happiest minds
The Mindful IT Company
Born **Digital** . Born **Agile**

Quality Assurance in the Era of LLM's

Methodologies for Evaluating and Validating Generative AI Systems

www.happiestminds.com

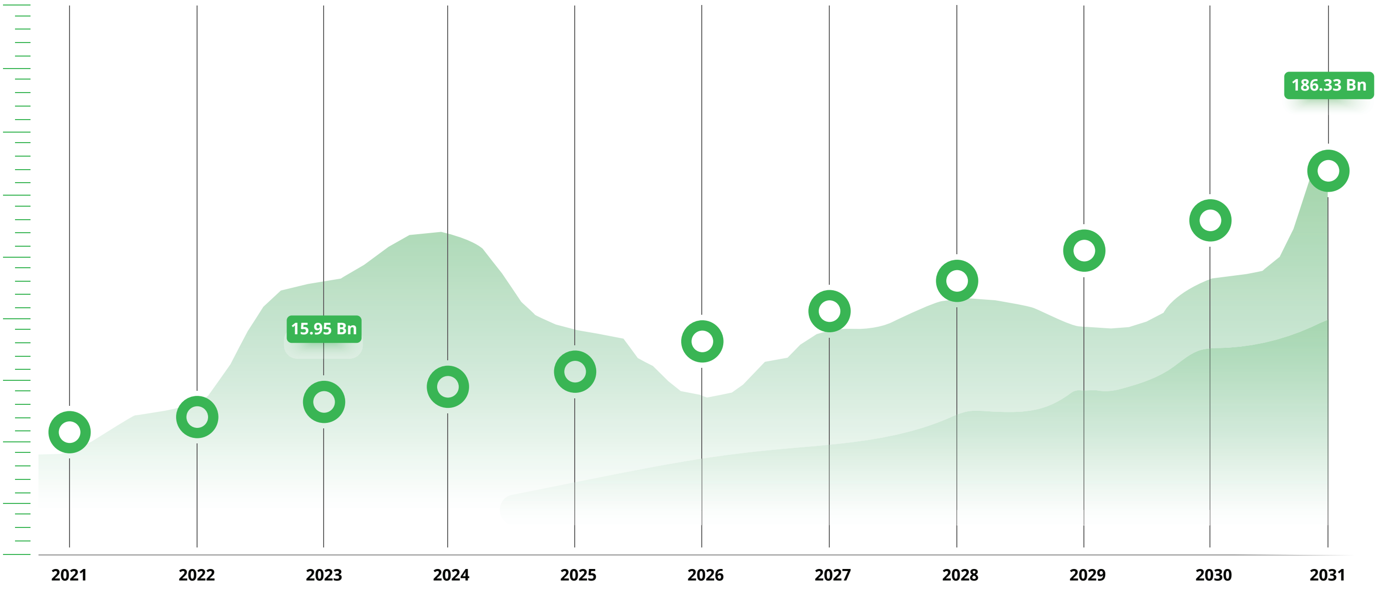
Table of Contents

Introduction

As generative AI (Gen AI) systems are being used more widely in software development, traditional testing strategies will have to transform. Gartner's recent survey results show that AI engineering is estimated to introduce new best practices for software engineering companies with 80% of them having AI-based testing strategies in place by 2025. This whitepaper unfolds the gaps in testing mindsets and presents strategies and tools for Gen AI application's successful trials. Proven by the generative AI market being valued at $186.33 billion by 2031, with a compound annual growth rate of 34.3%, the need to come up with reliable testing approaches is at high stakes.

Global Generative AI Market

Size, 2021-2023 (USD Billion)



Source: https://www.kingsresearch.com/generative-ai-market-478

Difference between traditional testing approach and Gen AI testing

Testing Generative AI (Gen AI) applications requires a new outlook, as it demands a fundamentally different approach compared to traditional software testing. We require a significant shift in mindset when testing Gen AI-based solutions or platforms. Here are some key differences in the testing mindset:

Traditional Software Testing Mindset	Context
Predictive outcomes	Traditional software has a set of anticipated outputs for a given set of inputs, which allows testing to concentrate on confirming compliance with previously set standards.
Scenario-based testing	The scenario-based testing is designed to ensure the software performs as required for various conditions and user interactions, including edge cases.
Defect identification	One of the primary responsibilities of quality assurance (QA) professionals is to identify defects in the product. QA teams are typically composed of skilled individuals who are trained to meticulously inspect and test products at various stages of development. The QA team uses a combination of manual inspection techniques and automated tools to uncover defects.
Comprehensive testing	Although 100% coverage may not be achievable. The aim is to test a wide range of scenarios, maximizing coverage to identify and resolve issues before the release.

GEN AI Testing Mindset	Context
Dynamic response generation	Like all AI systems, Generative AI system's output is based on the intricate patterns extracted from the training data. This calls for the need for innovative testing strategies. Unlike traditional software, GEN AI systems behave in a non-deterministic manner which generates diverse and different responses for the same inputs.
Intent-based testing	The testing approach will be focused on assessing whether the system's responses achieve the desired outcome, rather than adhering to predetermined output expectations.
Bias & fairness evaluation	It is very important to evaluate systems to ensure that there are no traces of Bias or fairness issues as this will stem from the training data which might contain societal and cultural biases.
Comprehensive testing	A broad spectrum of scenarios should be provided for testing, although it is not always feasible to achieve 100% test coverage. The purpose is to perform as many tests as possible to detect and correct potential issues before the system is released.
Human judgement/expert review	Human evaluation by subject matter experts plays a crucial role in validating the quality and accuracy of Gen AI responses.

Key challenges of GenAI testing

Evaluating large language models (LLMs) poses a challenge due to their black-box characteristics. With traditional software, it is easy to predict the outcome and thereby errors can be traced to specific segments of code. With LLMs, the scope of inputs and generated outputs is unimaginably large. This makes error identification and correction nearly impossible. The evaluation consists of verifying how well the LLM comprehends the intended message of the text and produces it in human language, as well as checking the context appropriateness of the responses. The primary goal of testing GenAI-based solutions is to identify and address issues related to Intents,bias, language misinterpretation, and system integration gaps.

The below snapshot captures the key challenges wrt testing LLM Output's

Retrieval Accuracy

Relevance Ranking: Identifying the most relevant documents from a large knowledge base for a given query, especially when dealing with ambiguous or complex questions

Semantic Understanding: Ensuring the retrieval system accurately captures the semantic meaning of both the query and the documents to retrieve truly relevant information

Data Quality Issue: Dealing with inconsistencies, outdated information or noisy data within the knowledge base

LLM Output Quality

Hallucination: Evaluating whether the generated response contains factual errors or fabricated information even if it appears coherent.

Context awareness: Assessing if the LLM properly integrates the retrieved context into its response, avoiding irrelevant or misleading information.

Answer completeness: Ensuring the generated answer addresses all aspects of the query and doesn't miss crucial details.

Bias: Ensure that the generated response is free from bias

System Design Challenges

Scalability: Handling large-scale knowledge bases and high query volumes while maintaining efficient retrieval performance.

Prompt engineering: Crafting effective prompts to guide the LLM to generate accurate and relevant responses based on retrieved information.

Data Management: Managing data from multiple sources and ensuring consistent formatting within the knowledge base

Non-Deterministic Behavior

By design, Gen AI systems produce unique and diverse outputs based on their training data and algorithms. Their non-deterministic nature means that identical inputs can yield different, yet valid, outputs.

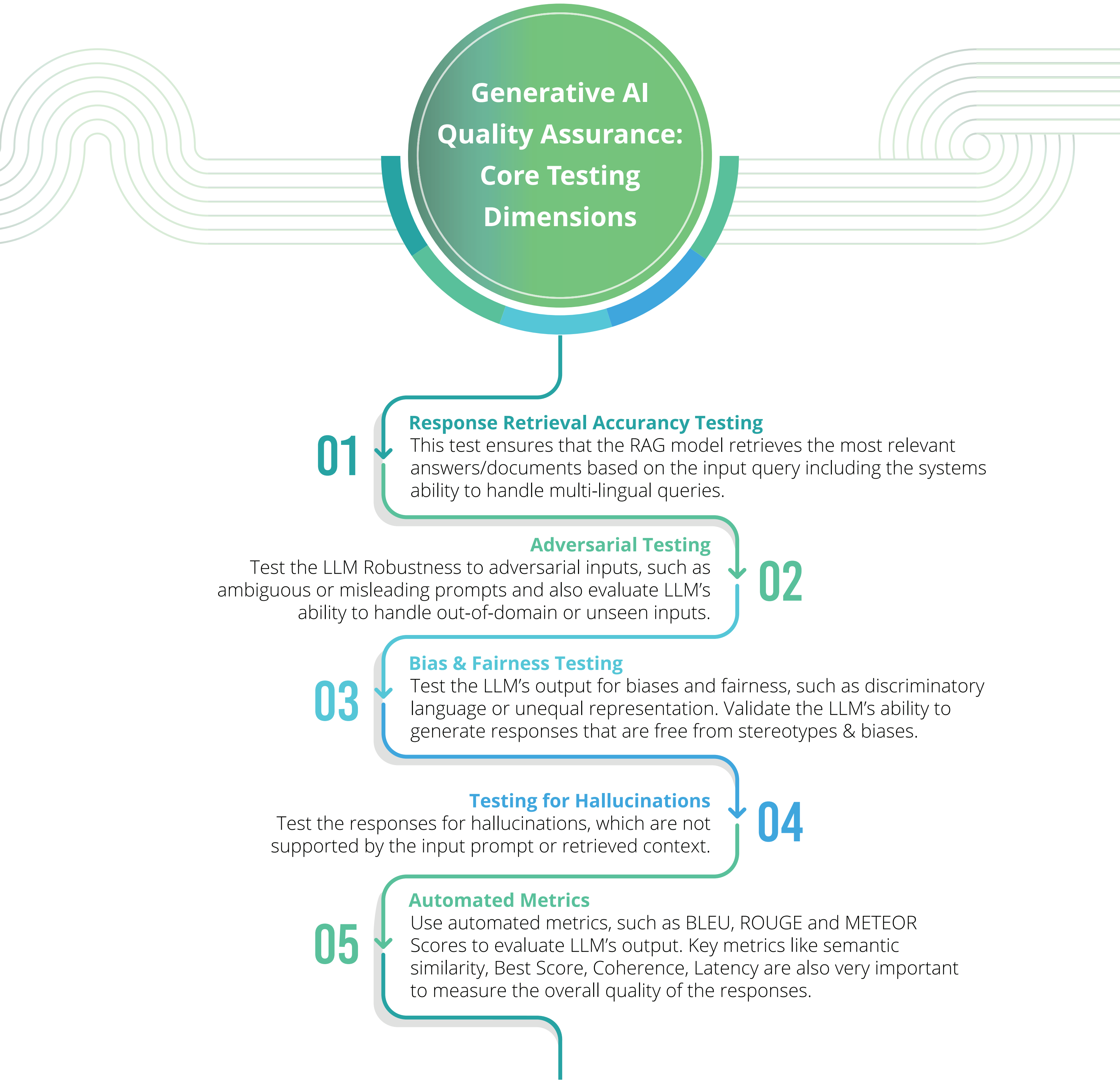
Prompt: Benefits of natural language processing (NLP)

Response 1: NLP enables computers to understand and interpret human language, facilitating applications like sentiment analysis, text summarization, and chatbots, which can improve customer experience, enhance decision- making, and boost operational efficiency

Response 2: By bridging the gap between human communication and machine understanding, NLP unlocks a wide range of possibilities, including

Testing techniques and strategies

The below snapshot indicates the overall paradigm shift in the thought process while testing a Gen AI solution.



Generative AI Quality Assurance: Core Testing Dimensions

01 **Response Retrieval Accurancy Testing**
This test ensures that the RAG model retrieves the most relevant answers/documents based on the input query including the systems ability to handle multi-lingual queries.

02 **Adversarial Testing**
Test the LLM Robustness to adversarial inputs, such as ambiguous or misleading prompts and also evaluate LLM's ability to handle out-of-domain or unseen inputs.

03 **Bias & Fairness Testing**
Test the LLM's output for biases and fairness, such as discriminatory language or unequal representation. Validate the LLM's ability to generate responses that are free from stereotypes & biases.

04 **Testing for Hallucinations**
Test the responses for hallucinations, which are not supported by the input prompt or retrieved context.

05 **Automated Metrics**
Use automated metrics, such as BLEU, ROUGE and METEOR Scores to evaluate LLM's output. Key metrics like semantic similarity, Best Score, Coherence, Latency are also very important to measure the overall quality of the responses.

Instead of testing for specific outputs, the focus shifts to evaluating whether the Gen AI system's outputs align with the intended purpose or user's intent. The goal is to ensure that the system produces relevant, coherent, and appropriate responses.

Illustrative Incidents:
The Price of Imperfect AI - Business Lessons from Quality Lapses

Intent Recognition & Response Accuracy

Example: If a virtual secretary like Alexa is not tested for its intents before deploying to users, it may start giving out irrelevant or even incorrect answers to the user questions.

Consequence: Users might get frustrated with the virtual assistant's performance, which might lead to loss of trust and hurt the company's reputation.

Adversarial Testing

Example: Just picture a situation where a facial recognition system is set up for security at a very busy airport. If it has not been put through severe testing against several tricky situations such as different lighting, angles or disguises-It could pose a problem for the passengers.

Consequence: A person in disguise could slip past the facial recognition system, bypassing security and gaining access to restricted areas

Testing for Inclusive AI: Bias & Fairness

Example: A hiring algorithm which is not tested for Bias might start discriminating Gender's which in turn might bring in Bias in hiring the candidates.

Consequence: The company might face a lawsuit for discriminatory hiring practices, which might tarnish its reputation and lead to financial losses.

AI Hallucinations: The Risk of Fabricated Facts

Example: A medical diagnosis system which has not been tested for hallucinations, might provide false information or incorrect diagnosis to its patients, which can backfire against the hospitals or the doctors.

Consequence: Patients might receive the wrong treatments, which could worsen their health and even lead to fatalities. The company could face a lawsuit for medical malpractice, damaging its reputation and causing financial losses

Human evaluation or Human-in-the-Loop (HITL)

The lack of comprehensive human testing has led to notable failures in generative AI solutions, emphasizing the critical role of human evaluation in ensuring AI systems meet their intended goals. Human testing provides a critical layer of evaluation, enabling developers to identify and address potential issues before AI systems are deployed.

Effective human validation is crucial for ensuring that AI systems are reliable, trustworthy, and aligned with our values. There are several key incidents that could have been averted with human validation, which underscores the risks of depending solely on AI algorithms.

Below are a couple of key incidents that could have been avoided if there had been human validation instead of solely relying on AI algorithms.

ChatGPT legal case mishap

Failure description: A lawyer used ChatGPT in court and cited fake cases that didn't even exist. The use of ChatGPT was discovered by the opponent's lawyers wherein they could not find any relevant documents for the cases which was quoted. This incident brought to light the dangers of completely relying on AI for essential tasks without any human checks.

Air Canada chatbot incident

Failure description: Air Canada's virtual assistant promised a discount that wasn't available to a passenger. The passenger was assured that he could book a full-fare flight for his grandmother's funeral and then apply for a bereavement fare discount. This led to a passenger purchasing a ticket under false recommendations, and in turn resulted in a legal dispute and financial penalties for the airline.

Key Quality Metrics and Validation Criteria for measuring GEN AI Outputs

When it comes to validating GEN AI Outputs for a RAG System (Retrieval-Augmented Generation), some of the key Validation Parameters and Quality Metrics are essential for ensuring that the outputs are accurate, relevant, and trustworthy. Some of the essential validation parameters to keep an eye include Relevance, accuracy, fluency, coherence, and completeness of the generated responses.

To measure how well the responses are generated, Below are some of the popular metrics which are used to assess the quality of the generated responses.

Response Reliability Metrics

Response Relevance

This metric quantifies how well the generated response addresses the user's query or prompt intent. High relevance indicates the system correctly interpreted the user's query.

Response Consistency

The metric is a method of determining how much of the explanation of the RAG system's response can be unquestionably identified as having come from the context documents. High consistency scores signify that a larger portion of the response is assured by data from source materials.

Hallucination Detection

Identifies instances where the model generates inaccurate or fabricated information not supported by the retrieved context. This metric is critical for maintaining system trustworthiness in knowledge-intensive applications.

Linguistic Quality Assessment

Coherence

Evaluates whether the generated content preserve's a logical flow, appropriate transitions between statements and also maintains the overall structure that can be read easily.

Fluency

Validates the grammatical correctness of the generated text, natural language usage, and has an appropriate style across different types of responses and length.

Completeness

Assesses whether the response covers all aspects of the query without omitting any critical information required by the user.

Ethical Safeguard Metrics

Toxicity Detection

Identifies potentially harmful content like hate speech, offensive language and inappropriate material that could hurt people or go against ethical standards This essential safety measure uses special classifiers to spot and filter out problematic content before users see it. This protects both companies and individuals from possible harm.

BIAS Detection

Bias testing checks if AI systems show favoritism or discrimination toward different groups or opinions. It looks at the outputs based on factors like gender, ethnicity, age, and income level to make sure the system works fairly for a wide range of people and situations.

Industry-Standard NLP Evaluation Metrics

BLEU (Bilingual Evaluation Understudy)

A precision-focused metric that quantifies similarity between machine-generated text and human reference text by calculating n-gram overlap. Originally developed for translation tasks but adaptable to general text generation quality assessment.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE Score are a set of metrics used to evaluate the quality of summaries or generated text by comparing them to reference texts. It primarily measures overlap in n-grams, word sequences, and word pairs between the generated and reference outputs.

METEOR (Metric for Evaluation of Translation with Explicit Ordering)

An advanced evaluation metric that incorporates semantic similarities including synonyms, stemming, and paraphrasing to align more closely with human judgment of quality. Provides a balanced assessment of precision and recall.

BERTScore (Bidirectional Encoder Representations from Transformers)

Leverages contextual embeddings from pre-trained language models to compute similarity scores between generated and reference texts at the token level. This semantic similarity approach often correlates better with human evaluations than lexical overlap metrics.

Implementation Considerations

For comprehensive validation of GenAI systems, Project teams should:

Use a variety of evaluation methods that cover different parameters.
Establish clear threshold values for each metric based on use case requirements.

Define clear target values for each evaluation method, aligned with specific use case goals and performance expectations.

Supplement automated metrics with periodic human evaluation to ensure metric alignment with business objectives.

This well-rounded approach helps ensure that GenAI systems produce outputs that are steady, reliable, and meet both technical and business needs while staying ethical

Regression testing strategy for Gen AI solutions

Regression testing plays a vital role in making sure that GEN AI chatbots deliver reliable and consistent responses, especially in settings where updates, fixes, and fine-tuning happen frequently. The main aim here is to ensure that any new changes don't accidentally harm the chatbot's performance or bring about unexpected problems

Baseline creation: Start by building a thorough test suite with set inputs and expected outputs. This will act as a benchmark to compare the chatbot's responses after any changes.

Automated testing pipelines: Setup automated regression testing suites which can effectively validate text based comparisons and semantic validations.

Semantic validation metrics: Leverage advanced metrics like BLEU, ROUGE, METEOR, and BERT scores to evaluate the response quality. We can also Incorporate metrics for coherence, relevance and fluency which is used to evaluate how well the language mode can produce output that flows smoothly, reads naturally and resembles human-like responses.

Bias and fairness testing: We need to continuously test the responses for biases & stereotypes; and also make sure that the fixes or updates do not introduce any discriminatory or toxic content.

Multi-turn conversations: Incorporate test scenarios for multi-turn conversations to verify if the context of the responses is preserved in subsequent queries.

Hallucination detection: Perform regular checks on the responses to ensure that the GEN AI chatbot generates accurate content and it does not fabricate responses.

Performance testing: Monitor the response times of the LLM and ensure that code change or configurations do not affect the latency of the chatbot during high loads.

Test automation of LLM outputs

While Large Language Models (LLMs) transform different industries, they also bring special challenges to the test automation strategy because of their unpredictable behavior and contextual processing. LLMs produce variable outputs unlike traditional software; so standard validation techniques prove insufficient. Our approach to maintaining the accuracy, coherence, factual correctness, and ethical standards of these models require innovative thinking.

Automated testing must expand its scope beyond keyword verification to incorporate semantic checks as well as contextual accuracy evaluation including detecting biases and hallucinations. To foster user confidence in GenAI solutions models, we need to use reference-based evaluation methods (such as BLEU, ROUGE, METEOR) alongside AI-assisted validation and human-reviewed testing procedures.

Some of the key use cases for test automation include:

Automated test case generation is a complex method to produce test prompts by using a script that is dynamically created based on different parameters and requirements. The method encompasses the abstraction of input requirements, for instance, the style (either formal, casual or technical), language preferences (English, Spanish, Hindi), and intended output format (structured responses, bullet points, narrative) to generate appropriate test prompts.

The automated prompt generator is an advanced framework for processing documents and generating question-answer pairs. It's designed to automatically create detailed Q&A pairs from a variety of formats, such as Word documents, Excel spreadsheets, PDFs, and even images. Project teams can really benefit from this, as they have the potential to boost productivity by over 70%.

Automated quality evaluation metrics: Automated evaluation metrics for LLM outputs include both quantitative and qualitative measures that enable the evaluation of the responses synthesized by large language models. These metrics are crucial for semantic accuracy, factual correctness, response relevance, and coherence of the generated responses. Some of the key quality metrics include ROUGE (which considers the overlap of texts), BLEU (which measures precision), and BERT scores (for semantic similarity). The main objective of these evaluation metrics is to ensure the high quality of LLM outputs, consistency, and reliability, as well as compliance with user expectations and business needs.

Automation testing of text-speech validations: This solution is all about checking how well AI chatbots can convert text into speech. It plays a crucial role in ensuring that the text generation and speech synthesis are up to par by automating the testing process for both the accuracy of the content and the quality of the audio. This is especially important in our current AI-driven world, where voice interactions are on the rise. It provides a thorough way to validate not just the meaning of AI responses but also how well they sound when spoken.

Future trends and considerations LLM testing

01 Trend: **Real-time adaptive testing**

Dynamic model updates: Since fine-tuning and continual learning are getting more and more popular, testing processes should be able to stay current with the real-time validation of model updates. Automated pipelines are recommended to be designed in a way that they can easily detect and validate the changes with no necessity of re-deployment.

02 Trend: **Explainability and interpretability testing**

Focus on explainability As the rules about AI accountability are becoming indisputable, LLMs must be able to provide the user with explainable outputs. Testing must ensure that the responses are simple to understand and that the source documents are correctly mentioned in the Retrieval-Augmented Generation (RAG) implementations.

Evaluating causal consistency: Testing frameworks will need to evolve to guarantee a logical connection between queries, retrieved documents, and the responses generated.

03 Trend: **Multi Model LLM Testing**

Cross-modal interactions: Due to the abilities provided by the GPT models of OpenAI including their ability to work on text image, and audio input/output, testing frameworks will be expected to check the consistency of the output on different models.

Scenario-based testing: The daily conditions will involve more and more such cases where inputs come in combination (such as a voice query combined with textual context), and hence the use of simulation environments will increase.

04 Trend: **Responsible AI testing**

Ethical validation: The importance of bias and ethical compliance will mature during course of time. Testing will become extremely crucial and has to become fully automated across all languages, culture and demographics.

05 Consideration: **Regulatory and compliance testing**

Industry-specific standards: With industries like healthcare and finance increasingly adopting large language models (LLMs), it's crucial to conduct compliance testing that aligns with regulations such as HIPAA and GDPR.

AI audits: Testing frameworks will play a key role in preparing for audits, making sure that traceability, data lineage, and ethical considerations are all properly validated.

Conclusion

When it comes to the testing of GenAI solutions, we have to change the way we think about the traditional software testing methods. The generative AI systems are complex, especially in settings of Retrieval-Augmented Generation (RAG). Therefore, they require a strong strategy, which includes high accuracy, consistency, and adaptability. To deal with challenges such as hallucinations, bias, and ever-changing responses, we have to adopt innovative automation and regression strategies. As LLMs become more integrated into critical applications, having a strategic, adaptable, and automation-first approach to testing will be essential for building trust and delivering high-quality AI solutions.

About the Author

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.

About Happiest Minds Technologies

Happiest Minds Technologies Limited (NSE: HAPPSTMNDS), a Mindful IT Company, enables digital transformation for enterprises and technology providers by delivering seamless customer experiences, business efficiency and actionable insights. We do this by leveraging a spectrum of disruptive technologies such as: artificial intelligence, blockchain, cloud, digital process automation, internet of things, robotics/drones, security, virtual/ augmented reality, etc. Positioned as 'Born Digital. Born Agile', our capabilities span Product & Digital Engineering Services (PDES), Generative AI Business Services (GBS) and Infrastructure Management & Security Services (IMSS). We deliver these services across industry groups: Banking, Financial Services & Insurance (BFSI), EdTech, Healthcare & Life Sciences, Hi-Tech and Media & Entertainment, Industrial, Manufacturing, Energy & Utilities, and Retail, CPG & Logistics. The company has been recognized for its excellence in Corporate Governance practices by Golden Peacock and ICSI. A Great Place to Work Certified™ company, Happiest Minds is headquartered in Bengaluru, India with operations in the U.S., UK, Canada, Australia, and the Middle East.

happiest minds
The Mindful IT Company
Born **Digital** . Born **Agile**

www.happiestminds.com

To know more about our offerings. Please write to us at **business@happiestminds.com**