Guide to Generative AI Evaluation

February 2025

pwc

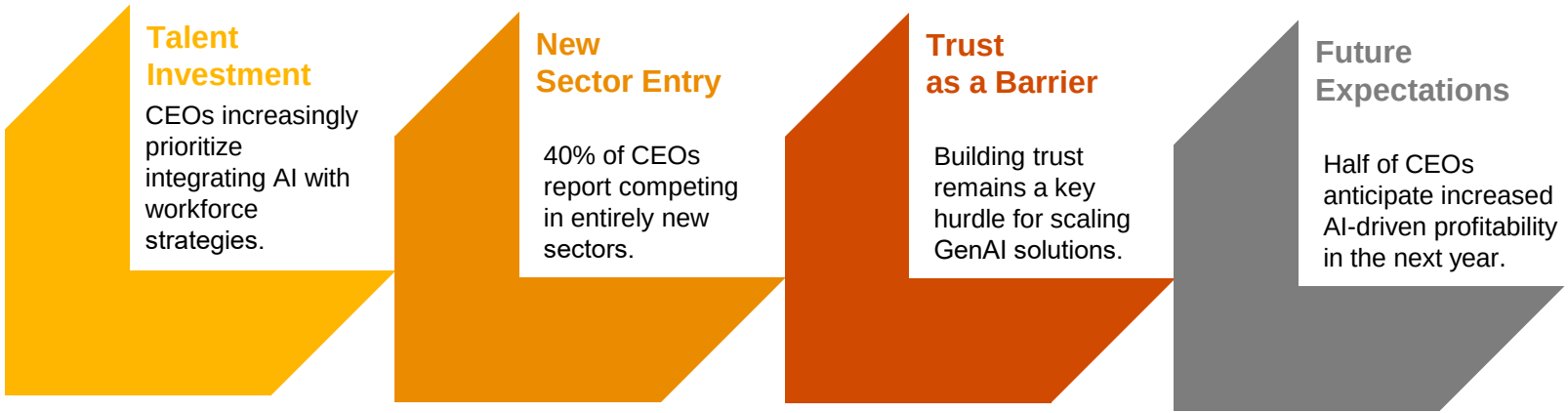
Agenda

1

Context

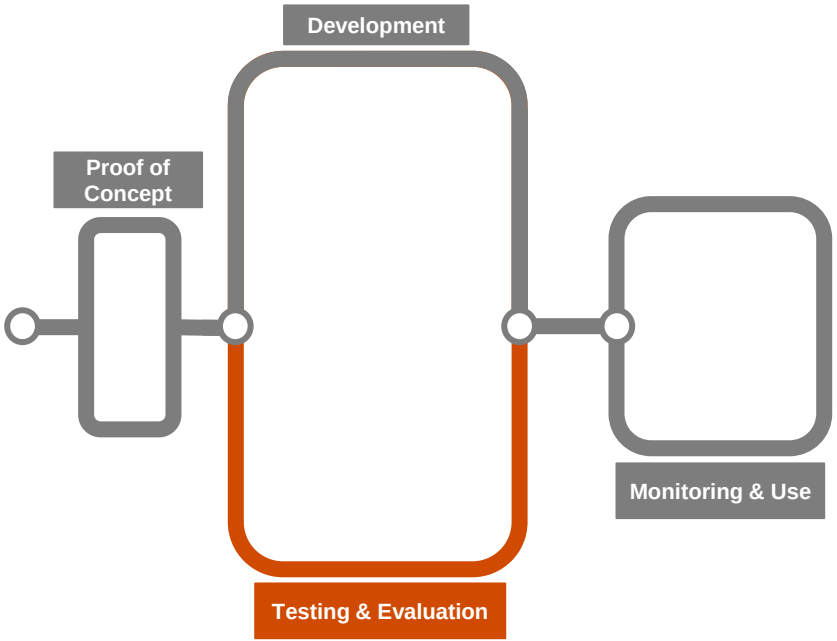
Context
The Rise of AI in Business Operations

AI technologies, including generative AI (GenAI), are revolutionizing business operations globally.
Companies are leveraging AI to enhance workforce **productivity and customer experiences, increase revenue, and to stay competitive** in evolving markets. Notably, **56% of CEOs report efficiency gains** from GenAI, with 34% seeing increased profitability.



To maximize AI's potential, businesses also need to adopt **advanced evaluation methods**, such as automated metrics for quality assessment and business impact measurement, to drive sustained innovation and value creation.

The Importance of Evaluating GenAI Outputs



Evaluation is a crucial phase in the deployment of AI solutions. For AI-driven automation to effectively assist companies in reaching their objectives, it must consistently produce **high-quality outputs**. Achieving this requires **thorough testing and evaluation** before implementing the AI solution. However, evaluating the outputs of Generative AI (GenAI) presents **unique challenges** not encountered in traditional software development.

2

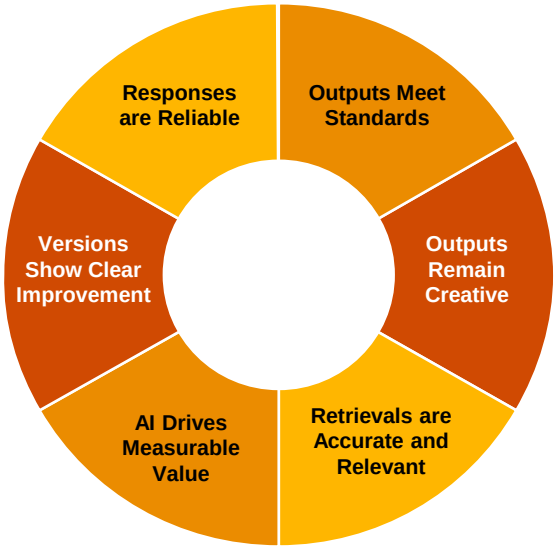
Typical Challenges

Typical Challenges of GenAI Evaluation

Response Consistency
- AI responses can vary across similar queries, leading to inconsistency and loss of trust in system reliability.
- Lack of consistent outputs can impact business workflows that rely on AI-generated content.

Version Benchmarking
- Comparing different AI versions to track improvements often lacks a structured approach
- Stakeholders struggle to gauge whether newer versions genuinely outperform older ones.
- Without proper benchmarks, development decisions risk being based on assumptions.

Business Impact
- Quantifying the real-world impact of AI systems such as ROI, cost savings, and productivity gains is challenging without clear metrics.
- This creates difficulty in justifying ongoing AI investments to decision-makers.



Quality Assessment
- Evaluating whether AI responses meet critical quality benchmarks (accuracy, coherence, and relevance) is complex.
- Without structured assessments, organizations risk deploying subpar AI systems that fail to meet user expectations.

Novelty and Diversity
- AI models risk producing repetitive and stale outputs over time, particularly in creative tasks.
- Overfitting to common patterns in training data limits innovation.
- Striking the balance between diversity and coherence remains a persistent challenge for GenAI.

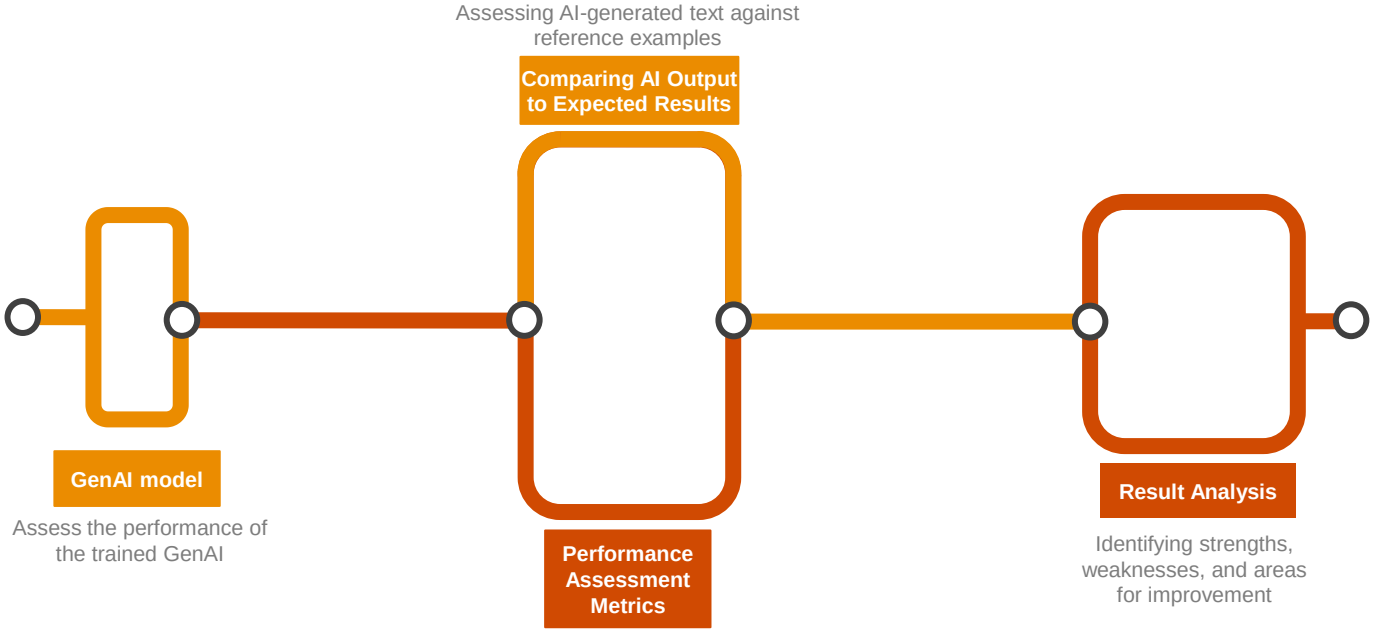
RAG-Specific Evaluation
- Maintaining context relevance is a persistent challenge, especially when the system retrieves unrelated or incomplete information.
- Maintaining answer faithfulness to retrieved content while ensuring user relevance is difficult.

3

Solution

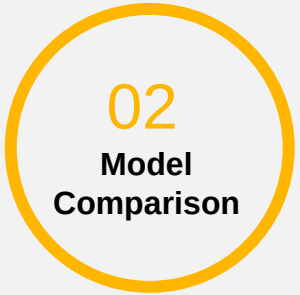
Evaluating GenAI Performance

Our focus is on applying rigorous evaluation metrics to assess GenAI performance, ensuring reliability, accuracy, and business impact.



Benefits of GenAI Evaluation

Manual evaluation is inefficient and prone to errors; automated measurement ensures accuracy and consistency while saving time, reducing costs, and enhancing evaluation precision.



Conduct thorough benchmarking to assess performance variations across different GenAI models, enabling data-driven selection and optimization.

Continuously assess whether model outputs maintain their relevance and accuracy following updates or version changes, ensuring consistent performance and alignment with business needs.

Implement robust evaluation frameworks to assess model performance, enabling early detection of issues and continuous improvement of AI solutions.

Leverage evaluation insights to make informed decisions about the performance, accuracy, and relevance of your AI models, ensuring they meet evolving business objectives.

4

Text Quality and Similarity Assessment

Text Quality and Similarity Assessment

1 BLEU Score

Measures n-gram overlap between generated and reference texts, focusing on precision-based similarity. This makes it particularly valuable for machine translation where exact matches of phrases indicate higher quality. Its ability to evaluate different n-gram sizes helps assess both local and broader textual consistency.

BLEU Evaluation

Reference texts

The liquidity report has been reconciled and published.

Generated texts

High BLEU Score:
The liquidity report was reconciled and published.
(Precise match of n-grams, phrase structure maintained)

Low BLEU Score:
Financial reports were finalized today.
(Significant deviation from reference, few n-gram overlaps)

2 ROUGE Score

ROUGE evaluates how well generated text captures content from reference text through recall-oriented measurement. It comes in multiple variants: ROUGE-N for n-gram overlap, ROUGE-L for longest common subsequence, and ROUGE-S for skip-gram co-occurrence.

ROUGE Evaluation

Reference texts

The liquidity report has been reconciled and published.

Generated texts

High ROUGE Score:
The reconciliation for liquidity has been completed and published.
(High recall: Retains key content despite changes)

Low ROUGE Score:
The finance team prepared reports today.
(Low recall: Only loosely related to the reference content)

3 METEOR

METEOR takes a more sophisticated approach by measuring text similarity through a weighted mean of precision and recall, while accounting for synonyms, stemming, and paraphrasing. Powerful for semantic evaluation, particularly in machine translation.

METEOR Evaluation

Reference texts

The liquidity report has been reconciled and published.

Generated texts

High METEOR Score:
The liquidity statement has been balanced and published.
(Recognizes synonyms for high match)

Low METEOR Score:
The system issued multiple financial reports.
(Low semantic overlap; different key ideas and terms.)

5

Diversity and Novelty Evaluation

Diversity and Novelty Evaluation

1 Self-BLEU

Self-BLEU measures diversity in generated text by calculating n-gram overlap between different outputs. Lower scores indicate greater variation, making it particularly valuable for evaluating creative writing, dialogue systems, and generative models where diverse responses are essential. High Self-BLEU suggests repetitive outputs, while low Self-BLEU reflects a broader range of expressions.

Self-BLEU Evaluation Example

Reference text

Loan approval requires credit evaluation and risk assessment.

Generated texts

Low Self-BLEU (High Diversity):
A thorough risk review and credit check determine loan eligibility.
(Varied phrasing, different word choices—indicating diverse generation.)

High Self-BLEU (Low Diversity):
Loan approval requires credit evaluation and risk assessment.
(Near-identical outputs—indicating low diversity)

2 Perplexity

Perplexity measures a language model's fluency and confidence by evaluating how well it predicts the next word in a sequence. Lower perplexity scores indicate more natural, coherent, and predictable text generation, while higher scores suggest uncertainty and less fluent outputs. Perplexity is widely used across various domains to compare model architectures and assess fluency.

Perplexity Evaluation Example

Reference text

Loan approval requires credit evaluation and risk assessment.

Generated texts

Low Perplexity (Fluent & Predictable):
Loan approval depends on assessing credit scores and evaluating risk factors.
(Natural, structured, and fluent—model is confident in its predictions.)

High Perplexity (Uncertain & Disfluent):
Approval credit risk needs loan for assessing require.
(Unnatural, disordered—model struggles with prediction.)

6

Business Impact Metrics

Business Impact Metrics

1 ROI Metrics

Provide a comprehensive view of business performance by measuring cost savings, time efficiency, and overall business value. These metrics are essential for making informed investment decisions and tracking performance over time. Through systematic analysis of resource utilization and financial returns, organizations can better justify projects and allocate resources effectively.

ROI Metrics Example

Scenario: A bank implements an AI-driven fraud detection system to reduce fraudulent transactions.

Investment: 2 million in AI infrastructure and model development.
Cost Savings: Reduced fraud losses by 5 million annually.

Time Efficiency: Manual fraud investigations decreased by 40%, freeing analysts for high-risk cases.

Business Value: Improved customer trust and regulatory compliance, reducing legal risks.

ROI Calculation:
ROI = (Savings - Investment) / Investment × 100 = (5M - 2M) / 2M × 100 = 150% ROI

2 User Satisfaction

Focus on understanding the real-world impact through user feedback and behavior patterns. By tracking feature adoption rates, usage patterns, and customer retention, these metrics provide valuable insights into product-market fit and guide feature prioritization. The combination of quantitative usage analytics and qualitative feedback creates a robust framework for measuring product success.

User Satisfaction Example

Scenario: A fintech app introduces a personalized budgeting feature.

Feature Adoption Rate: 70% of active users engage with the feature within the first month.

Usage Patterns: Users who engage with budgeting tools log in 3x more frequently.

Customer Retention: Churn rate among engaged users drops by 25%.
User Feedback: 85% of surveyed users report the feature helps them manage expenses better.

Business Impact:
The high adoption and engagement indicate strong product-market fit, while improved retention justifies further investment in personalization features.

7

RAG – Specific Evaluation

Core Quality Metrics in RAG Evaluation

RAG systems must be evaluated across three fundamental dimensions to ensure reliable performance.

1 Context Relevance

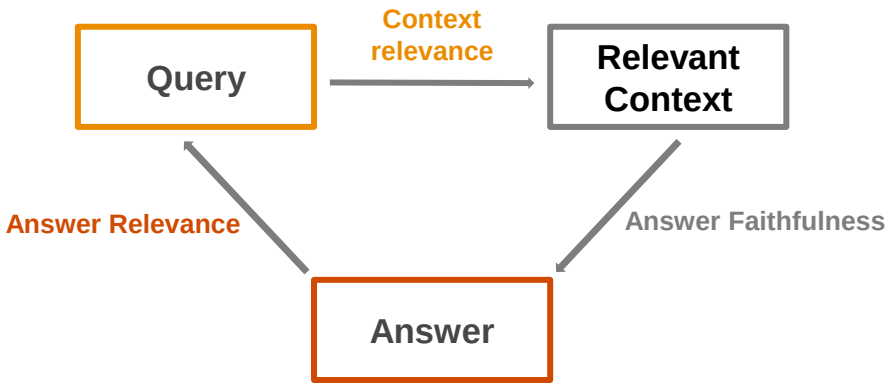
Measures how well the system retrieves appropriate information. This includes assessing the precision of retrieved context and its specificity to the query. While powerful for information retrieval assessment, it faces challenges with domain dependency and query ambiguity. Best practices involve comparing multiple retrieval methods and maintaining updated indices.

2 Answer Faithfulness

Examines the system's ability to generate responses that accurately reflect the retrieved context. This crucial metric ensures factual accuracy and proper source attribution, though it can be challenging to validate complex inferences across multiple sources. Implementing robust source tracking and cross-reference verification is essential.

3 Answer Relevance

Evaluates how well responses align with user queries and meet information needs. This metric focuses on response completeness and usefulness, though it must contend with subjective criteria and complex query handling. Success requires integrating user feedback and task-specific benchmarking.



RAG-Specific Evaluation

Required System Capabilities

For real-world deployment, RAG systems must demonstrate four essential abilities.

1 Noise Robustness

Systems must be capable of filtering out irrelevant or misleading information, ensuring they can focus on providing accurate responses. This is achieved through rigorous noise injection testing, which helps assess the model's ability to maintain relevant output in noisy or ambiguous contexts.

2 Negative Rejection

Effective systems should recognize when they lack the necessary knowledge to provide a reliable answer and reject such queries accordingly. This requires clear rejection criteria that guide the system in managing user expectations and avoiding the generation of incorrect or misleading responses.

Noise Robustness Example

Question

What is the largest planet in the solar system?

External documents contain noises

Jupiter is the largest planet in our solar system.

Earth is the third planet from the sun.

RAG ↓

Jupiter

Negative Rejection Example

Question

What is the currency of Japan?

External documents are all noises

The Euro is used in many European countries.

The US dollar is the currency of the United States.

RAG ↓

Insufficient information to answer

Required System Capabilities

Required System Capabilities

For real-world deployment, RAG systems must demonstrate four essential abilities.

3 Information Integration

The system must integrate information from diverse sources effectively, ensuring coherence and accuracy across the gathered data.
This involves using advanced algorithms to combine insights while minimizing conflicts between sources, enabling the generation of well-rounded responses.

4 Counterfactual Robustness

The system should identify and correct any misinformation in its outputs. Achieving this requires the use of source weighting mechanisms and regular updates to the fact database, supported by expert review, to ensure that responses remain grounded in verified and reliable information.

Information Integration Example

Question

What are the two main components of the Central Nervous System?

External documents contain noises

The brain is a primary component of the Central Nervous System.

The spinal cord is also a major component.

RAG ↓

Brain and spinal cord

Counterfactual Robustness Example

Question

What is SEPA in the context of European banking?

Counterfactual external documents

SEPA is a European agreement for free trade across borders.

SEPA primarily deals with the elimination of tariffs on goods.

RAG ↓

Factual errors. SEPA is the Single Euro Payments Area for euro transfers in EU

Industry Standard Benchmarks

Leading frameworks for standardized RAG evaluation include.

1 RAGAS Framework

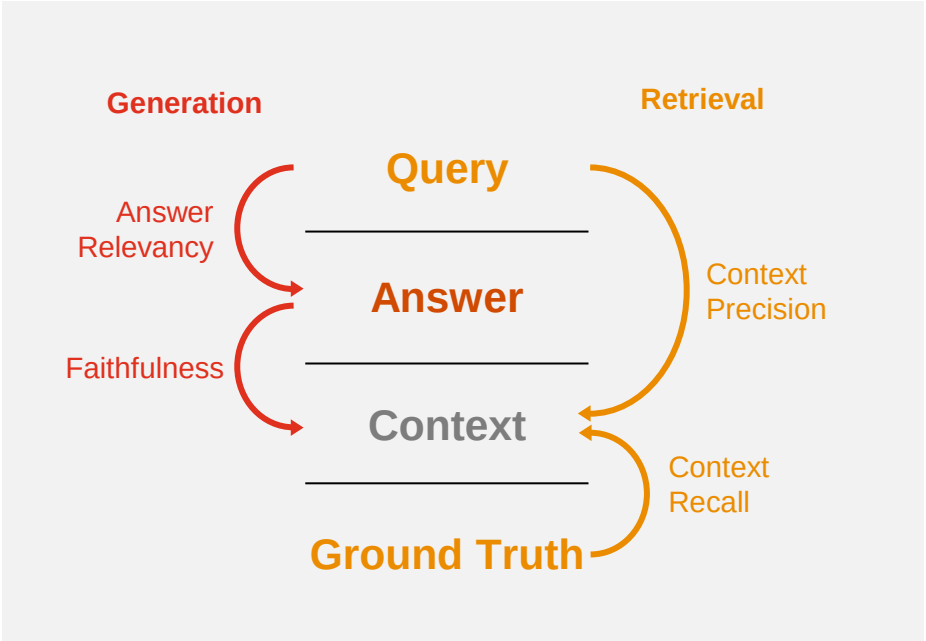
Serves as a comprehensive evaluation tool, combining automated LLM evaluation of faithfulness, relevance, and context quality. Its standardized scoring system makes it ideal for consistent performance tracking. The framework's integrated approach allows organizations to identify and address weaknesses across multiple dimensions of RAG system performance.

2 ARES Benchmark

Concentrates on evaluation through its comprehensive testing suite, emphasizing retrieval accuracy and answer generation quality. With multi-domain testing capabilities and standardized metrics, it provides robust performance assessment across different use cases and industries.

3 TruLens

Specializes in deeper verification aspects, focusing on truth verification, source attribution, and answer consistency validation. Its implementation features automated analysis systems, detailed ground truth comparisons, and continuous performance tracking mechanisms.



8

PwC Added Value

Meet the PwC Team Ready to Support You in AI Adoption

Understanding Industry Challenges

We recognize the key challenges companies face when adopting AI solutions. Our team focuses on practical and effective AI applications that address real-world needs and drive tangible business value.

Experienced Professionals

Our team consists of highly skilled experts with extensive experience in AI, data analytics, and risk management. We collaborate with businesses to implement AI strategies that deliver measurable impact.

Specialization in Banking & Insurance

We focus on leveraging AI in the banking and insurance sectors, helping organizations enhance efficiency, improve risk assessment, and create personalized customer experiences through advanced AI models.

Commitment to Innovation & Trust

Beyond technology, we prioritize transparency, trust, and responsible AI use. We work closely with our clients to ensure AI solutions align with regulatory standards, ethical considerations, and business goals.

9

Contacts

Interested?
Contact us.

Petr Novák
PwC Czech Republic
T: +420 602 383 972
M: petr.novak@pwc.com

Thank you!