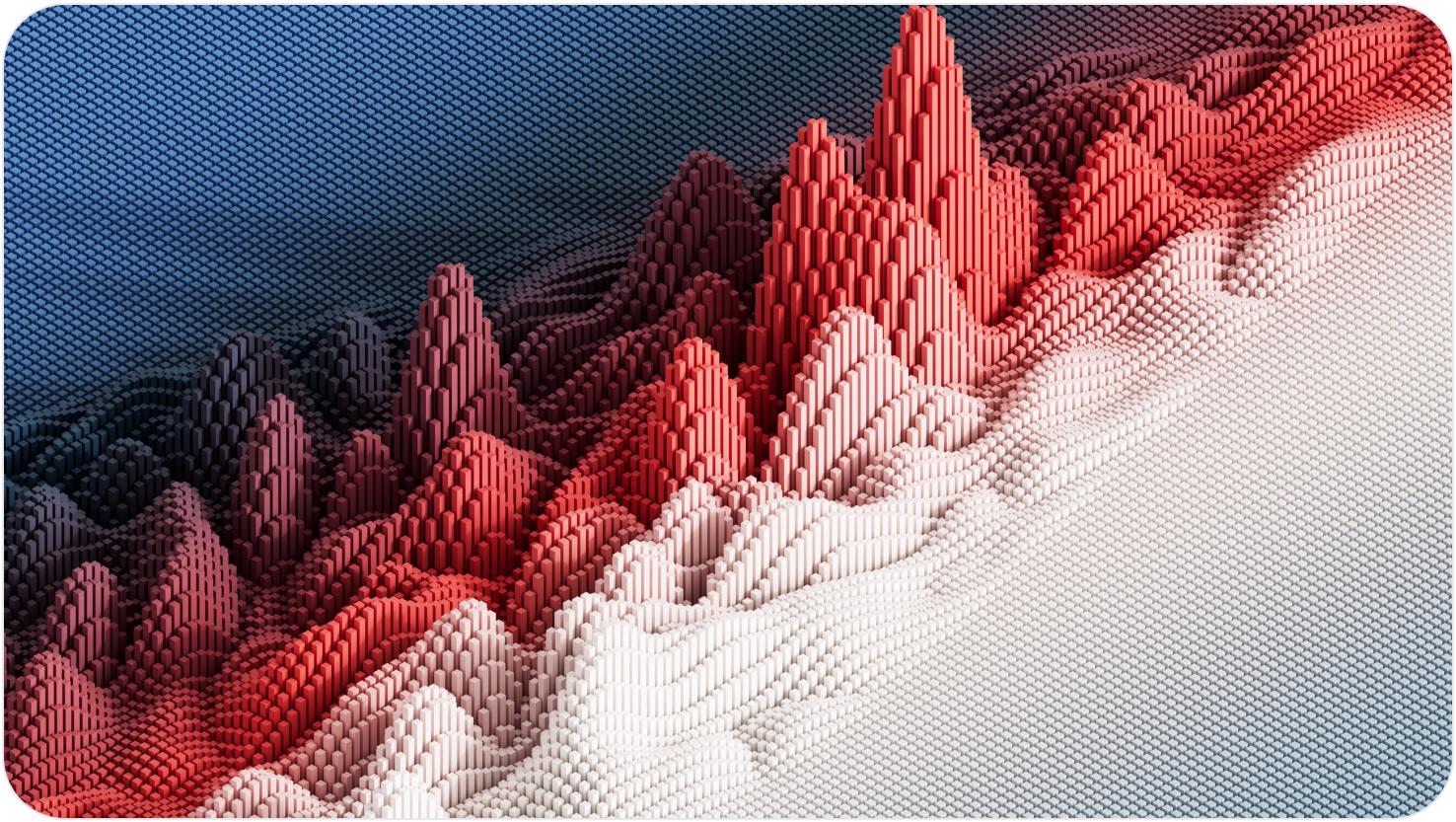
AI & Machine Learning

Measuring gen AI success: A deep dive into the KPIs you need

November 25, 2024



Hussain Chinoy
Technical Solutions Manager, Applied AI Engineering

Amy Liu
Head of AI Solutions, Value Creation

Looking to measure the effectiveness of your gen AI? Here's the KPIs that'll help track model accuracy, operational efficiency, user engagement, and

As the saying goes, “You can’t manage what you don’t measure.” Key performance indicators, or KPIs, are the bedrock of both business and technology success, providing a set of clear metrics you can use to track the progress your teams and projects are making. When adopting generative AI, KPIs remain critical for evaluating success, helping to objectively assess the performance of your AI models, align initiatives with business goals, enable data-driven adjustments, and demonstrate the overall value of an AI project.

We’ve previously discussed that gen AI requires developing a new set of AI metrics and approaches to measure the performance of your gen AI projects. Still, we find that most organizations are using the same computation-based model quality KPIs they relied on for tracking other types of AI technologies, unaware of the other crucial metrics they need related to system performance and adoption. They also don’t spend enough time measuring business value, often confusing operational efficiency gains with end state goals.

In this post, we’ll dive deeper into the KPIs that are essential for measuring the effectiveness of gen AI, including all the specific metrics and how to use them to make the most of your investments.

Model quality KPIs

Model quality metrics are crucial for understanding the accuracy and effectiveness of your AI model's output. Computation-based model metrics are very effective

parameters. If you were to evaluate a product search AI model, these metrics provide a straightforward way to compare outputs against a reference dataset. For instance, precision measures how relevant the products surfaced are to the search query, recall measures all the relevant products and how many were captured by the model, and F1 score provides a balanced average score between them.

Gen AI's unique ability to generate a wide variety of unbounded outputs — original, unexpected, and, in some cases, even harmful content — requires more subjective evaluation. Model-based metrics use auto-raters — large language models designed for evaluation — as a judge to assess performance based on descriptive evaluation criteria to assess creativity, accuracy, and relevancy. Judge models can automatically analyze a model's responses based on an evaluation template to measure the quality of outputs.

There are two common methods to produce quantitative metrics about a model's performance:

1. **Pointwise metrics** assess the candidate model's output based on the evaluation criteria. These metrics work well in cases where it's not difficult to define a scoring rubric. For example, the score could be 0-5, where 0 means the response does not fit the criteria, while 5 means the response fits the criteria well.
2. **Pairwise metrics** involve comparing the responses of two models and picking the better one to create a win rate. This is often used when comparing a candidate model with the baseline model. These metrics work well in cases where it's difficult to define a scoring rubric and preference is sufficient for evaluation.

The ROI of gen AI

Google Cloud surveyed more than 2,500 business and technology leaders to discover where and how they are seeing the biggest returns on their AI investments. One thing is clear: Those who aren't deploying AI today risk falling behind tomorrow.

Read it now.

Currently, model-based metrics are still somewhat experimental, and many organizations use human evaluators to manually evaluate models. However, model-based evaluation with autoraters, calibrated with human raters to ensure quality, is an area worth following closely. As they allow for richer evaluation criteria, including:

- **Coherence:** Measures the model's ability to generate a coherent response based on the prompt.
- **Fluency:** Measures the model's language mastery based on the prompt.
- **Safety:** Measures the level of harmlessness in a response.
- **Groundedness:** Measures the ability to provide or reference information included only in the prompt.

- **Verbosity:** Measures a model's conciseness and its ability to provide sufficient detail without being too wordy or brief.
- **Text quality:** Measures how well a model's responses convey clear, accurate, and engaging information that directly addresses the prompt.
- **Summarization quality:** Measures the overall ability of a model to summarize text.

When to use them: Model-based metrics are a good fit when evaluating models that produce unstructured and unbounded outputs, such as long-form text, complex code, or images, which make comparison to a reference dataset more difficult. By tracking these diverse metrics, you gain a comprehensive understanding of your model's strengths and weaknesses, enabling you to make targeted improvements and ensure high-quality output.

System quality KPIs

As we've written before, you'll need to invest in an end-to-end AI platform to harness the full potential of gen AI and enable your organization. These platforms will need to seamlessly integrate all the key components needed to develop, fine-tune, deploy, and manage models at scale while also measuring the performance of various aspects of an AI system, including deployments, responsiveness, and resource utilization.

System metrics focus on the operational aspects of your AI system, ensuring it runs efficiently, reliably, and at scale to support the needs across your organization. They offer

Deployment metrics

Tracking how many pipelines and model artifacts are deployed provides insights into your AI platform's capacity, governance, and organization wide impact. Here are some of the most commonly used metrics:

- **Number of deployed models**: This metric measures the number of models currently serving predictions to users or applications. This metric may indicate if you are a builder or a buyer when it comes to AI.
- **Model time to deployment:** This metric measures the average time it takes to deploy a new model or update an existing one, measuring the velocity of your deployment processes. This metric can help highlight bottlenecks in your deployment pipeline.
- **Percentage of automated pipelines**: This measures the percentage of automated workflows throughout the entire lifecycle of your AI models. This metric helps understand how much manual effort is required and where to invest in automation.
- **Percentage of models with monitoring**: This measures the number of deployed models actively monitored for changes in data distribution or model performance degradation. This metric is essential for maintaining model effectiveness over time.

Reliability and responsiveness metrics

Tracking how quickly your AI platform responds to requests is critical for user experience and maintaining model and application performance. Here are some of the most commonly used metrics:

- **Uptime**: The percentage of time your system is available and operational. A higher uptime indicates greater reliability and availability.
- **Error Rate**: The percentage of requests that result in errors. Understanding error types can provide valuable insights into underlying system issues, such as quotas, capacity, data validation, user input error, and more.
- **Model Latency**: The time it takes for your gen AI model to process a request and generate a response. This metric can identify subpar user experiences and signal needs for hardware upgrade
- **Retrieval Latency**: The time it takes for your system to process a request, retrieve additional data from applications, and return a response. Optimizing retrieval latency is crucial for applications that rely on real-time data.

Throughput and Utilization

Tracking throughput and resource utilization can reveal the processing capacity of your system. These metrics can help you optimize performance, manage costs, and allocate resources more effectively. Here are some of the most commonly used metrics:

need for burst capacity to accommodate high request volumes and minimize HTTP 429 errors (Too Many Requests).

- **Token throughput:** The volume of tokens an AI platform per unit of time. As foundation models introduce new modalities and larger context windows, this metric is critical for ensuring appropriate sizing and usage.
- **Serving nodes:** The number of infrastructure nodes or instances handling incoming requests. This metric helps monitor capacity and ensure adequate resources are available to meet demand for steady state and peak time periods.
- **GPU/TPU accelerator utilization:** Measures the percentage of time specialized hardware accelerators like GPUs and TPUs are actively engaged in processing data. As AI infrastructure usage grows, this metric is crucial for identifying bottlenecks, optimizing resource allocation, and controlling costs.

When to use them: The metrics you choose will depend on both your AI platform and the type of gen AI model you choose. For example, using proprietary models like Gemini means metrics, such as required serving nodes, target latency and accelerator utilization, are handled for you by a Google managed service so you can focus on building your application via simple APIs. If using open models and hosting it yourself, you may need to incorporate a wider range of system quality metrics to help you identify bottlenecks and optimize your AI system's performance.

Operational metrics measure the impact of your AI system on your business processes and outcomes. These metrics will differ by solution and industry. When you make changes to AI systems, there is not always one direction of movement; improving one KPI can sometimes impact another. For instance, for retailers, cart size may increase with a more engaging chatbot, but time-to-cart, a previously important metric to keep low, may increase. Therefore, context and industry expertise is critical when interpreting any changes.

Here are some examples of business operational metrics used for tracking impact across the most common gen AI use cases in different industries.

Customer service modernization

In nearly every industry, enterprises are adopting gen AI to modernize their customer service for its ability to enhance personalized experiences, and boost employee productivity. The primary industries adopting these use cases include telecommunications, travel, financial services, and healthcare. Here are some common metric examples:

- **Call and chat containment rates:** Measures how many incoming calls or chat interactions were handled and resolved by AI solutions. This metric signals your organization's capacity for using AI automation to deflect inbound inquiries, manage future demand, and scale efficiently.
- **Average handle time:** The amount of time both human and AI agents spend to resolve a customer inquiry. This metric is often used when implementing

- **Customer churn and customer satisfaction score (CSAT):** Measures the rate at which customers stop doing business and how happy your customers are with products and services. There's often a strong inverse correlation between these two metrics — as satisfaction increases, churn often decreases.
- **Human agent churn and satisfaction:** Measures the rate at which human agents leave a company and how happy they are with their jobs. Identifying factors that contribute to agent satisfaction can help to reduce agent churn, leading to higher retention and reducing the cost of hiring and onboarding new agents.

Product, service, and content discovery

Gen AI can significantly enhance the ability for customers to discover new products, services, or content based on personalized recommendations and more intuitive AI-powered search experiences. The primary industries adopting these use cases include retail, quick service restaurant (QSR), and travel. Here are some common metric examples:

- **Click-through rate (CTR):** Measures how many users click on a product, service, or content after seeing it. This metric indicates the relevance of search results or surfaced recommendations.
- **Time on site (TOS):** Measures the length of time a customer spends on a website or application. This metric indicates user engagement and satisfaction in different ways. For example, longer TOS in page views or watch time reflects elevated engagement rates for media customers. Lower TOS on product or search

- **Revenue per visit (RPV):** Measures the total amount of revenue for a specific period per unique visitors. This metric indicates how effective monetization is at converting customers for each site or app visit. Other metrics like click-through-rate (CTR), add-to-cart, conversion rate, and average cart size can also affect RPV.
- **Visit volume:** Measures the total number of unique users visiting the site or engaging with an application. This metric can be a broad indicator for customer experience, satisfaction, or net promoter score (NPS). The amount of traffic can reflect overall business growth, target audience research, and marketing campaign success.

Intelligent document understanding and processing

Leveraging gen AI, many industries are also accelerating their ability to extract data from unstructured documents, such as PDFs, invoices, contracts, reports, forms, and more. The primary industries adopting these use cases include financial services, healthcare, and manufacturing. Here are some common metric examples:

- **Processing time:** Measures the amount of time it takes to process and extract data from a document, typically including steps such as quality assurance and validation.
- **Process capacity:** Measures the maximum amount of output a process can achieve, both under ideal conditions and considering factors like downtime or equipment failure. This metric indicates how effective a process is at handling high volumes of documents without increasing costs. A bottleneck in your process will typically set the limit for process capacity.

to meet other applications. The metric helps you understand whether the AI application can be used for more tasks, such as knowledge search, analysis, or data licensing.

When to use them: Business operational metrics can help you connect technical model quality with downstream financial impact, allowing you to understand whether or not your AI initiatives are generating tangible value for your business. Monitoring operational metrics requires close collaboration across multiple teams, with business stakeholders helping to interpret the results, data science teams optimizing model outputs, and developers ensuring that AI features function as expected.

Adoption KPIs

The broad potential of gen AI has introduced the need for a new set of adoption metrics for tracking its adoption and use across organizations. Unlike predictive AI technologies, which are often integrated directly into an application, gen AI success hinges heavily on changes in

engage with them. Similarly, AI-enabled employee productivity tools can only drive productivity gains if employees adopt them into their daily workflows.

Here are some of the most commonly used metrics:

- **Adoption rate:** The percentage of active users for a new AI application or tool. Movements in this metric can help identify if low adoption is caused by lack of awareness (consistently low), or lack of performance (high adoption dropping to low).
- **Frequency of use:** Measures how often many queries a user sends to a model on a daily, weekly, or month basis. This metric gives insight to the usefulness of an application and the types of usage.
- **Session length / Queries per session:** Measures the average duration of a user's interaction with an AI model, and may shed light on its entertainment value or effectiveness at retrieving answers.
- **Query length:** The average number of words or characters per query. This metric illustrates the amount of context users are submitting to generate an answer.
- **Thumbs up or thumbs down feedback:** Measures satisfaction and dissatisfaction of a customer's interaction. This metric can be leveraged as human feedback to refine future model responses and improve output quality.

When to use them: Adoption metrics provide insights into how your gen AI application is being used and its impact on user behavior, revealing whether people find it valuable and identify areas for improvement. In addition to these metrics, it's also helpful to supplement other approaches, such as surveys or focus groups to gain

provide an overall picture of a model's accessibility, reliability, and usability.

Business value KPIs

One of the biggest challenges facing executives and business leaders is proving the value of gen AI investments. More and more, there is a growing need for ways to quantify the impact of gen AI and demonstrate that it is delivering return on investment (ROI). Business value metrics complement your business operational and adoption metrics, helping to translate them into financial metrics that quantify the overall impact of your AI initiatives on your organization.

Some of the most common examples include:

- **Productivity value metrics:** Captures productivity enabled by AI by measuring concrete improvements, such as average call handling times, document processing times, and time saved with tools.
- **Cost savings metrics:** Illustrates IT and services efficiencies from AI applications by measuring legacy licensing costs against AI solutions, call and chat containment rates, and cost savings for hiring and onboarding.
- **Innovation and growth metrics:** Assesses the role of AI in driving new products, services, or business models by measuring document processing capacity, knowledge extensibility, and improvements in work, communication, or asset quality.
- **Customer experience metrics:** Captures the impact of AI on customer satisfaction and loyalty by

- **Resilience and security metrics:** Evaluates how well gen AI can withstand disruption and protect sensitive data by measuring application downtime or scalability, reduced security risks, and improvements in detection and response.

How to use them: Once you have developed a solid plan for understanding business operational and adoption KPIs, your finance team can help translate them into concrete financial impact metrics. You will also need to consider the costs of building and maintaining gen AI to gain a complete understanding of your ROI. This includes evaluating cost drivers, such as data size and complexity, usage volume, the number of models, model size and complexity, and all the relevant resources needed to develop and maintain applications. Gen AI costs can vary significantly, so it's critical to leverage models that best fit the performance, latency, and financial requirements of your use case.

Putting KPIs for gen AI to work

Let's explore how these metrics can be applied in a practical context by examining how you might use them to assess the utility and value of a real-world gen AI use case.

Imagine a consumer food delivery company introducing a new AI-powered chatbot to handle common customer support inquiries, such as help with food orders, billing and payment, and general subscription and service questions.

performance and effectiveness of your gen AI projects. Tracking the right metrics across model quality, system performance, operational efficiency, adoption, and business value can enable you to make smarter decisions and realize the full potential of gen AI in your organization.

In addition to the authors, Mikhail Chrestkha greatly contributed to this post.

Posted in AI & Machine Learning

Related articles

AI & Machine Learning

Building AI that benefits humanity: A conversation with Google DeepMind's leaders

By Chau Mai • 6-minute read

AI & Machine Learning

Data engine: McLaren's Lando Norris and Oscar Piastri on their F1 data and AI edge

By Matt A.V. Chaban • 6-minute read

AI & Machine Learning

The Prompt: What is long context — and why does it matter for your AI?

By Warren Barkley • 6-minute read

AI & Machine Learning

Ask OCTO: Closing the AI skills gap

By Will Grannis • 9-minute read