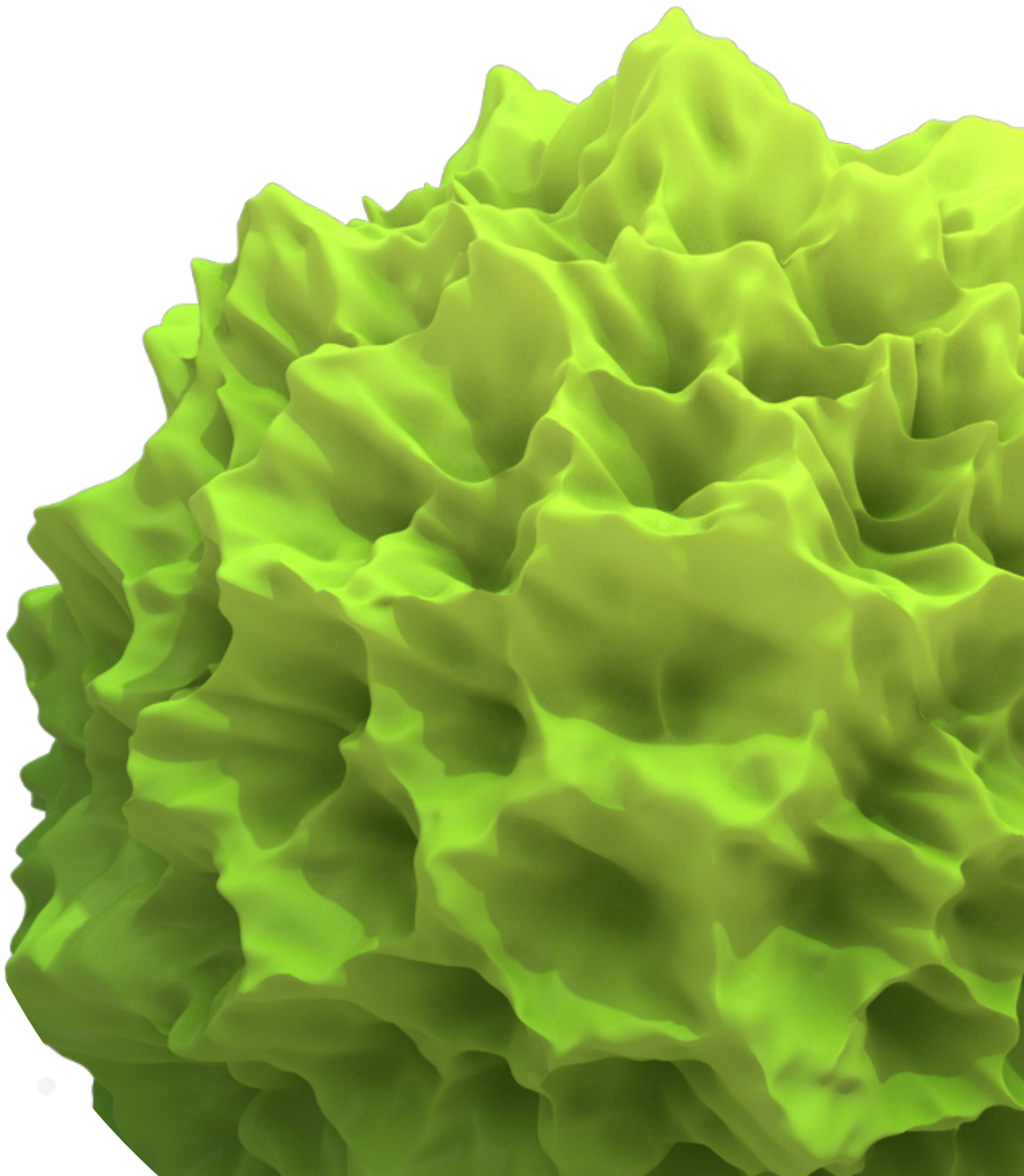
Deloitte.

Generative AI Model Validation

Challenges, Approaches and Future Innovations

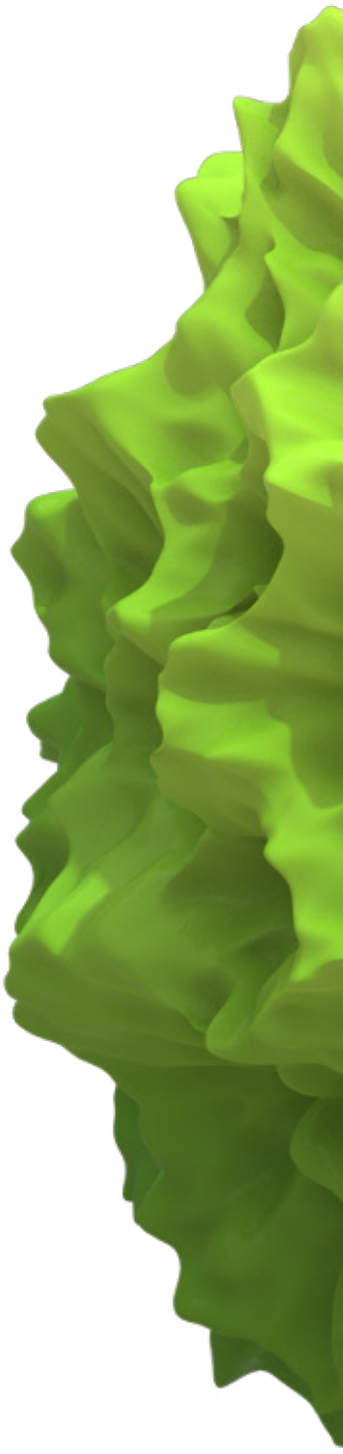


Contents

Executive Summary

Generative Artificial Intelligence (GenAI) is transforming how organizations create content, analyse data, and make decisions. However, ensuring that these models operate reliably and responsibly poses challenges. Unlike traditional machine learning models, GenAI presents unique validation challenges stem from factors such as high model complexity, their “black-box” nature, non-deterministic outputs and hallucinations. These factors make it difficult to consistently assess model performance, detect and mitigate risks such as bias, hallucinations, or harmful content, and maintain stakeholder confidence.

This white paper takes a close look at the current landscape of GenAI model validation, examining both established industry-standard practices and the limitations that can affect the trustworthiness of these models. It underscores that systematic, well-designed validation frameworks are essential for unlocking the full potential of GenAI while maintaining ethical and operational integrity. We also explore emerging strategies and future evolution of validation processes. By highlighting these insights, we aim to provide practical guidance for organisations to build confidence in their GenAI systems and deploy them responsibly.



The Approach to Generative AI Model Validation

A robust GenAI validation strategy is fundamental for a successful AI transformation and implementation across the organisation, by ensuring that AI systems outputs are trustworthy and aligned with the organisations' objectives. It is not merely a technical exercise but a continuous responsibility throughout the AI system lifecycle, from data preparation, model development and training to ongoing monitoring. An end-to-end validation process helps organisations manage risk, maintain ethical and performance standards, and build long-term confidence in their AI systems.

The following fundamental principles underpin responsible AI development and deployment, and should guide the validation process:

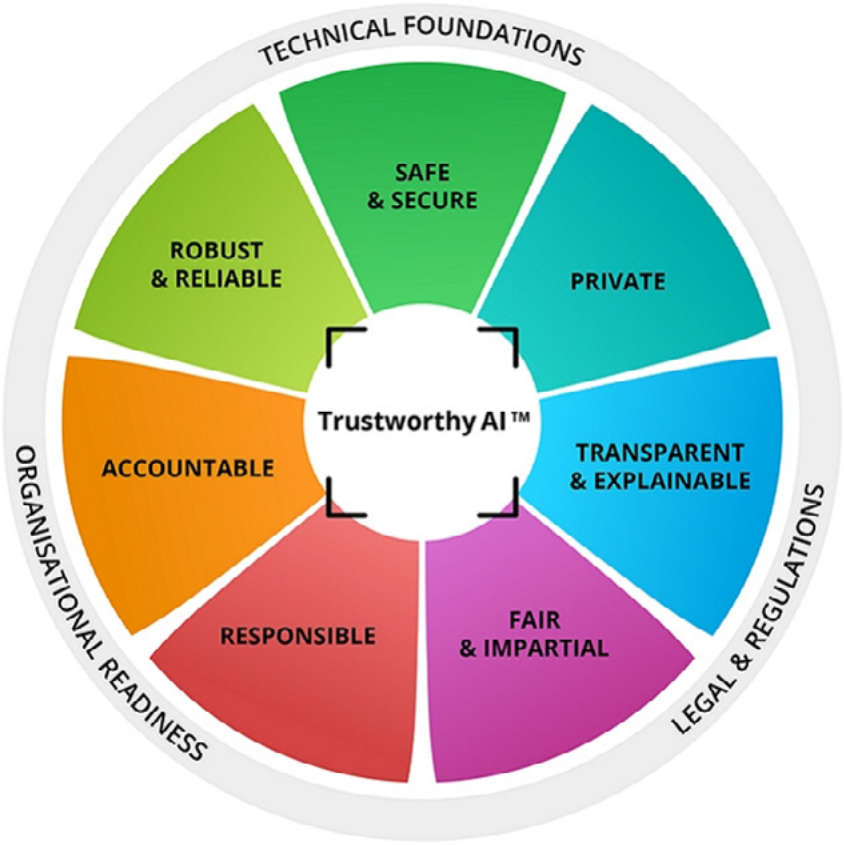


Figure 1: Deloitte's Trustworthy AI Framework

- **Private –** AI system must protect user privacy by ensuring that data is used only for its stated purpose and duration, and that sensitive personal and confidential data are protected using appropriate measures that adhere to relevant privacy laws.
- **Transparent and Explainable –** AI systems should be explainable and auditable, which mandates a clear understanding of how AI models function and the underlying mechanisms by which they arrive at their decisions.
- **Fair and Impartial –** Mitigate biases inherent in training data and model outputs is crucial to prevent discriminatory outcomes, which requires the use of diverse datasets and continuous monitoring for emergent biases.
- **Responsible –** AI systems must be developed and deployed in a socially responsible and sustainable way, considering long-term societal impacts beyond immediate applications.
- **Accountable –** Robust policies should be in place to define and enforce accountability for decisions derived from or influenced by technological systems, reinforcing governance and responsible use.
- **Robust and Reliable –** AI systems should produce accurate and consistent outputs, demonstrates resilience to errors, and rapidly recover from unforeseen disruptions or misuse.
- **Safe and Secure –** AI systems must be proactively designed and operated to protect against potential harm, and aim for equitable application, access, and outcomes. It should also include security controls against adversarial attacks.

Beyond Traditional AI Model Validation

Validation techniques traditionally used for "classical" AI models often fall short when applied to GenAI models, owing to their intrinsic complexity. While traditional methods focus on fixed metrics like predictive accuracy against known ground truth (AKA golden standard reference data), GenAI demands broader criteria.

GenAI models operate as "black-box" systems, producing non-deterministic and sometimes inconsistent outputs even for identical or similar inputs, and can generate plausible yet factually incorrect outputs due to their lack of real-world grounding. They can also inherit perpetual biases from training data, while their black-box nature limits interpretability and hinders error detection. Beyond technical limitations, GenAI models raise complex ethical and operational challenges that go beyond traditional AI validation.

These challenges, necessitate a more qualitative, human-centric, and continuous validation approach that integrates ethical and societal dimensions.

Key Challenges in GenAI Model Validation

In this subsection, we highlight the key challenges faced in validating GenAI models. These challenges fall into four main categories, reflecting the technical, ethical, and operational complexities of building trustworthy GenAI systems.

Data-Centric Challenges

Data-centric challenges in generative AI validation stem from the fundamental opacity and scale issues surrounding the vast volumes of data that these models ingest, typically sourced from publicly accessible internet content without transparent disclosure of specific sources, composition, or quality control measures. The training of Large Language Models and other generative systems demands massive, high-quality datasets often comprising billions or trillions of tokens, yet the proprietary nature of these datasets means that users, researchers, and even regulators lack visibility into the data used, hindering assessment, auditability, and accountability.

Furthermore, these web-scraped datasets likely contain personal and sensitive information, raising significant privacy and consent issues, making compliance with regulations like GDPR and CCPA difficult to verify. This creates a fundamental tension between the technical demands of large-scale GenAI and the essential principles of transparency, accountability, and responsible data stewardship.

1. Training Data Quality and Representativeness:

Massive datasets, often scraped without comprehensive quality control, lead to noise, errors, and embedded biases (e.g., geographic, cultural). Furthermore, these datasets are snapshots, risking outdated information and significant domain gaps.

2. Data Contamination and Leakage:

The sheer scale makes it hard to prevent evaluation data from inadvertently entering training sets, leading to inflated performance metrics and models overfitting to benchmarks.

3. Ground Truth Establishment:

Many GenAI outputs, especially creative ones, lack objective ground truth, requiring subjective human judgment. Establishing ground truth is complex for cross-modal content and highly dependent on context, intent, and application.

4. Evaluation Dataset Construction:

Creating comprehensive evaluation datasets for open-ended generative tasks is practically impossible. Human annotation introduces issues of consistency, bias, and scalability, while evolving models necessitate continuous dataset updates.

Model-Intrinsic Challenges

Model-intrinsic challenges stem from the fundamental architecture, training mechanisms, and behavioural characteristics of generative models themselves, representing obstacles that are inherent to how these models function and pose significant barriers to effective validation. GenAI models are designed to create novel content through complex, multi-layered neural architectures that incorporate stochastic processes, emergent behaviours, and sophisticated reasoning capabilities that were not explicitly programmed. These intrinsic characteristics make GenAI model powerful and versatile but simultaneously create validation challenges that cannot be addressed through conventional testing approaches.

1. Output Variability and Stochastic Behaviour

Most GenAI models are stochastic, producing varied outputs from identical inputs, which complicates traditional validation's need for repeatability. Generation parameters (e.g., temperature, top-k, and nucleus sampling) and even minor prompt changes can drastically alter outputs making it difficult to establish robust validation protocols.

2. Emergent Behaviours and Capabilities

GenAI models often exhibit unpredictable capabilities not explicitly trained for, often appearing only at scale, with these emergent behaviours being difficult to predict, test for, or validate comprehensively during development. The ability of these models to learn from examples further varies performance based on the quality, relevance, and presentation of in-context examples. Furthermore, these models may exhibit complex reasoning behaviours that are difficult to validate systematically.

3. Hallucination and Factual Accuracy

Models frequently generate plausible but factually incorrect information (hallucinations) and may incorporate/blend information from training data without proper attribution, which are hard to detect automatically and can further complicate accuracy verification.

4. Safety and Alignment Challenges

GenAi Models may generate content that is harmful, offensive, or dangerous, even without explicit prompting. GenAi model behaviour may also drift from intended objectives, particularly in complex tasks, and can be manipulated through carefully crafted adversarial prompts to bypass safety measures and generate prohibited content or access sensitive information.

Ethical, Legal, and Societal Challenges

Ethical, legal, and societal challenges in generative AI validation represent a complex landscape where technical validation intersects with broader questions of responsibility, compliance, and social impact that extend far beyond traditional performance metrics. Operating within intricate social contexts, GenAI outputs can influence behaviour, perpetuate biases, and blur lines of authorship and intellectual property. Although existing regulatory frameworks are evolving to keep pace with technological advances, they remain fragmented across different territories. This creates geographic complexity requiring a robust horizon scanning to navigate the shifting regulatory landscape. Beyond compliance, growing societal concerns about misinformation, privacy, and manipulation demand validation approaches that consider human values and social norms. Addressing these issues requires interdisciplinary collaboration to assess GenAI's broader implications. Key challenges include:

1. Bias and Fairness:

Validating fairness in generative AI is inherently complex, which requires assessing outputs across multiple demographic groups and use cases. Defining and measuring bias and fairness can be contentious, with stakeholders often holding conflicting definitions of equitable treatment. Key challenges include lack of clear equality and discrimination legislations, securing representative datasets, particularly for marginalised communities, and establishing universal fairness criteria given cultural differences. Furthermore, intersectional bias demands validation approaches that detect and measure bias across multiple attribute combinations, carefully avoiding the reinforcement of existing stereotypes.

2. Intellectual Property and Copyright:

GenAI models trained on vast web data risk inadvertently reproducing copyrighted material, posing significant intellectual property infringement challenges. Detecting this goes beyond direct copying, extending to 'sufficient similarity' and navigating subjective legal concepts like fair use and transformative work. The global nature of GenAI means compliance with diverse national copyright laws, while the sheer volume of generated content makes manual review impractical, necessitating imperfect automated detection systems.

3. Privacy and Data Protection:

Validating GenAI compliance with privacy regulations (e.g., GDPR, CCPA) is challenging. It requires ensuring models don't inadvertently reproduce personally identifiable information (PII) from training data. Detecting PII is difficult, extending beyond obvious details to subtle embedded, inferred, or behavioural patterns that could identify individuals. Cross-border data transfer regulations add complexity, as models deployed globally face conflicting privacy frameworks. Furthermore, the 'right to be forgotten' and data deletion mandates create ongoing validation challenges for removing PII from model outputs post-training.

4. Transparency and Explainability:

Regulatory frameworks increasingly demand AI explainability, but GenAI models pose unique challenges due to their complex, stochastic nature. Validating this requires not only technical interpretability but also ensuring explanations are understandable and meaningful to diverse stakeholders (users, regulators, affected parties). Further complicating matters are the potential for AI-generated explanations, whose trustworthiness must be assessed, and the need for explanations to account for GenAI's dynamic, context-dependent generation process.

Operational & Security Challenges

Operational and security challenges in GenAI validation bridge the gap between theoretical model capabilities and real-world deployment. These systems face diverse threats, operational constraints, and complex production realities, demanding consistent performance across varied environments, defence against adversarial attacks, and service quality under fluctuating demand. GenAI's resource intensity, stochastic behaviour, and unpredictable outputs amplify security concerns and operational complexity, especially at scale. Effective validation requires comprehensive testing, anticipating evolving threats, and continuous assurance throughout the operational lifecycle. Key challenges include:

1. Model Deployment challenges

Validating GenAI across different versions and deployment environments is challenging due to varying hardware, software, and optimisation. Updates and fine-tuning can subtly alter behaviour, necessitating comprehensive validation within CI/CD pipelines to detect regressions and emergent issues, despite the computational demands of testing every configuration.

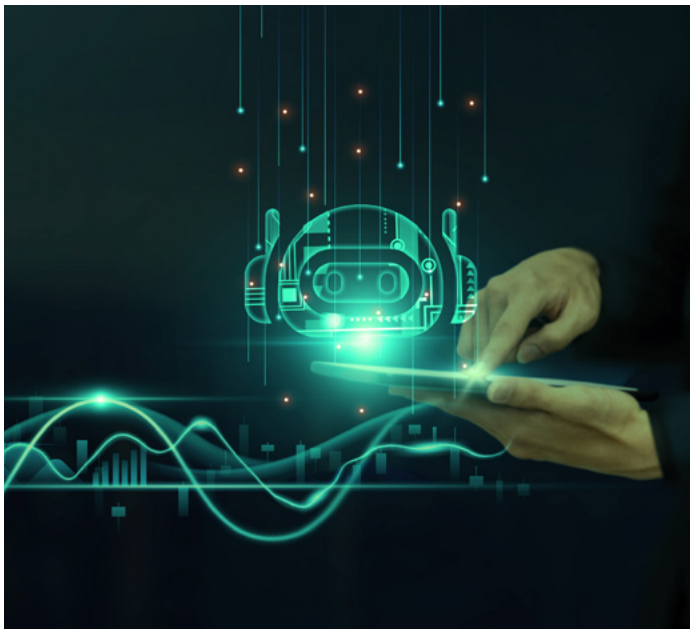
Additionally, a major operational challenge is the rapidly increasing and often unpredictable cost of token usage. As GenAI models charge for both input and output tokens, costs can rise sharply with long prompts, multi-turn interactions, retrieval-augmented systems, and agentic workflows. Without proper validation of prompt optimisation, output limits, and usage governance, organisations may face budget overruns that erode the expected business value of AI deployment.

2. Adversarial Attacks and Security challenges

GenAI faces unique security threats like prompt injection, model extraction, and data poisoning. Validating security requires ongoing red-teaming efforts to identify vulnerabilities across multiple attack vectors, balancing security with utility. The complexities of third-party providers of GenAI systems and global accessibility create numerous attack surfaces, demanding specialised validation approaches.

3. Monitoring and Anomaly Detection

Continuous production monitoring is crucial for detecting anomalous behaviour, performance degradation, and emergent issues. However, defining "normal" for high-dimensional generative outputs is difficult. Monitoring systems must distinguish legitimate model evolution from problematic changes (e.g., bias, quality degradation) while operating with minimal latency across global deployments, all while managing false positives and maintaining privacy.



Current Approaches to GenAI Model Validation

Validating GenAI models requires a combination of established techniques and emerging practices to ensure they deliver reliable, safe, and meaningful outcomes. In this section we outline the commonly accepted industry approaches along with their limitations.

Quantitative Metrics for GenAI Model Validation

Quantitative metrics remain the most common way to evaluate GenAI models, typically divided into reference-based and reference-free approaches. Reference-based metrics compare generated text against human-annotated ground-truth, while reference-free metrics attempt to assess quality without such references.

Reference-Based Text Metrics

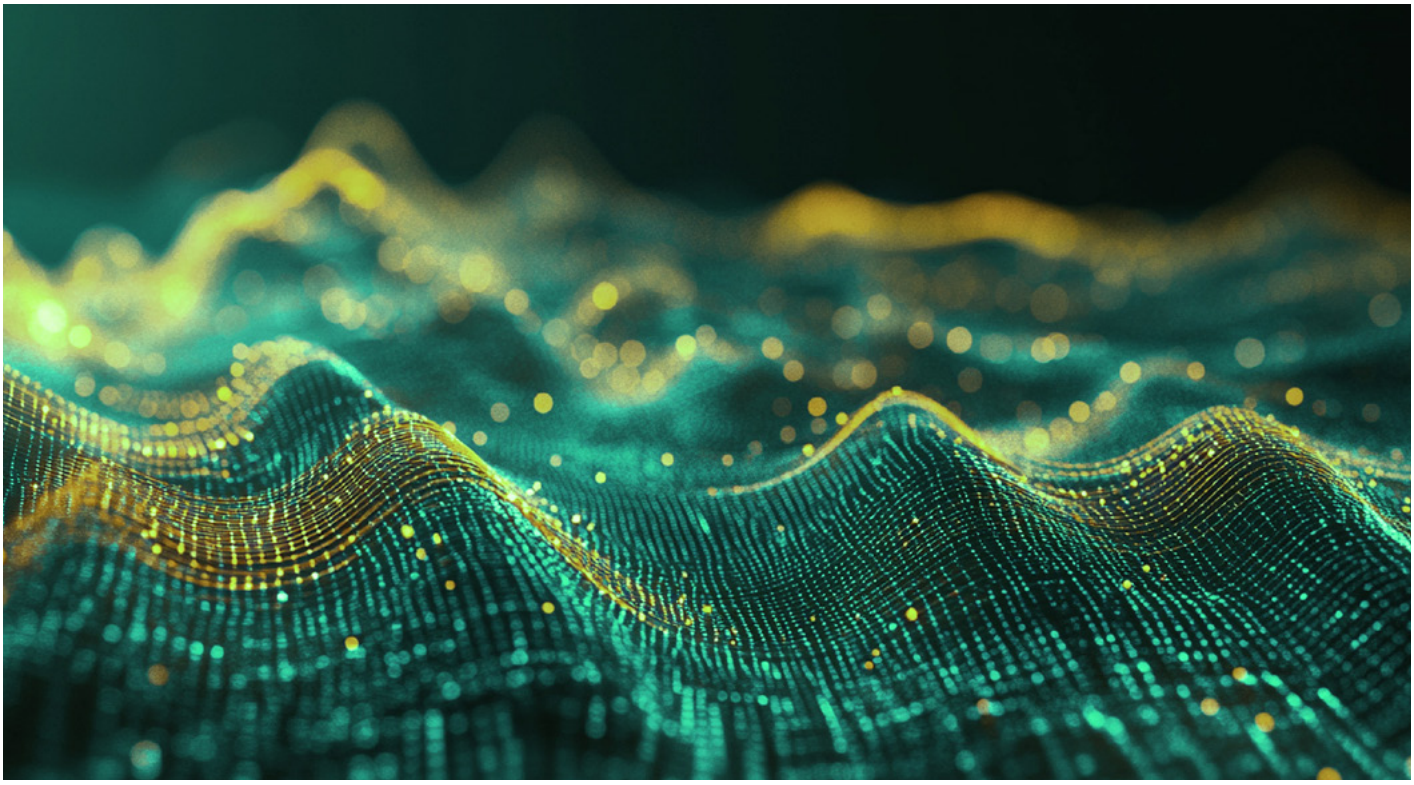
These metrics compare the generated text against one or more human-written reference answers. They are generally fast, interpretable, and effective for benchmarking in controlled tasks. A summary of most used metrics can be found in Table 1

Table 1: Summary of Reference-based Text Metrics

Metric	Application	Method	Strengths/ Limitations
BLEU (Bilingual Evaluation Understudy)[1]	Machine translation, summarisation	Measures n-gram overlap (precision) between the generated text and reference text.	**Strengths:** Fast, highly interpretable, and remains the standard for benchmarking in machine translation **Limitations:** Penalises valid paraphrases; only measures lexical similarity; lacks correlation with human judgement for fluency.
ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[2]	Summarisation, abstractive tasks	Measures the longest common subsequence (LCS) overlap, focusing on recall.	**Strengths:** Fast, good for evaluating informativeness and content preservation, strong focus on recall. **Limitations:** Highly dependent on reference quality; may favour longer, less concise summaries; ignores word order beyond the LCS.

Metric	Application	Method	Strengths/ Limitations
METEOR (Metric for Evaluation of Translation with Explicit Ordering)[3]	Machine translation, text generation	Measures semantic matches based on exact word, stemmed word, and synonym matching.	**Strengths:** Incorporates synonyms and stemming for better matching, shows stronger correlation with human judgement than BLEU in single-reference tasks. **Limitations:** Requires external resources (e.g., WordNet); still struggles with diverse paraphrasing and deep semantic understanding.
BERTScore[4]	Translation, dialogue, summarisation	Measures semantic similarity by computing cosine similarity between BERT embeddings of tokens in the candidate and reference texts.	**Strengths:** Captures semantic similarity using deep embeddings, generally shows high correlation with human judgement. **Limitations:** Heavily relies on the quality of the underlying BERT model; is not fully faithful to token positions or syntactic structure.
MoverScore[5]	Text generation, paraphrasing	Measures the minimum "cost" to transform the generated text's word embeddings into the reference text's embeddings.	**Strengths:** Excellent at handling paraphrasing and capturing sentence-level semantic meaning due to its grounding in Word Mover's Distance. **Limitations:** Computationally more expensive than BERTScore; results are highly sensitive to the pre-trained word embeddings used.
Sentence Mover's Similarity[6]	Text generation, summarisation, paraphrasing	Measures the distance between generated and reference text by comparing sentence embeddings, adapting Word Mover's Distance to the sentence level.	**Strengths:** Effectively captures semantic similarity at the sentence level, providing robustness against small, local word changes. **Limitations:** Computationally expensive due to the underlying transport mechanism; heavily dependent on the quality of the sentence embeddings used.
Perplexity[7] & cross-entropy[8]	Language modelling, fluency tracking	Measures the predictability or uncertainty of a language model over a sample of text (its inverse probability).	**Strengths:** Effective for internal model comparison, rapid optimization, and confirming model fluency during training/fine-tuning. **Limitations:** Valuable only during training/fine-tuning; does not align well with task-specific quality (e.g., factual accuracy, coherence).

Metric	Application	Method	Strengths/ Limitations
F1 Score (Token/ Span Level)[9]	QA, named entity recognition (NER)	Harmonic mean of precision and recall, often calculated on the overlap of tokens/ spans between generated output and reference answer.	**Strengths:** Highly relevant for extracting specific information (spans); balanced measure that rewards both correct inclusion (recall) and avoiding irrelevant content (precision). **Limitations:** Sensitive to small errors in boundary detection; does not capture semantic similarity or fluency, only exact token overlap.
Levenshtein Distance[10]	Spelling correction, speech recognition	Measures the minimum number of single-character edits (insertions, deletions, substitutions) required to change the generated text into the reference text.	**Strengths:** Provides a simple, highly interpretable, and precise measure of character-level error rate and textual proximity. **Limitations:** Extremely strict; does not tolerate paraphrasing, and a single character error can be disproportionately penalized in short strings.



Reference-Free Text Metrics

When reference are unavailable, reference-free metrics become important.
These metrics evaluate the quality of generated content without reliance on a ground-truth reference, often by comparing the output to the source document or analysing the statistical properties of the output itself. A summary of most used metrics can be found in Table 2

Table 2 Summary of Reference-Free Text Metrics

Metric	Application	Method	Strengths/ Limitations
SUPERT[11]	Abstractive summarisation	Measures the semantic overlap between salient parts of the generated summary and the source document.	**Strengths:** Focuses on semantic faithfulness to salient points of the source, useful for evaluating highly abstractive tasks. **Limitations:** Requires accurate identification of salient sentences; less interpretable than lexical metrics.
BLANC[12]	Summarisation	Measures the informativeness and coverage of the summary using a Masked Language Model (MLM) to compare information content between source and summary.	**Strengths:** Highly correlated with human judgements of informativeness; model-based, capturing deep semantic content and coverage. **Limitations:** Computationally intensive due to the multiple masking and prediction steps involved; relies on the performance of the underlying language model.
Rough-C[13]	Summarisation	Measures the extent to which the content units (e.g., -grams) in the source document are represented in the generated summary.	**Strengths:** Simple, effective for quickly assessing coverage and information transfer from source to summary. **Limitations:** Purely lexical and shallow; does not measure fluency, coherence, or semantic equivalence of paraphrases.
Distinct-N[14]	Open-ended generation, dialogue	Measures the number of unique n-grams (typically unigrams and bigrams) in the generated corpus, assessing diversity.	**Strengths:** Simple, effective, and fast measure of generation diversity and fluency in open-ended and conversational tasks. **Limitations:** High diversity does not guarantee coherence, relevance, or quality; sensitive to corpus size.
Self-BLEU[15]	Open-ended generation	Measures the stability of a model's output by comparing one generated response against all others in the batch.	**Strengths:** Directly quantifies the repetitiveness and lack of diversity across a model's output corpus. **Limitations:** Only indicates how repetitive the model is; does not assess semantic quality or fidelity to the prompt.

Metric	Application	Method	Strengths/ Limitations
SummC (Summarisation consistency)[16]	Summarisation, factuality	Employs a Natural Language Inference (NLI) model to classify if sentences in the summary are entailed by the source document.	**Strengths:** Directly targets factual consistency (non-hallucination), which is critical for trustworthy text generation. **Limitations:** Accuracy is limited by the underlying NLI model's performance; struggles with complex or nuanced facts and knowledge inference.
FactCC (Factuality Correctness)[17]	Summarisation, Data-to-Text	Utilises a fine-tuned BERT model to classify if a generated statement is factually consistent with its source document.	**Strengths:** Directly assesses factual consistency and hallucination risk, providing a binary or probabilistic truth score. **Limitations:** Dependent on the training data and quality of the classifier; struggles with complex logical inferences.
DAE (Disentangled AE)[18]	Summarisation, factuality	Entailment-based metric that checks if generated text is logically supported (entailed) by the source document.	**Strengths:** Provides a robust, fine-grained check of logical entailment and faithfulness to the source text. **Limitations:** Requires training a dedicated entailment classifier; may mistakenly penalise valid paraphrases that are not strictly logically entailed.
SRLScore (Semantic Role Labelling Score)[19]	Summarisation, factuality	Measures factual consistency by comparing the structured semantic roles (predicates, arguments, agents) extracted from the generated text and the source.	**Strengths:** Offers fine-grained factuality assessment by focusing on "who did what to whom," providing structured verification. **Limitations:** Relies heavily on accurate SRL parsing, which can be noisy; more computationally complex and slower than pure sequence matching.
QA-FactEval (QA-based Factuality Evaluation)[20]	Summarisation, factuality	Measures the semantic overlap between salient parts of the generated summary and the source document.	**Strengths:** Highly interpretable and strong correlation with human judgement for factual correctness and faithfulness. **Limitations:** High computational cost due to the required question generation and answering stages; accuracy depends on the underlying QA model.

Metric	Application	Method	Strengths/ Limitations
QuestEval[21]	Summarisation, question answering (QA)	Measures the semantic overlap between salient parts of the generated summary and the source document.	**Strengths:** Focuses on semantic faithfulness to salient points of the source, useful for evaluating highly abstractive tasks. **Limitations:** Requires accurate identification of salient sentences; less interpretable than lexical metrics.
Task-Specific Metrics (e.g., Answer Relevancy)	Dialogue systems, code generation, controlled output	Automated evaluation of adherence to specific rules (e.g., must output JSON), user intent satisfaction, or factual relevance to the prompt.	**Strengths:** Direct measure of performance on the specific goal of the task, essential for goal-oriented systems. **Limitations:** Requires custom definition and implementation for every task; may not generalize well across different domains.

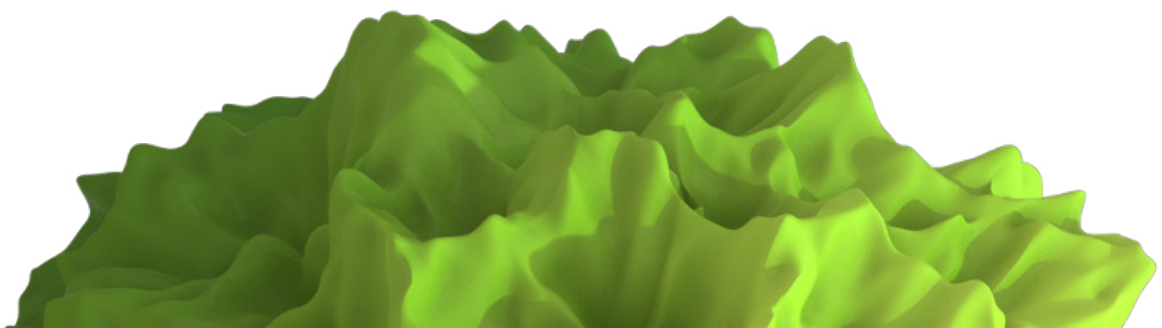
Image-Based Metrics

As GenAI expands beyond text, evaluation of image outputs presents distinct challenges, requiring metrics that capture not only low-level fidelity but also perceptual quality and semantic consistency. These metrics address the challenges of evaluating visual outputs, focusing on low-level fidelity (pixel-based) and high-level realism and diversity (feature-based).

Table 3: Summary of common Image-Based Metrics

Metric	Application	Method	Strengths/ Limitations
PSNR (Peak Signal-to-Noise Ratio)[26]	Image restoration, denoising	Measures the ratio between the maximum possible power of a signal and the power of corrupting noise (pixel difference).	**Strengths:** Fast and simple calculation, effective for measuring low-level reconstruction accuracy and fidelity in tasks like super-resolution. **Limitations:** Excellent for fidelity, but highly sensitive to small shifts; does not correlate well with human perceptual quality.
SSIM (Structural Similarity Index Measure)[22]	Image quality assessment	Measures the perceived change in structural information (luminance, contrast, and structure).	**Strengths:** Measures structural change and shows better correlation with human perception of image quality than PSNR. **Limitations:** Better than PSNR for perceptual quality but still a low-level metric; requires pixel-aligned reference images.

Metric	Application	Method	Strengths/ Limitations
LPIPS (Learned Perceptual Image Patch Similarity)[23]	Style Transfer, Image-to-Image	Measures the distance between feature representations extracted from a pre-trained deep neural network.	**Strengths:** Excellent correlation with human judgement for perceptual similarity, widely used for image style and fidelity tasks. **Limitations:** Correlates well with human judgements; however, the score is dependent on the architecture and quality of the pre-trained feature extractor.
FID (Fréchet Inception Distance)[24]	Generative adversarial networks (GANs), diffusion models	Measures the statistical distance between the feature distribution of real images and generated images using the network.	**Strengths:** Considered the gold standard for measuring realism and diversity (distribution quality) of generated image sets. **Limitations:** Computationally expensive; highly sensitive to the size and quality of the generated and real image samples (sample size bias).
Inception Score (IS)	Generative adversarial networks (GANs)	Measures the quality (image clarity) and diversity (entropy of label predictions) of generated images.	**Strengths:** Fast calculation, provides an aggregate score of both clarity and class diversity in a single number. **Limitations:** Highly sensitive to the "inception" network's ability to classify objects; tends to be maximised by models that generate visually sharp, but not necessarily diverse, images.
Precision and Recall (Generative Models)[25] Learned Aesthetic Scores[26]	Open-ended image synthesis	Measures the diversity (Recall) and realism (Precision) of the generated image set relative to the real data distribution.	**Strengths:** Provides separate, interpretable measures for mode coverage (recall) and fidelity (precision) of GAN/ Diffusion models. **Limitations:** Requires defining a reliable, high-dimensional feature space; calculation can be complex and sensitive to hyperparameters.



Model-Assisted Evaluation (LLM-as-a-Judge)

More recently, LLM-based evaluation methods have been introduced, where an LLM itself is used as judge. These approaches rely on prompting an LLM to assess generated text. Several prompting frameworks have been proposed such as Reason-then-Score (RTS), MCQ Scoring, Head-to-Head Scoring[27], which can be used to systematically evaluate the quality, accuracy and preference of model outputs across tasks such as summarisation, translation, and question answering.

Other task-specific frameworks have also been developed: GEMBA[28] is a GPT-based metric designed for translation quality assessment, while G-Eval[29] leverages Chain-of-Thought prompting (CoT[30]) to systematically evaluate summarisation outputs. These LLM-as-a-judge methods[31] show promise in capturing semantic nuance and task-specific criteria more effectively than traditional metrics. However, challenges include its resource intensiveness, a lack of standardisation across different models and datasets, and the potential for biases in the LLMs rate themselves (e.g., towards verbosity, position, or specific styles). Human oversight remains necessary to calibrate and validate the AI judge's performance, as the LLM-as-a-judge can still suffer from the systematic bias, "hallucinated" reasoning, and a lack of nuanced context that only a human can calibrate to ensure the judge aligns with real-world intent rather than just statistical patterns.

Qualitative Assessment and Human Evaluation:

While automated quantitative metrics provide valuable objective measures of GenAI performance, they often fall short in capturing the nuanced aspects of output quality. Qualitative assessment complements quantitative evaluation by examining characteristics such as creativity, coherence, style, relevance, and user experience. This type of assessment often involves expert review, thematic analysis, or content evaluation, where outputs are examined for attributes that are difficult to measure numerically. It can also reveal subtle strengths and weaknesses that quantitative metrics might miss, like the narrative flow in generated text or the aesthetic appeal in generated images.

Human evaluation is a cornerstone of qualitative assessment, providing structured judgements on model outputs. Key approaches include:

1. Expert Review

Engaging domain specialists or editors to provide in-depth feedback on linguistic quality, technical accuracy, readability, and professional tone. This is crucial for identifying hidden errors and biases that automated systems often overlook.

2. User Studies

Involving independent groups of users to rate the coherence, fluency, and comprehensibility of AI outputs, often using structured scales (e.g., Likert scales). These studies gauge practical utility and user satisfaction in real-world scenarios.

3. Human-in-the-Loop (HITL)

A collaborative strategy where human expertise is integrated throughout the AI system's lifecycle. Humans actively participate by providing feedback, evaluating performance, and offering critical insights. This continuous involvement enhances accuracy, mitigates bias, improves transparency, builds user trust, and ensures ongoing adaptation and improvement of the AI system, providing contextual and ethical judgement that machines cannot replicate. Common techniques within these approaches include rating outputs along defined dimensions (e.g., clarity, informativeness), pairwise comparisons to benchmark improvements, and task-specific assessments to measure practical utility. While human evaluation is invaluable, it has limitations. It is time-consuming, costly, and subject to variability due to evaluator bias or inconsistency. Structured guidelines, multiple evaluators, and aggregation of scores are commonly employed to improve reliability.

Beyond Output Metrics: Technical and Operational Validation Dimensions

While evaluating the quality of generated outputs is important, extending the validation to their intrinsic technical characteristics and operational performance is essential. This delves into the model's underlying design, behaviour, and deployment readiness, encompassing dimensions such as architecture, robustness, throughput, explainability, consistency, and bias and fairness.

GenAI Architecture Assessment

Architecture evaluation focuses on the model's fundamental design, size, and computational footprint, involving profiling metrics such as parameter count, memory usage, and operational efficiency. These provide objective insights into resource demands, scalability, and deployment feasibility, guiding decisions on infrastructure and cost.

Robustness and Consistency:

Robustness assesses the model's ability to maintain stable performance and avoid errors when faced with unexpected, noisy, or adversarial inputs, or shifts in data distribution. This is often tested through stress testing and adversarial attacks to uncover vulnerabilities. **Consistency,** on the other hand, evaluates whether the model produces stable and coherent outputs when given repeated or semantically similar inputs. High consistency is vital for tasks requiring reliable and repeatable results, such as dialogue or summarisation. Together, these evaluations ensure the model behaves dependably across various conditions and over time, though neither guarantees factual accuracy nor covers all real-world complexities.

Throughput and Efficiency

This evaluates the model's inference speed and operational efficiency, typically measured by latency and batch processing performance. High throughput is vital for interactive or high-volume applications, directly impacting user experience and system scalability. It provides a clear indication of operational efficiency, though its measurement is heavily dependent on hardware and the specific deployment environment, which can limit generalisability.

Explainability (XAI):

Explainability examines the degree to which a model's reasoning or generation process can be interpreted and understood by humans. Techniques such as attention visualisation, feature attribution (e.g. SHAP[32] and LIME[33]), and LLM-generated rationales (e.g., Chain-of-Thought prompting) are employed. These methods foster transparency, build trust, and aid in regulatory compliance and debugging. However, explanations can sometimes be incomplete, approximate, or even misleading, requiring careful interpretation.

Bias and Fairness:

This critical dimension involves analysing model outputs across various demographic groups or sensitive attributes to identify and mitigate unfair or discriminatory outcomes. Validation employs fairness metrics, scenario-based testing, and targeted evaluations. These assessments are essential for ethical deployment and regulatory adherence. However, fairness metrics are inherently incomplete, sensitive to dataset biases, and cannot capture the full spectrum of societal or ethical considerations.

Security:

Beyond robustness, security validation specifically addresses potential vulnerabilities to malicious attacks, data breaches, or intellectual property theft. This includes evaluating resistance to prompt injection, data exfiltration, and model inversion attacks. Ensuring the model's integrity and the confidentiality of its training data and outputs is paramount, especially in sensitive corporate environments.

Future Directions and Emerging Innovations

This section outlines key advancements and future directions in GenAI model validation, focusing on robust methodologies for benchmarking, human oversight, and continuous monitoring to build trust and mitigate risks.

Development of Unified Benchmarks

Establishing robust benchmarks, that are specific to the use case and evaluating performance under diverse real-world conditions are critical for ensuring the reliability and generalisability of the GenAI model. Without rigorous benchmarking, models risk performing well only in limited, artificial scenarios, leading to inaccurate or unreliable real-world results. This section details the development of comprehensive benchmarks and testing methodologies to assess model robustness and resilience.

HELM (Holistic Evaluation of Language Models):

The Holistic Evaluation of Language Models (HELM) project[34] introduces a novel, multi-faceted evaluation framework designed to provide a more complete and nuanced understanding of LLM capabilities and limitations, primely on textual data. HELM is a comprehensive benchmark evaluating 30 prominent language models across 42 diverse scenarios using seven key metrics (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency). Its novel scenarios, including those for trustworthiness and neglected dialects, offer a more holistic assessment. Publicly available data and tools foster community involvement, creating a "living benchmark" that adapts to LLM advancements. While primarily focused on textual data, HELM significantly advances LLM evaluation.

UniEval

UniEval provides a comprehensive framework specifically designed for unified multimodal models[35]. Unlike existing task-specific benchmarks, UniEval uses a single benchmark (UniBench) and metric (UniScore) without requiring extra models, images, or annotations. UniBench offers high diversity with 81 fine-grained tags and is more challenging than existing benchmarks, with UniScore strongly correlating with human judgment.

LightEval

LightEval[36] addresses the need for a flexible, accessible evaluation toolkit for a broader range of LLMs, including unimodal models. This open-source, Hugging Face-developed framework offers a unified toolkit for performance assessment across various benchmarks and settings, integrating seamlessly with the Hugging Face ecosystem.

Overall, HELM offers a comprehensive, multi-faceted approach, while UniEval and LightEval provide more specialised and streamlined evaluations. The ongoing development of these benchmarks is crucial for fostering transparency, enabling fair comparisons, and accelerating the development of reliable AI systems.

Automated and Continuous Validation Frameworks

A well-defined validation framework and continuous monitoring are essential for maintaining the long-term reliability and trustworthy operation of GenAI models. This includes clear procedures for ongoing assessment, adaptation, and addressing emerging issues, allowing for proactive risk identification and mitigation.

The unique characteristics of Gen AI necessitate the development of comprehensive and adaptable validation frameworks. Current tools like Encord Active, Deepchecks, and HoneyHive.ai offer capabilities for model testing, performance evaluation, and bias assessment[37]. Future innovations will likely expand these automated evaluation tools, reducing reliance on human judgment while enhancing efficiency and consistency[38]. Post-deployment, continuous monitoring will track performance, identify incidents, and trigger swift responses to harmful or inappropriate outputs, ensuring ongoing quality and safety in real-world conditions[39].

Self-Improving AI Validation Systems

A significant future direction involves the development of self-improving AI agents that can autonomously learn from feedback and experiences, refining their knowledge or skills without needing human engineers to rewrite code[40]. Recent research demonstrates that large language models can effectively self-improve through self-judging, leveraging the inherent asymmetry between generating and verifying solutions[41]. This involves models generating practice problems, solving them, and evaluating their own performance without external validation. This approach suggests a potential paradigm shift towards AI systems that continuously improve through self-directed learning, potentially accelerating progress in domains with scarce training data or complex evaluation requirements

Advancements in Explainable AI (XAI) for GenAI

Explainable AI (XAI) techniques are crucial for bridging the gap between complex AI models and human understanding, fostering trust and usability. Given the "black-box" nature of many GenAI models, which makes their decision logic non-transparent, XAI aims to provide insights into how specific inputs influence outputs and the reasoning behind generated content[42]. Future advancements will focus on developing GenXAI techniques that produce verifiable explanations, enhancing understanding and knowledge acquisition from data[43]. For increasingly complex GenAI systems, particularly Multimodal Large Language Models (MLLMs), Multimodal eXplainable AI (MXAI) is developing. MXAI integrates multiple modalities (e.g., image-text, audio-text) for prediction and explanation task[44]. MXAI aim to validate that outputs are trustworthy, adapted to specific project conventions, and capable of providing key insights into the intrinsic interpretability of the model's reasoning.

Conclusion

The rapid advancement of generative AI models has created unprecedented opportunities across a wide range of industries. However, it also introduces crucial challenges in validating these models effectively. In this paper, we outlined the key challenges encountered in validating generated AI models, spanning performance and evaluation, transparency, deployment and maintenance, as well as regulatory and ethical compliance. We also reviewed current industry-standard validation approaches used to mitigate these challenges, highlighting both their strengths and significant limitations. Moreover, we explore future directions and emerging innovations for advanced validation approaches, focusing on benchmarking and robustness, validation framework and monitoring, human oversight and ethical consideration, which aim to support and improve validation of generative AI models.

The implications of our findings reveal that generative AI model validation can no longer be treated as static, post-hoc process. Instead, it should evolve in a continuous, multi-dimensional discipline integrated throughout the model lifecycle, from design and training to deployment and monitoring. In addition, developing a generalised validation framework that can be tailored and adapted to use-case-specific context is essential. Furthermore, validation should be consistent and objective, focusing not only on the model itself but also on the surrounding components integrated within the system. It is worth noting that GenAI model validation is not just about technical checks. It also involves ensuring that these systems align with human goals, are safe to use, behave responsibly, and comply with evolving regulatory standards.

Get in touch

Richard Tedder
Partner
AI Assurance
rtedder@deloitte.co.uk

Avtar Benning
Director
AI Assurance
abenning@deloitte.co.uk

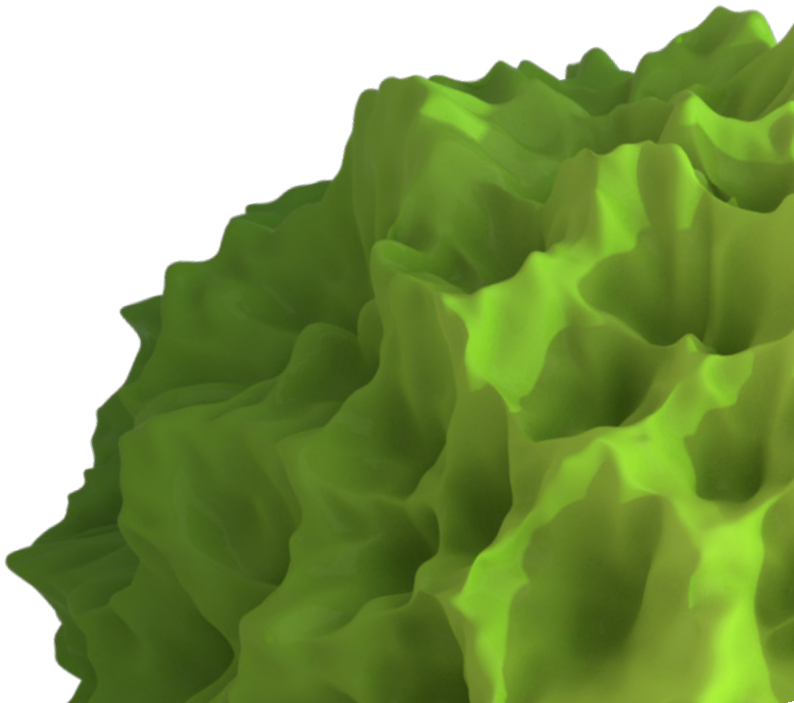
Ala Alfakara
Associate Director
AI Assurance
aalfakara@deloitte.co.uk

References

1. Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., "Bleu: a method for automatic evaluation of machine translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, 2002.
2. Lin, C.Y., "Rouge: A package for automatic evaluation of summaries," Text Summarization Branches Out, pp. 74-81, 2004.
3. Banerjee, S. and Lavie, A., "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," Proceedings of the ACL Workshop on Iintrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65-72, 2005.
4. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y., "Bertscore: Evaluating text generation with berT," arXiv preprint arXiv:1904.09675., 2019.
5. Zhao, W., Peyrard, M., Liu, F., Gao, Y. and Meyer, C., "MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance," arXiv preprint arXiv:1909.02622, 2019.
6. Clark, E., Celikyilmaz, A. and Smith, N. A., "Sentence mover's similarity: Automatic evaluation for multi-sentence texts," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2748-2760, 2019.
7. Jelinek, F., Mercer, R. L., Bahl, L. R. and Baker, J., "Perplexity—a measure of the difficulty of speech recognition tasks," The Journal of the Acoustical Society of America, vol. 62(S1), pp. S63-S63, 1977.
8. Xu, C., Zhu, Z., Wang, J., Wang, J. and Zhang, W., "Understanding the role of cross-entropy loss in fairly evaluating large language model-based recommendation," arXiv preprint arXiv:2402.06216, 2024.
9. T. T. Vu, "Understanding LLM Evaluation Metrics: Best Practices for Reliable LLM Assessment," Medium, 12 May 2025. [Online]. Available: https://medium.com/@tungvu_37498/understanding-llm-evaluation-metrics-best-practices-for-reliable-llm-assessment-3fce1fa48251. [Accessed 2026].
10. Microsoft, "A list of metrics for evaluating LLM-generated content," Microsoft, 2025. [Online]. Available: https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics. [Accessed 2026].
11. Gao, Y., Zhao, W. and Eger, S., "SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization," arXiv preprint arXiv:2005.03724, 2020.
12. Vasilyev, O., Dharnidharka, V. and Bohannon, J., "Fill in the BLANC: Human-free quality estimation of document summaries," arXiv preprint arXiv:2002.09836, 2020.
13. Sai, A. B., Mohankumar, A. K. and Khapra, M. M., "A survey of evaluation metrics used for nlg systems," vol. 55(2), pp. 1-39, 2022.
14. Li, J., Galley, M., Brockett, C., Gao, J. and Dola, B., "A diversity-promoting objective function for neural conversation models," arXiv preprint arXiv:1510.03055, 2015.
15. Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J. and Yu, Y., "Texygen: A benchmarking platform for text generation models," The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1097-1100, 2018.

16. Laban, P., Schnabel, T., Bennett, P. N. and Hearst, M. A., "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization," Transactions of the Association for Computational Linguistics, vol. 10, pp. 163-177, 2022.
17. Kryściński, W., McCann, B., Xiong, C. and Socher, R., "Evaluating the factual consistency of abstractive text summarization," arXiv preprint arXiv:1910.12840, 2019.
18. Goyal, T. and Durrett, G., "Evaluating factuality in generation with dependency-level entailment," arXiv preprint arXiv:2010.05478, 2020.
19. Fan, J., Aumiller, D. and Gertz, M., "Evaluating factual consistency of texts with semantic role labeling," arXiv preprint arXiv:2305.13309, 2023.
20. Fabbri, A. R., Wu, C. S., Liu, W. and Xiong, C., "QAFactEval: Improved QA-based factual consistency evaluation for summarization," arXiv preprint arXiv:2112.08542, 2021.
21. Scialom, T., Dray, P. A., Gallinari, P., Lamprier, S., Piwowarski, B., Staiano, J. and Wang, A., "QuestEval: Summarization asks for fact-based evaluation," arXiv preprint arXiv:2103.12693, 2021.
22. Hore, A. and Ziou, D., "August. Image quality metrics: PSNR vs. SSIM," 20th International Conference on Pattern Recognition, pp. 2366-2369, 2010.
23. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. and Wang, O., "The unreasonable effectiveness of deep features as a perceptual metric," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586-595, 2018.
24. Yu, Y., Zhang, W. and Deng, Y., "Frechet inception distance (fid) for evaluating gans," China University of Mining Technology Beijing Graduate School, vol. 3(11), 2021.
25. Sajjadi, M. S., Bachem, O., Lucic, M. and Bousquet, O., "Assessing generative models via precision and recall," Advances in Neural Information Processing Systems, vol. 31, 2018.
26. Lu, X., Lin, Z., Jin, H., Yang, J. and Wang, J. Z., "Rating image aesthetics using deep learning," IEEE Transactions on Multimedia, vol. 17(11), pp. 2021-2034, 2015.
27. Shen, C., Cheng, L., Nguyen, X. P., You, Y. and Bing. L., "Large language models are not yet human-level evaluators for abstractive summarization," arXiv preprint arXiv:2305.13091, 2023.
28. Kocmi, T. and Federmann, C., "Large language models are state-of-the-art evaluators of translation quality," arXiv preprint arXiv:2302.14520, 2023.
29. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C., "G-Eval: NLG evaluation using GPT-4 with better human alignment," arXiv preprint arXiv:2303.16634, vol. 12, 2023.
30. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V. and Zhou, D., "Chain-of-thought prompting elicits reasoning in large language models," Advances in Neural Information Processing Systems, vol. 35, pp. 24824-24837, 2022.
31. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H. and Wang, S., "A survey on llm-as-a-judge," arXiv preprint arXiv:2411.15594, 2024.

32. Nohara, Y., Matsumoto, K., Soejima, H. and Nakashima. N., "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," Computer Methods and Programs in Biomedicine, vol. 214, p. 106584, 2022.
33. Zafar, M. R. and Khan, N., "Deterministic local interpretable model-agnostic explanations for stable explainability," Machine Learning and Knowledge Extraction, vol. 3(3), pp. 525-541, 2021.
34. Bommasani, R., Liang, P. and Lee, T., "Holistic evaluation of language models," Annals of the New York Academy of Sciences, vol. 1525(1), pp. 140-146, 2023.
35. Li, Y., Wang, H., Zhang, Q., Xiao, B., Hu, C., Wang, H. and Li, X., "Unieval: Unified holistic evaluation for unified multimodal understanding and generation," arXiv preprint arXiv:2505.10483, 2025.
36. "LightEval Deep Dive: Hugging Face's All-in-One Framework for LLM Evaluation," [Online].
37. S. Oladele, "encord," 2024. [Online]. Available: https://encord.com/blog/model-validation-tools/.
38. L. Ramamoorthy, "Evaluating Generative AI: Challenges, Methods, and Future Directions," International Journal For Multidisciplinary Research, 2025.
39. P. Walsh, "Validating Generative AI: A Practical Framework For Reliability And Compliance in Life Sciences," Clincal Leader, 2025. [Online]. Available: https://www.clinicalleader.com/doc/validating-generative-ai-a-practical-framework-for-reliability-and-compliance-in-life-sciences-0001.
40. J, "Self-Improving Data Agents: Unlocking Autonomous Learning and Adaptation," Powerdrill, 2025. [Online]. Available: https://powerdrill.ai/blog/self-improving-data-agents.
41. T. Simonds, K. Lopez, A. Yoshiyama and D. Garmier, "Self Rewarding Self Improving," arXiv, 2025.
42. F. Mumuni and A. Mumuni, "Explainable artificial intelligence (XAI): from inherent explainability to large language models," arXiv, 2025.
43. J. Schneider, "Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda," arXiv, 2024.
44. W. A. F. T. e. a. Shilin Sun, "A Review of Multimodal Explainable Artificial Intelligence: Past, Present and Future," arXiv, 2024.



Deloitte.

Designed and produced by Creative Studio at Deloitte. RITM2547599

MAKING AN IMPACT THAT MATTERS since 1845