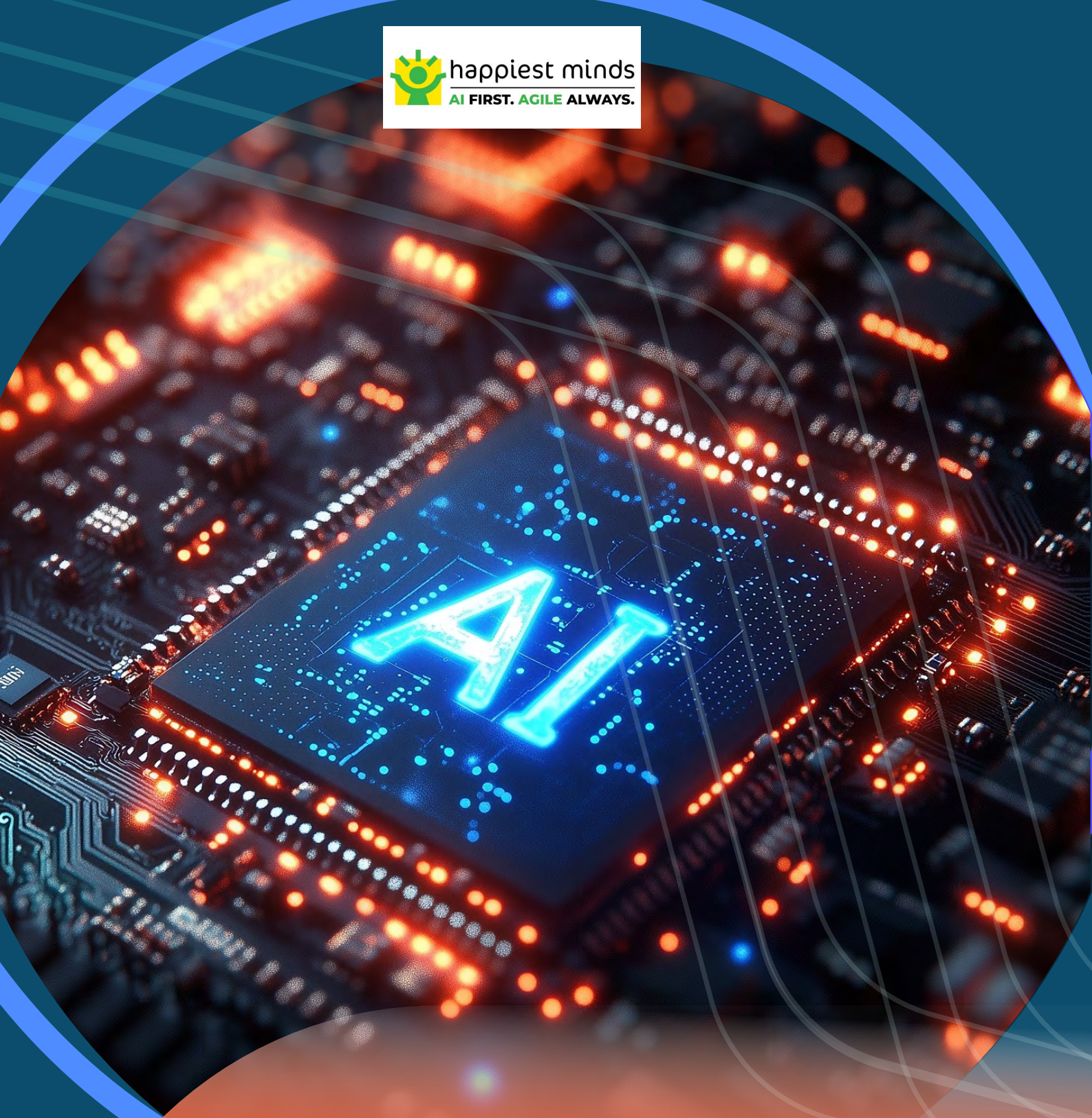
happiest minds
AI FIRST. AGILE ALWAYS.



Measuring GEN AI Success:

Evolving Beyond Traditional Quality Benchmarks for Modern Systems

Table of Contents

Introduction

AI Powered systems have undergone a profound transformation with the advent of RAG (Retrieval-Augmented Generation) systems. These systems harness the power of search engines and large language models (LLMs) to deliver improved answers. There are still significant challenges in measuring the quality of the responses. The traditional methods of evaluating these systems no longer fit into the new landscape.

It is not only a technical issue, but also makes life difficult for the teams that develop these systems. Systems that provide high scores in traditional tests fail when the real users use them. At the same time, the systems that users are enthusiastic about might receive poor scores in automated tests. This disparity makes it clear that it is essential to find new means of measuring the quality of the responses.

How We Traditionally Measured AI-Powered QA Systems

Traditional question-answering systems were evaluated using metrics borrowed from other areas like machine translation. Below are the few of the popular metrics which were used.

BLEU and ROUGE Scores

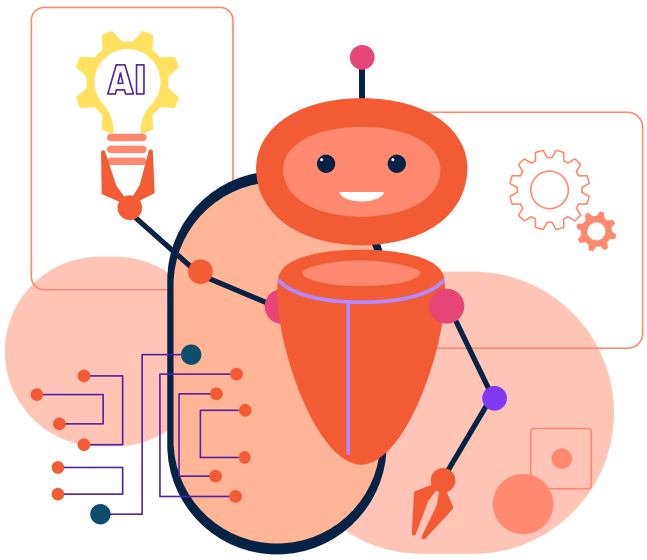
Compare how many words and phrases match between the system's answer and the expected "correct" answer. The more words that match, the higher the score.

Exact Match Accuracy

Verifies if the system's answer is the same as the expected one. If the correct answer is "Paris" and the system says "Paris," it will be 100%. If it says "Paris, France," it will be 60%.

Bert - F1 Scores

Attempt to make a trade-off between precision (the number of correct words in the answer) and recall (the number of correct words that were included).



These metrics worked well for older QA systems that simply found and copied text from documents. If someone asked, "Who was the first person on the moon?" and the correct answer was "Neil Armstrong," it was easy to check if the system got it right. But GEN AI systems work differently. They don't just copy text; they read multiple documents, understand the information, and write new answers in their own words. This creates problems for traditional metrics.

Why GEN AI Systems Are Different

RAG systems have two major components: one is a retrieval system that locates relevant documents, and the other is a generation system that answers questions based on those documents. This two-step procedure gives rise to problems:

- **Synthesis Over Extraction**

Instead of merely extracting text, GEN AI systems merge information from various sources.They could go through five different articles on a topic and compose a single, consistent answer that covers the most important points.

- **Natural Language Generation**

GEN AI systems generate responses in normal, conversational language.They are not like machines that just feed on facts; they function more like human beings while generating answers.

- **Dynamic Knowledge GEN AI Systems**

The current information can be obtained since GEN AI systems can retrieve from present databases & websites, which is not possible for traditional systems that can only be aware of what they have been trained on.

For instance, if you ask a question like "What are the health benefits of green tea?", a traditional system might only give you a sentence like "Green tea contains antioxidants," copied from somewhere. On the other hand, an GEN AI system might give sentences like this: "Green tea has a number of health benefits, and the most important among them is the high antioxidant content. The research indicates that it can help in inflammation reduction, support heart health, and even decrease the chances of getting certain types of cancer.

Traditional metrics often give the GEN AI answer a poor score because it doesn't match the simple reference answer, even though it's clearly more helpful and informative.

The Hallucination Problem Explained

Here is an example to illustrate the hallucination problem

Imagine your GEN AI system is asked, "**What's the population of New York**?" The system retrieves documents about New York, but due to a processing error, it couldn't read the retrieved information properly. Instead, it "hallucinates" (makes up) the answer "8.5 million people" based on its training data.

Now, if the actual population of New York is around 8.5 million, traditional metrics would give this answer a high score because it matches the expected answer. But this is dangerous because:

- The system didn't actually use the retrieved information
- It got lucky with a guess
- Next time, it might hallucinate a completely wrong answer
- We can't trust the system to work reliably

This is "**hallucination masked as accuracy**". The system is working correctly according to our metrics, but it's actually broken. Traditional metrics can't tell the difference between a system that correctly processed retrieved information and one that just got lucky with a guess.

Limitations of Traditional Metrics in measuring GEN AI Responses

Challenge 1:

Penalizing Good Paraphrasing

Traditional metrics cannot differentiate when the words in the text are changed, though the meaning and the semantic similarity are the same. For example, when the retrieved documents stated that the "Company revenue grew 15% in Q3" and the GEN AI system wrote "the company saw 15% growth in third-quarter revenue," traditional metrics would give a low score because the words are not identical.

Challenge 2:

Missing Citation Requirements

Nowadays, GEN AI systems can tell you where their data is sourced from. A good answer may be, "The 2023 annual report...", but traditional metrics ignore whether sources are cited properly.

Challenge 3:

Ignoring Reasoning Quality

When questions need to merge information from multiple documents, traditional metrics can't tell if the system did the reasoning correctly. The final answer might be right by accident, even if the logic was wrong.

Challenge 4:

Context Misunderstanding

A system might provide correct information, but not quite answer the specific questions. The traditional metrics will still give a high score to such a system if the facts are correct, thereby overlooking that it did not comprehend what the user was requesting.

Better Ways to Evaluate GEN AI Systems

In reference to these problems, researchers and organizations are developing new evaluation methods:

Component-Based Evaluation

Instead of just validating the final response, the parts of the GEN AI system can be evaluated separately - Did the system pull in relevant documents? Did it provide a good answer based on those documents?

Faithfulness Checking

Leverage AI systems or even validate manually to verify that the answer generated is supported by the retrieved documents. This helps in preventing hallucinations.

Multi-Dimensional Quality Assessment

Evaluate answers across multiple quality dimensions rather than a single response.

- Accuracy: Is the information factually correct?
- Relevance: Does it answer the question asked?
- Coherence: Does the answer flow logically and is well structured?
- Completeness: Does it cover all the important aspects of the question?
- Appropriateness: Is the tone and level of detail suitable for the context?

Human-in-the-Loop

Take advantage of both automated testing and human reviewers who can evaluate the parameters like helpfulness, truthfulness, bias, & hallucinations that machines alone cannot do efficiently.

Real-World Performance Metrics

Track the ways in which users interact with the system in real life. Will the answers be helpful to them? Do they continue to ask clarifying questions? Do they complete their tasks successfully?

Practical Steps for Better GEN AI Evaluation

Set Clear Goals
Identify and define clearly the success criteria for your specific use case. For example, A customer service bot has different requirements than a research assistant.

01

Multiple Metrics
Don't rely on just a single type of metric. Combine traditional metrics like Bleu, Rouge, Bert scores, with new RAG-specific metrics, and with human observations. Include both Deterministic & Non-Deterministic metrics.

02

Real Users Test
The feedback of real users is always the best way to validate our solution. Find ways to collect and analyze how people interact with your system

03

Monitor Continuously
Keep track of performance over time, especially when changes and updates happen to your knowledge base.

04

Concentrate on the Problem Areas
Pay special attention to controversial topics like any recent events, complex queries about problems where they are most likely to occur.

05

Importance of Trust and Reliability

Trust is a major factor that traditional metrics hardly consider. GEN AI systems are often used for important decisions – medical advice, legal, financial advice, etc. The users must ensure that the system is not only correct but also extracted from reliable sources.

Traditional metrics are incapable of measuring trust. They can't figure out if a system is always trustworthy or if it sometimes tricks with a confident tone and shows incorrect answers. They cannot validate if the system is providing honest and trustworthy responses.

Building & testing reliable GEN AI systems require evaluation strategies that go beyond accuracy scores for validating reliability, consistency, transparency, and unbiased answers.

Looking Ahead: The Future of AI Evaluation

As AI systems become more sophisticated and widely used, evaluation methods continue to evolve. We're likely to see

Domain-Specific Benchmarks

Evaluation frameworks tailored to specific industries and use cases, rather than using generic metrics.

Automated Faithfulness Detection

Improved AI systems that can reliably detect when the generated responses are not supported by any reliable sources.

Continuous Learning from Feedback

Systems get better at evaluation by learning from user interactions and feedback over time.

Multi-Modal Evaluation

Going forward, GEN AI systems will be capable of operating with images, videos, and other media types; evaluation will have to go beyond text-only metrics.

Industry- Specific Evaluation Challenges

Health Care Systems : Safety and liability concerns

Healthcare systems face unique evaluation challenges because incorrect answers can be fatal for the patients or the reputation of hospitals/doctors. Traditional accuracy metrics cannot detect if a system properly indicates uncertainty in the medical advice problem or if it informs users that they have to consult a doctor. The evaluation also must make sure that the system is not going to give any incorrect references, contradictions, or wrong drug recommendations. Healthcare GEN AI evaluation requires specialized medical expertise and must comply with regulations like HIPAA and FDA guidelines.

Legal Systems : Precision vs. comprehensiveness

Legal systems must find a middle ground between being thorough and being precise since a case that is missing an important legal precedent could be a disaster. Traditional metrics can't tell if the system gets legal nuances right. The evaluation must assess if the system appropriately qualifies its answers with and does not hallucinate when providing answers. Citation accuracy is extremely important since the lawyers must check the sources; and if the citations are incomplete or wrong, the entire legal arguments get weakened. The problem lies in gauging the depth of legal reasoning and not just the correct factual one. Traditional metrics like Bleu, Rouge and Meteor scores will not be able to handle these responses.

Financial Systems: Regulatory compliance requirements

The Financial systems are obligated to abide by the regulations that are very rigorous concerning investment advice risk disclosures and warnings. Traditional evaluation methods can't determine if the system correctly includes the required disclaimers about the risks of investment or if it has avoided giving personalized financial advice or stock recommendation. The evaluation must verify that the system is not making any promises about guaranteed returns or giving recommendations that might be considered as fraudulent. The Securities & Exchange commission has set conditions for AI systems which provide financial information to adhere to financial guidelines set by the governing body.

Customer Service Applications : User satisfaction vs. accuracy

Customer service systems faces the challenge that technically accurate answers might not satisfy frustrated customers who need empathy and relevant answers from the AI System. Traditional metrics give a primary focus on the correctness of the information, but they do not consider the response tone if it is suitable for a situation with upset customers. The response evaluation must assess if the system can handle edge cases, company policy exceptions, and when to transfer to human agents. Customer satisfaction scores often matter more than technical accuracy, as a perfectly accurate but unhelpful response can damage customer relationships. The system's ability to understand context, emotion, and urgency becomes as important as providing correct information

Conclusion
Measuring What Matters

The disparity between the GEN AI system's results and traditional QA metrics indicates a fundamental gap. As AI systems become more sophisticated, measuring and quantifying the quality metrics must change as well. Traditional metrics served us well when the applications and systems were simple, but they are no longer suitable for complex tasks and multi-step reasoning, which the GEN AI systems perform.

The result is not to discard all the existing metrics, but to acknowledge their limitations and complement additional metrics such as accuracy, coherence, relevance, context similarity, and a human- in- the- loop approach rather than just validating the response with Bleu or Rouge scores.

The future of AI evaluation isn't about finding perfect metrics, but building comprehensive evaluation strategies that can meet our needs to measure quality. Systems that are helpful, accurate, trustworthy, and reliable in the real world.

Author Bio

Prashanth TV is the Practice Director at Happiest Minds Technologies, currently leading the Quality Assurance (QA) practice within the Generative AI Business Unit. With over 19 years of experience in Software Testing and Quality Engineering, he brings deep expertise in test automation, solution development, and strategic QA consulting. Prashanth has played a pivotal role in building robust testing frameworks, accelerators, and methodologies tailored for emerging technologies, including Generative AI.

happiest minds
AI FIRST. AGILE ALWAYS.

www.happiestminds.com

About Happiest Minds

Happiest Minds Technologies Limited (BSE, NSE: HAPPSTMNDS) is an AI First, customer-centric digital engineering company committed to delivering 'Happiest People . Happiest Customers'. With an integrated approach that spans from chip to cloud, Happiest Minds delivers secure and scalable solutions across product engineering, cybersecurity, analytics , and automation platforms. Happiest Minds brings purpose and precision to every engagement, helping enterprises solve complex business challenges and fast-track their digital evolution across industry sectors such as Banking, Financial Services & Insurance (BFSI), EdTech, Healthcare & Life Sciences, Hi-Tech and Media & Entertainment, Industrial, Manufacturing, Energy & Utilities, and Retail, CPG & Logistics.

Happiest Minds has been honored by both the Golden Peacock Awards and the Institute of Company Secretaries of India (ICSI) for its exemplary Corporate Governance practices. Guided by its mission of 'Happiest People . Happiest Customers' and consistently recognized as a great place to work, Happiest Minds is headquartered in Bengaluru, India, with a global presence across the Americas, UK, Europe, Australia, the Middle East, Africa, and Asia.

For more information contact us at **business@happiestminds.com**