

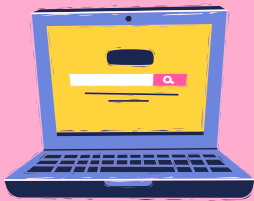
Gen AI Metrics & KPIs

A quick cheatsheet for crafting Metrics and KPIs for Generative AI products / features. 2 categories of Metrics to consider:

- TECHNICAL METRICS
- BUSINESS METRICS

Technical Metrics

Model Metrics



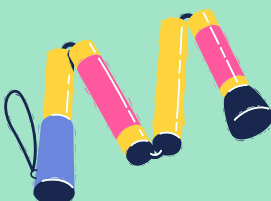
Measuring the underlying model quality

Application metrics



Measuring the Gen-AI end-user application quality

System Metrics



Measuring the overall system's tech stack quality

LLM Model Metrics

Metric	Description
SuperGLUE	Boolean Questions, CommitmentBank, Choice of Plausible Alternatives (COPA), Multi-Sentence Reading Comprehension (MultiRC), Reading Comprehension with Commonsense Reasoning Dataset (ReCoRD), Recognizing Textual Entailment (RTE), Word-in-Context: (WiE), Winograd Schema Challenge (WSC), Broad Coverage Diagnostics, Analyzing Gender Bias in Models
BLEU (BiLingual Evaluation Understudy)	Evaluates the output of your LLM application against annotated ground truths (or, expected outputs). It calculates the precision for each matching n-gram (n consecutive words) between an LLM output and expected output to calculate their geometric mean and applies a brevity penalty if needed.
ROUGE (Recall-Oriented Understudy for Gisting Evaluation)	Evaluates text summaries from NLP models, and calculates recall by comparing the overlap of n-grams between LLM outputs and expected outputs. It determines the proportion (0-1) of n-grams in the reference that are present in the LLM output.
METEOR (Metric for Evaluation of Translation with Explicit Ordering)	Calculates scores by assessing both precision (n-gram matches) and recall (n-gram overlaps), adjusted for word order differences between LLM outputs and expected outputs. It also leverages external linguistic databases like WordNet to account for synonyms. The final score is the harmonic mean of precision and recall, with a penalty for ordering discrepancies.
CIDEr (Consensus-based Image Description Evaluation)	Measures the similarity between a generated caption and the reference captions, and it is based on the concept of consensus: the idea that good captions should not only be similar to the reference captions in terms of word choice and grammar, but also in terms of meaning and content

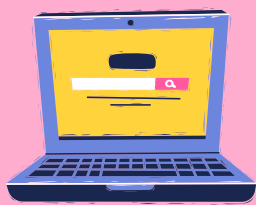
Gen AI Metrics & KPIs

A quick cheatsheet for crafting Metrics and KPIs for Generative AI products / features. 2 categories of Metrics to consider:

- TECHNICAL METRICS
- BUSINESS METRICS

Technical Metrics

Model Metrics



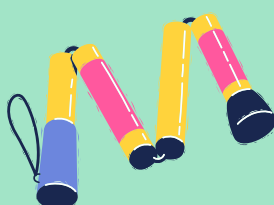
Measuring the underlying model quality

Application metrics



Measuring the Gen-AI end-user application quality

System Metrics



Measuring the overall system's tech stack quality

Application Metrics

Metric	Description
Error rate	The percentage of responses provided by the model, which are incorrect or invalid. Human evaluation helps define and generate this metric
Latency	The time delay between when a query is submitted to the model and when it returns the response. This includes the parallel processing capabilities of the model, model architecture, deployment infrastructure and availability
Accuracy range	The baseline expectation for precision accuracy thresholds for the model to meet. For this metric, it is often helpful to establish a red team to analyze and challenge your model
Safety Score	The number of harmful categories and topics that may be considered sensitive for the business
Bias	Checks and controls for social biases related to gender, race, and other factors
Groundedness	Measure assesses the correspondence between claims in an AI-generated answer and the source context, making sure that these claims are substantiated by the context. Even if the responses from LLM are factually correct, they'll be considered ungrounded if they can't be verified against the provided sources (such as your input source or your database)
Relevance	Measure assesses the ability of answers to capture the key points of the context. High relevance scores signify the AI system's understanding of the input and its capability to produce coherent and contextually appropriate outputs. Conversely, low relevance scores indicate that generated responses might be off-topic, lacking in context, or insufficient in addressing the user's intended queries
Coherence	Assesses the ability of the language model to generate text that reads naturally, flows smoothly, and resembles human-like language in its responses
Fluency	Assesses the extent to which the generated text conforms to grammatical rules, syntactic structures, and appropriate vocabulary usage, resulting in linguistically correct responses

Gen AI Metrics & KPIs

A quick cheatsheet for crafting Metrics and KPIs for Generative AI products / features. 2 categories of Metrics to consider:

- TECHNICAL METRICS
- BUSINESS METRICS

Technical Metrics

Model Metrics



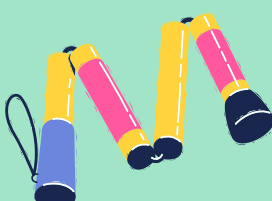
Measuring the underlying model quality

Application metrics



Measuring the Gen-AI end-user application quality

System Metrics



Measuring the overall system's tech stack quality

System Metrics

Metric	Description
Data relevance	The degree to which all of the data is necessary for the current model and project. Be warned, extraneous data can introduce biases and inefficiencies that can lead to harmful outputs.
Data and AI asset and reusability	The percentage of your data and AI assets that are discoverable and usable.
Throughput	The volume of information a gen AI system can handle in a specific period of time. Calculating this metric involves understanding the processing speed of the model, efficiency at scale, parallelization, and optimized resource utilization
System latency	The time it takes the system to respond back with an answer. This includes any ingress- or egress-based networking delays, data latency, model latency, and so on
Integration and backward compatibility	The upstream and downstream systems APIs available to integrate directly with gen AI models. You should also consider if the next version of models will impact the system built on top of existing models (not just limited to prompt engineering)
Real-time updates	System that's updated with recent information can contribute more broadly and produce better results
Cost	System that's sustainable as the application and user base scales
Compute and Infrastructure sustainability	Scalable compute, training infrastructure and data pipeline as the application scales and matures over time

WWW.PRODUCTBULB.COM

Gen AI Metrics & KPIs



GEN AI PRODUCT METRICS

Measuring the end product’s user metrics, business value, by isolating or zooming out from the technical metrics.

Business Metrics

Metric	Description
Adoption rate	The percentage of active users over the lifetime of a campaign or project divided by the total intended audience
Frequency of use	The number of times queries are sent per user on a daily, weekly, or monthly basis.
Session length	The average duration of continuous interactions
Queries per session	The number of queries users submit per session
Query length	The average number of words or characters per query
Abandonment rate	The percentage of sessions ended before users find answers
User satisfaction	Surveys assessing user experience or other customer satisfaction metrics, such as Net Promoter Score (NPS)