of samples from your model(s) with Vertex AI Batch Prediction.

↓

Evaluate and explain: Assess model behavior with Vertex AI Evaluation & Explainability.

↓

Compare: Track and compare experiments with Vertex AI Experiments. Iterate on model development with Tensorboard.

↓

Decide and do: Determine if and how models need to be improved with customization or augmentation techniques.

3 Improve model behavior

Customize models with Vertex AI Model Builder

Tune foundation models with SFT and RLHF.

Improve performance and reduce cost with Distilation.

Improve data quality and enable efficient model deployment with Feature Transform Engine and Controlled Parameter Reduction.

Automate processes with Vertex AI Pipelines.

Efficiently customize complex models with massive datasets with Ray on Vertex.

Train predictive models with custom training, AutoML, and Hyperparameter Tuning, also available on Vertex AI Model Builder.

Create multiple **Model Versions**

Augment models with Vertex Agent Builder

Orchestrate interactions between your models, tools, and users with LangChain on Vertex AI.

Grounding on Vertex AI offers varying degrees of simplicity and customization, from user-friendly to highly customizable solutions. Select the best fit for your requirement with Vertex Search, RAG API, Feature Store, and Vector Search. Ground on your own data or use Grounding with Google Search.

Connect models to the real world and expand their capabilities with Vertex AI Extensions and Function Calling.

Create AI agents using natural language or a code-first approach

4 Release, validate, and deploy

Build, test, and deploy on Cloud Build, a serverless CI/CD platform.

Deliver continuously with custom targets using to Cloud Deploy.

Orchestrate and automate workflows with Vertex AI Pipelines.

Prepare data that fuels your models, ensuring quality, consistency, and reliability with BigQuery and Dataform.

5 Serve and monitor

Deploy models for online serving with Vertex AI Endpoints, enabling low-latency and scalable access to your gen AI capabilities.

Ensure responsible use of open models with the Moderate Text API which automatically detects and flags inappropriate or harmful content generated by your models.

Traditional observability like logging, monitoring, and tracking is available via the Cloud Operations Suite.

Monitor the performance and health of your gen AI models with Vertex AI Model Monitoring.

Inference Logs are a goldmine of information for improving your gen AI systems: understand model behavior, diagnose issues, and ensure responsible AI.

Govern | Dataplex Catalog and Lineage, Cloud Build repositories, Artifact Registry, IAM, VPC-SC, CMEK, Org Policies, Resource Manager, Cloud Billing, Cloud Assist Inventory and Audit Logs, Vertex ML, Metadata

Google Cloud endava

Ready to get started?

Contact us