FEATURE

8 metrics to measure GenAI's performance and business value

When gauging the success of generative AI initiatives, metrics should be agreed upon upfront and focus on the performance of the model and the value it delivers.

By **George Lawton** Published: **21 Nov 2024**

Generative AI models are all the rage these days, so it's easy to get caught up in the hype and fall short of delivering real value. The only way to make these emerging models deliver more value is to measure what matters to your organization and improve on the results.

The trouble is that there are many ways to quantify generative AI initiatives and models, from performance and accuracy to precision, utility and ROI. Agreeing on the appropriate metrics upfront can mean the difference between a costly experiment and bottom-line value.

Enterprises should focus on two key types of metrics for generative AI, according to Christine Livingston, a managing director in the emerging technology practice at Protiviti. The first type is related to the performance of the model itself, such as response time, precision and accuracy.

These measures should also be considered in the context of the use cases and capabilities a model requires.

The second type of metric is the value delivered. "It is important to establish an anticipated benchmark for generative AI, such as reducing the time and effort spent to perform tasks, increasing throughput, or increasing demand generation," Livingston said. These metrics inform the enterprise of where and when to invest further in generative AI initiatives and ensure models continue to deliver value over time.

THIS ARTICLE IS PART OF

What is Gen AI? Generative AI explained

Which also includes:

- **8 top generative AI tool categories for 2025**
- **Will AI replace jobs? 17 job types that might be affected**
- **25 of the best large language models in 2025**

Customize metrics for specific use cases

Quantifying the performance of generative AI models can be challenging due to the vast number of use cases, said Buddhi Jayatilleke, chief data scientist at Sapia, a human resources AI platform. And the performance of LLMs can be defined in different ways across the development lifecycle. Various metrics provide more value across pre-training, fine-tuning and reinforcement learning from human feedback processes.

"The gold standard for evaluating generative models often involves human judgment, but this can be costly and slow, particularly at scale," Jayatilleke said. One approach is to use human evaluation sparingly and strategically, perhaps to validate and calibrate automated metrics. It's also helpful to set up feedback loops to learn from users, and to keep abreast of the latest research on new evaluation benchmark data sets to add to the automated tests.

Metrics can vary by the job function and service offered, said Todd Johnson, president of U.S. enterprise applications at Nexer Group, an IT services and consulting firm. Customer service scenarios can use traditional metrics such as customer retention, customer satisfaction and Net Promoter Score. Productivity scenarios can be more

difficult to measure directly. For example, the task completion rate needs to be balanced against work quality measures.

Johnson recommends carefully rolling out new tools or models by creating test groups to provide baselines and evaluate different tools. "Ultimately, it's still going to be a human decision as to whether these tools make us better at our job or not," he said.

8 areas to be evaluated with generative AI metrics

Here are eight areas -- and some of their metrics and benchmarks -- that should be tracked and evaluated to gauge the success of enterprise generative AI programs and projects.

1. ROI

From a financial perspective, measuring ROI helps determine if and how a machine learning (ML) program or project delivers meaningful value. But this value can come from various benefits such as increased sales, profits, productivity or customer engagement. "Many generative AI-based projects in organizations are still in the research stage, making it hard to predict their exact value," said Matan Libis, vice president of products at SQream, a data preparation platform for ML.

Libis recommends setting clear KPIs for success and allocating a sufficient budget for research.

2. Goal completions

Measuring goal completions, which reflect how many tangible and desirable business outcomes a model achieves, is another way to evaluate generative AI achievements, according to Israel Krush, CEO and co-founder at conversational AI platform Hyro.

For example, in an app that helps patients schedule physician appointments, Hyro measures how many such appointments the app handled end-to-end. Developing a new measure for each use case can show when it's delivering value. "It's never been easier to add a fun and engaging chatbot to your website," Krush said, "but if you don't really understand what it's there for, what's the point of having it?"

3. Fidelity

Another empirical method to gauge the success of generative AI-based systems in organizations is fidelity, which assesses the similarity between generated output and real data. A high fidelity score indicates the model's proficiency in producing realistic and accurate results. Libis said this aspect is critical for building trust in this

technology, as both organizations and their customers rely on these models to serve them faithfully and avoid misinterpretation.

However, achieving maximum ROI and fidelity together might not always be possible. Sometimes, improving one can come at the cost of the other. Libis has found it helpful to build a shared understanding of the application's specific requirements to strike the right balance between ROI and fidelity based on project goals and priorities.

4. Task performance

It's also helpful to assess task performance, Jayatilleke said. This area looks at how well the model responds to a given prompt, such as summarizing some text, solving a math problem or performing common-sense reasoning. A task-specific benchmark data set and related metrics can help evaluate how well the model performs against the benchmark.

Jayatilleke often uses the Massive Multitask Language Understanding, or MMLU, benchmark that covers subjects across STEM, the humanities, social sciences and more, ranging from an elementary to advanced professional level. Other relevant metrics include the following:

- **Generation consistency**, which measures whether similar prompts in the same context can lead to almost semantically similar responses.
- **Prompt sensitivity**, which measures how detailed a prompt needs to be to get the optimal response from the LLM.

5. Safety

Safety metrics help test for risks such as ethical concerns about generative AI, truthfulness, toxicity and security. This can include measuring the prevalence of racially biased responses, AI hallucinations or leaks of confidential information. Benchmarks such as TruthfulQA can complement human expert testing. Running multiple automated tests covering various concerns is the best practice, according to Jayatilleke. But as the training data and training parameters change, these benchmarks might not be able to capture the new learnings of the model.

6. Personality

At Sapia, Jayatilleke's team set up their own metrics to assess the personality projected by the different versions of the OpenAI GPT models. Their metrics related to personality dimensions or the emotional intelligence of the generative AI models that power chatbots to help better understand and compare their behavior. They found

significant differences between GPT-2, ChatGPT -- built on GPT-3.5 -- and GPT-4. This research helped them establish a baseline that could be used to evaluate how further adjustments to these models affect interactions with users.

7. Accuracy

Accuracy measures how well a model's predictions or outputs align with the desired results, which is not always easy to assess. "LLMs, in general, have an accuracy problem, and no one has been able to determine a standard method for evaluating an LLM's quality in this regard," said Yonatan Geifman, CEO and co-founder of Deci, a deep learning development platform.

It's often easier to assess accuracy in domains such as coding using benchmarks including the HumanEval database. In other domains, there are multiple evaluation methods to choose from, including the following:

- **Perplexity** is a metric that evaluates a language model's ability to predict the next word in a word sequence.
- **Inception score**, or IS, is a mathematical algorithm that measures the quality of generative AI images.
- **Fréchet inception distance**, or FID, analyzes images generated by generative adversarial networks for realism and diversity.
- **Precision** is a metric that measures the number of correct predictions made by a generative AI model.
- **Recall** details the ratio of positive samples classified by a model as positive vs. the total number of positive samples generated.
- **F1 score** measures an AI model's accuracy using the precision and recall factors.
- **Bilingual evaluation understudy**, or BLEU, is a metric used to automatically evaluate machine-translated text against reference text.
- **Recall-oriented understudy for gisting evaluation**, or ROUGE, measures the quality of a machine-translated summary against one created by a human.
- **Metric for evaluation of translation with explicit ordering**, or METEOR, scores machine translations based on word-to-word matching against a reference translation.
- **Consensus-based image description evaluation**, or CIDEr, measures a machine-generated sentence against human-generated information that is known

to be real and true.

- **Manual evaluation** is when a human compares machine-generated results on a case-by-case basis.

It's also common to ask users to count the number of likes or suggestions they accept.

The more significant challenge, Geifman said, lies not in choosing which metrics to use, but in selecting the appropriate evaluation data sets on which to apply these metrics. Inaccuracies are being discovered in some of the most commonly used data sets for evaluation. There are also concerns that large models such as GPT-4 were exposed to these evaluation data sets during training, and therefore they cannot be used as an objective evaluation data set.

8. Inference speed

Inference speed quantifies the model's performance in terms of speed and efficiency at runtime.

The latency of the model is typically measured in iterations per second, which directly affects the inference cost of the model. Lower latency leads to reduced compute cost, a smaller carbon footprint and an improved overall user experience. "It is important to consider model speed early in the process, as slow inference performance can become a major barrier for business scalability and operational cost efficiency," Geifman said.

Challenges in setting up a metrics program

It's important to note that many challenges can arise when setting up a metrics program for generative AI models. Some of these top challenges, according to Doug Ross, vice president and head of insights and data at Sogeti, part of Capgemini, include the following:

Subjectivity

Generative models often create outputs, such as text or images, meant to be consumed and interpreted by humans. Human evaluation is subjective, and individuals might disagree as to the quality of the result. This can make it difficult to establish a benchmark or baseline for evaluation. One way to overcome this is to crowdsource using multiple human evaluators to assess the results.

Bias

Models can mimic or even amplify biases in training data or methods, which can have ethical consequences. Monitoring for bias is an important step, and more diverse training data sets are helpful. A 2023 Capgemini Research Institute report, "Harnessing the Value of Generative AI: Top Use Cases Across Industries," found that 51% of surveyed organizations cited a lack of clarity on underlying training data as a challenge stalling generative AI implementations.

Scalability

As models and data sets grow, the computational power and attendant expenses can also increase. Efficiency and financial metrics can help plan for and address scalability challenges.

Attacks

Some models are susceptible to attacks, such as jailbreaks, that leak information or provide unwanted outputs. Red team testing of models is helpful to determine if additional safeguards are needed.

Selecting the right metrics

Given the various ways to gauge model accuracy, performance and efficiency, choosing the right metrics becomes a key factor in a production deployment.

Monitoring

Continuous checks of model output will help ensure the model does not drift or that answers do not change over time. This can be done with automated testing and human oversight.

The future of generative AI metrics

Tools to assess performance and benefits are a work in progress. Experts expect to see advances on many fronts.

Veracity can be a challenging, if not impossible, metric to establish for LLMs, Livingston observed. She predicts significant developments in the ability to monitor and measure the authenticity and veracity of responses and outputs to determine drift and bias movements.

Libis expects to see more progress on explainability and interpretability metrics, while Jayatilleke believes it is time to consider better metrics to measure the carbon footprint of training generative models. It will also be important to develop metrics to

assess the contamination of future training data sets with AI-generated content as those become ubiquitous, he said.

Johnson predicts that additional telemetry will become part of generative AI tool sets to provide detailed data on how the AI model is used. It will also provide insight into how the model is performing and guidance on where there is room for improvement.

Ross expects to see improved performance evaluation metrics that better capture model accuracy and generalized performance. Some promising new metrics include precision and recall; FID+, an enhancement on the Fréchet inception distance algorithm; and learned perceptual image patch similarity, or LPIPS, distance.

"A combination of better evaluation techniques, more diverse training data sets and transparent architectures are likely to be among the directions used in improving objective measures of model performance," Ross said.

↘ Next Steps

AI content generators to explore

The best large language models

Generative AI challenges that businesses should consider

Attributes of open vs. closed AI explained

Top resources to build an ethical AI framework

Dig Deeper on AI technologies

What are vision language models (VLMs)?

By: Alexander Gillis

Amazon Q, Bedrock updates make case for cloud in agentic AI

By: Beth Pariseau

What is Fréchet inception distance (FID)?

By: George Lawton

How businesses can measure AI success with KPIs

By: Jerald Murphy

Latest TechTarget
resources

BUSINESS ANALYTICS

CIO

DATA MANAGEMENT

ERP

Search Business Analytics

Super Bowl teams show analytics gaining ground in NFL

As data-driven decision-making gains ground in pro football, both the Eagles and Chiefs have gained advantages from using ...

MicroStrategy adds personalization to GenAI-powered bot

The longtime independent analytics vendor's latest update aims to provide users with a more tailored query and response ...