SEPTEMBER 2024

Guide to Measuring AI Performance

This guide aims to equip those assessing AI systems in the public sector with an understanding of the possible metrics at their disposal to evaluate AI performance.

GovAI COALITION

Introduction

As the use of AI systems in government continues to grow, it is critical that those tools are adequately assessed during the public procurement process. To conduct such assessments, those in government need an appropriate set of metrics they can reliably use to evaluate the efficacy and contextual merit of AI systems. Many vendors rely on technical metrics (e.g., accuracy or ROUGE scores) for performance assessment in AI FactSheets, while public agencies often use additional non-technical criteria when assessing AI tools.

The selection of appropriate metrics is further complicated by the unique context of each AI use case. For instance, the appropriate performance of a metric for an AI-based translation tool – say, reading level requirement – may vary depending on whether the tool is used for emergency response (3rd-grade reading level) or public communications (8th-grade reading level).

This guide aims to equip those assessing AI systems in the public sector with an understanding of the possible metrics at their disposal to evaluate AI performance. Measuring the performance of AI systems is a comprehensive task, and while this guide does not address all aspects of such measurement, it provides an initial set of guidance for agencies seeking to benchmark AI systems.

How do vendors measure performance?

Vendors often use traditional AI metrics when assessing their tools. These metrics are based on established AI model types like regression (continuous outputs), classification (categorical outputs), and natural language processing (e.g., LLMs). Figure 1 outlines for each model type: a simple definition, common metrics, and use case examples.

As you view the chart, keep in mind that many AI systems draw from multiple model types, so multiple metrics across model types may be needed in their assessment.

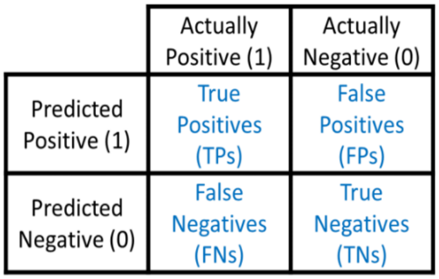
Model type	Definition	Common metrics	Examples
Classification	Used to categorize data into predefined labels or classes.	• Accuracy • Precision • Recall • F1 Score • F2 Score • F-beta Score • Area Under the Curve-Receiver Operating Characteristic Curve (AUC-ROC)	• Medical diagnosis • Sentiment analysis • Spam email detection • Object detection
Regression	Used to find a relationship between variables and continuous numerical outcome.	• Mean Squared Error (MSE) • Root Mean Squared Error (RMSE) • Mean Absolute Error (MAE) • R-squared	• House prices • Projected sales • Cyber risk • Weather patterns • Future stock prices
Clustering	Used to categorize data without predefined labels.	• Silhouette score • Dunn index • Rand index	• Market segmentation • Search results • Anomaly detection • Identifying themes

Ranking	Used to rank a series of inputs.	• Mean Average Precision (MAP) • Normalized Discounted Cumulative Gain (NDCG) • Precision at K	• Search engine results • Procurement recommendations • Case management • Emergency services
Generative AI	AI models that are used to create new content, such as text, images, and audio	• Flesch Kincaid Readability (measures the grade-level of written output) • Word error rate (WER)	• Chatbots that communicate with the public • Text-to-speech systems • Image generators (i.e., DALL-E)

Figure 1: A table covering general categories of AI models, their definitions, common metrics used, and examples of AI models.

Definitions of common metrics and concepts

Confusion Matrix (used for Accuracy, Precision, Recall, F-1 Score):



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

[1]

Figure 2: Confusion matrix.

Accuracy, precision, and recall are commonly used technical metrics that are built from the confusion matrix in Figure 2. It may be helpful to ask vendors if the system has a higher false positive rate or a higher false negative rate, as it is common for AI systems to exhibit one or the other. There is a well-documented tradeoff between false positives and false negatives. For a particular use case (e.g., automated counting of hikers in the woods), a high false positive rate (i.e., the model tends to mistake deer for people and overcount the number of hikers) does not matter as much as the low false negative rate (i.e., the model rarely mistakes a hiker as a deer). For safety purposes, overcounting hikers matters much less than undercounting the number of hikers in the woods.

Some use cases, like gunshot detection, require both low false negatives (i.e., the model rarely misses detecting an actual gunshot) and low false positives (i.e., the model rarely categorizes similar sounds, like fireworks, as gunshots) for the technology to be considered effective. In such cases, we suggest using the F1 score in your evaluation.

Figure 3 is a chart of common metrics with definitions and a general good range. Keep in mind that this range will vary based on the context of the use case.

Drift

Across all model types, drift is a common concern. While not necessarily a metric, drift is the phenomenon of models decreasing their performance over time. Tolerances and breaches of

[1] https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/

metrics that are set by humans external to the AI system will help detect drift.[2] There are several types of drift, including target drift, concept drift, covariant drift, label drift, and algorithmic drift, that are further explained in the Figure 3. Common models where drift is of particular concern includes spam detection, anomaly detection, predicting future stock prices (e.g., models during COVID-19 experienced drift).

Common Metrics	Metric Definition	Good Range
Accuracy	The percentage of correct predictions out of all predictions made.	85%-100% (depending on complexity of task)
Precision	How often the model is correct when it predicts a certain class. Calculated as $True\ Positives/(True\ Positives + False\ Positives)$	70%-100%
Recall	How often the model finds the relevant items for a particular class, or the true positive rate. Calculated as $True\ Positives/(True\ Positives + False\ Negatives)$	70%-100%
F1 Score	A measure that averages precision and recall such that both must be high for the F1 score to be high.	• 1 = perfect performance • 0.9 = very good • 0.8 - 0.9 = good • 0.5 - 0.8 = ok • <0.5 = poor performance
AUC-ROC	Measures how well the model distinguishes between positive and negative classes.	0.7-1 (values closer to 1 are better)
Mean Squared Error (MSE)	The average error of predicted values (the distance between predicted and actual values), giving more weight to larger errors. While similar to MAE, this metric penalizes larger errors more.	Lower is better (task-dependent)

[2] https://www.datacamp.com/tutorial/understanding-data-drift-model-drift

Root Mean Squared Error (RMSE)	The square root of the MSE.	Lower is better (task-dependent)
Mean Absolute Error (MAE)	The average difference between predicted and actual values.	Lower is better (task-dependent)
R-squared	A score of how well the model fit the data. This metric can be increased simply by adding more independent variables to the model.	0.7-1 (values closer to 1 are better)
Silhouette Score	Measures how well data points are grouped into clusters, with higher scores indicating an object is a good fit for its own cluster and a poor fit for other clusters.	Range: -1 to 1 • 0.7 - 1 = strong • 0.5-0.7 = reasonable • 0.25-0.5 = weak • 0 = Clusters not meaningfully distinguishable from each other • -1 to 0: Clusters are incorrectly assigned
Dunn Index	A higher score indicates compact clusters that are further away from each other	Range: 0 to infinity Higher values are better
Rand Index	Compares the clustering results to a true set of labels to measure how similar they are. The higher the better.	Range: 0 to 1 0.7+ is good
Mean Average Precision (MAP)	Commonly used in ranking tasks (i.e., information retrieval and object detection models). A high score in object detection means that the model detects multiple objects well in the same image. A high	0.5-1 (closer to 1 is better)

	score for information retrieval means that the model returns a list of documents that are relevant and ranked well.	
Normalized Discounted Cumulative Gain (NDCG)	Metric used for information retrieval by comparing model's ranked list to ideal order. Higher score indicates relevant items are closer to the top of the list.	0.5-1 (closer to 1 is better)
Precision at K	Measures how many items with the top K positions are relevant.	70%-100%
Target Drift	Indicated by changes in the distribution of the target variable over time.	Minimal drift is better
Concept (Model) Drift	Degradation of a model's performance over time due to changes in the statistical properties of the data or the relationship between the input and output variables. For example, a model predicting snow shovel sales may perform poorly in the summer if trained only on winter data.	Minimal drift is better
Covariant (Data) Drift	Indicated by changes in the distribution of the input features or changes in the statistical properties of data inputs.	Minimal drift is better
Label Drift	Indicated by changes of in the performance due to the meaning or definition of labels changing over time. For example, a spam email detector may perform poorly when the definition of a spam email changes.	Minimal drift is better
Flesch Kincaid Readability	Measures the grade level of written output.	Target according to agency guidelines. For example, the City of San José requires all public communications to be at an 8th-grade reading level or lower, and

		at a 3rd grade reading level or lower in emergency situations.
Algorithmic drift	Indicated by changes in the performance or the system's processing of the same data results during learning and technical enhancements of an algorithm. For example, algorithmic drift occurs when updates or modifications to the algorithm leads the system to provide different results when fed the same inputs it was given previously. This is not drift that occurs due to changes in the data, targets, labels.	Minimal drift is better.

Figure 3: This table lists out common metrics, their definitions, and the performance value for the AI model to be considered "good". Keep in mind that this is a general range, and specific values will vary for different contexts.

How can agencies measure performance?

The metrics used by vendors are typically quantitative metrics. We suggest using a mix of both quantitative and qualitative metrics that capture both the technical capability of the AI system in addition to how well it is suited to the unique context of the use case and the agency's business and operational needs.

Non-quantitative metrics include:

1. Customization and flexibility of the model
2. Ethical considerations (e.g., environmental impact)
3. Privacy considerations
4. User engagement and satisfaction
5. Model explainability
6. Model effectiveness
7. Model transparency
8. Ease of updating the model

The list of quantitative and non-technical metrics above are generalized, and as you use them in your assessment of AI systems, you will need to adapt them to the use case at hand. For example, Figure 4 adapts commonly used metrics like accuracy and recall, to unique use cases. Further detail on metrics, especially non-quantitative metrics, will be detailed in a future version of this document.

Use Case	Example metrics
Question and answer	• Accuracy of statements in response • Relevance of response to question
Document Summarization	• Quality of summary • Relevance of summary to document, as determined by user • ROUGE score
Meeting summarization	• Correctly recording decisions reached in meeting • Accuracy when assigning next steps or action items
Document Retrieval	• Relevance and recency of the documents recommended (i.e., providing the documents that an experienced staff-member familiar with the document space would have provided).
Memo or policy drafting	• Time required to edit or correct the document • Adherence to agency writing standards • Flesch-Kincaid Score
Translation	• Cultural sensitivity of translation • Reading level of translation • BLEU score
Coding	• Conciseness of code generated • Ability to find and resolve bugs • Time saved by software engineers

Figure 4: This is a table demonstrating specific use cases and metrics that may be used to assess AI tools meant to address these use cases.

Get involved

This guide is an initial step towards future work around measuring the performance of AI systems, and was developed in collaboration with government, civil society, and industry members of the GovAI Coalition. If you would like to contribute to future iterations of this work and continue conversations on measuring AI performance, you can join the Coalition’s Vendor Agreements Working Group (under the Adoption Support Committee) and the Use Cases Committee. Register for these committees through the GovAI Coalition registration form. If you have any questions on how to get involved, contact us at digitalprivacy@sanjoseca.gov.