Persistent

Evaluation Framework for Generative AI Applications

An innovative approach to enterprise-grade application development

As the AI landscape continues its evolution within enterprises, Generative AI (GenAI) and Large Language Models (LLMs) have become pivotal in enhancing enterprise applications. Patterns like retrieval augmented generation (RAG), querying SQL databases, and agentic workflows are gaining popularity in the building of next-gen applications. What's needed, however, is a unified approach to evaluate these GenAI-enhanced applications. While multiple leaderboards like LMSYS Arena do a suitable job evaluating individual models, what enterprises require is a way to evaluate GenAI-based applications built using these models through a rigorous evaluation.

In this white paper, we present an enterprise-focused Evaluation Framework designed to assess the performance of GenAI-based applications on unstructured and structured data utilizing both black box and white box testing approaches, with the capability to assess agent trajectory for agentic workflows.

Evaluation Framework: An Approach for Methodical Data-Based Assessments

A major challenge in the evaluation of GenAI applications is determining the right metrics by which to evaluate generated data. Unlike traditional machine learning (ML) where metrics like precision, recall, and F1-score provides a good view of performance, GenAI needs more sophisticated metrics and evaluation. For example, it may easily be the case that a generated answer has no matching words to the ground truth text but may still be 100% the correct answer. Therefore, new metrics that test the essence of generated data are required.

The Evaluation Framework engineered by Persistent is designed for enterprises to methodically assess GenAI application performance. It automates the creation of test data, evaluates answer quality using intuitive application-oriented metrics, and utilizes production data to drive ongoing application enhancement. GenAI applications can utilize diverse types of data as a knowledge base, including but not limited to text, structured data, graph data, and image data. Each modality has specific test data requirements and evaluation metrics. However, this framework offers a unified experience across these various data types.

Test Data

Evaluation of any application requires test data, and this holds true for GenAI applications as well. The creation of pairs of "a question and its ground truth answer (one based on real truthful data)" is a critical component of the Evaluation Framework. Traditional testing involves generating test data for various conditions to encompass numerous test scenarios. For GenAI applications, the test data includes questions, correct answers, and optionally, precise context from the knowledge base. Thus, akin to the "test scenarios" in traditional testing methodology, the Evaluation Framework provides different question types to assess the application from multiple perspectives.

Unstructured or Text Data

For unstructured or test data, the test data point comprises a question and the answer. Examples of question types for unstructured data in the Evaluation Framework are highlighted in Figure 1.

Question Type	Description	When to Use
Contextual	This represents the most straightforward type of question, where the context matching the given question is located within a single document to produce the answer.	Employ these questions to assess the fundamental functionality and interconnections of all components within an application.
Multi-context	In this scenario, chunks relevant to the question are identified in several documents, and the final response is formulated by amalgamating this information.	Employ these questions to assess the efficiency of an application in consolidating information from multiple documents.
Summarization	The question necessitates an application to condense a larger portion of the text.	At times, the chunking process can result in information loss, which can negatively impact the generation of summaries for specific topics covered in a PDF document. Utilize these questions to test and improve an application's summarization capabilities.
Tabular	The questions are formulated based on tabular data extracted from provided PDF documents.	Should the tabular data in the PDF contain significant information, employ these questions to examine and enhance an application's capability to process tables within PDF documents.
Conversational	A series of questions is created by first generating a main seed question, followed by generating follow-up questions based on the answers provided for the seed question.	Utilize these questions to evaluate an application's ability to engage in a dialogue with the user, leveraging its conversation history.

Figure 1: Evaluation Framework question types for unstructured data. (Source: Persistent)

The framework facilitates the easy integration of new question types, tailored to the specific requirements of the application.

Structured Data

For structured data, each test data point comprises a question, its corresponding SQL query crafted to fetch the requisite data, and the answer — specifically, the result set generated by executing that SQL query on the provided sample data. Here, the types of questions are determined based on the complexity of the questions, as seen in Figure 2.

Question Type	Description	When to Use
Simple	Simple questions on a single table with filterby conditions.	Utilize these questions to evaluate the basic functionality and interconnections of all components within an application. Furthermore, they aid in determining if the query is selecting the correct set of columns.
Aggregate	Questions involving use of aggregate functions on a single table.	Utilize these questions to evaluate whether any English terms in the question, which indicate aggregated data, are accurately translated into corresponding aggregate functions in an application.
Multi-table	Questions that require a joining between two tables or use of subquery clause.	Employ these questions to check an application's ability to handle more complex data spread across multiple tables.

Figure 2: Evaluation Framework question types for structured data. (Source: Persistent)

Evaluation of GenAI applications on graph data is like that of structured data. For graph data, question types are derived based on the number of hops required in the query to obtain the answer. Also, in place of SQL, a cypher query is generated as part of test data.

Agent Trajectory Evaluation

More recently, GenAI applications are increasingly deploying agentic workflows, primarily due to the adaptability that they provide. In this setup, an agent is tasked with generating answers to questions, leveraging a specified set of tools effectively. Interestingly, not all these tools necessitate access to data — a calculator or a code interpreter serves as a case in point. The Evaluation Framework introduces a distinctive set of metrics for conducting white box testing on these applications. This is accomplished by assessing the trajectory undertaken by the master agent to formulate an answer, thus offering profound insights into an application's operation and performance.

Test Data Generation

There are multiple ways to prepare training data for evaluation.

Autogenerated Question Set: The Evaluation Framework platform has the capability to generate questions using LLMs, drawing from the seed data provided. Given the non-deterministic nature inherent in LLMs, the quality of generated questions may vary. To ensure the highest quality, the platform incorporates a manual maker-checker process. This feature enables users to review and rectify any autogenerated questions or answers, thereby enhancing the accuracy and reliability of the generated content.

Manually Compiled Question Set: Users have the flexibility to conduct multiple "question generation" iterations, cherry-picking specific questions from each run to compile bespoke question sets. For instance, one could create a "hardening" question set by selecting the most challenging questions across all runs. This method allows for a tailored approach to meet specific needs or objectives.

Imported Question Set: Should the user already possess a pre-prepared list of questions and answers, they have the convenience of importing them directly into the platform. This imported data can then be utilized effectively in the evaluation process.

Evaluation

Once the golden data containing sample questions and ground truth answers is prepared, the next step is to use it to evaluate incrementally built versions of the RAG application. Users can download this data, test it on the application, and collect the generated answers. These application-generated answers can then be uploaded onto the platform for evaluation.

Evaluation Metrics for Unstructured or Text Data

The Evaluation Framework employs lexical, semantic, and Responsible AI evaluation metrics to provide a holistic assessment of the responses.

/ **Lexical Evaluation**

- BLEU: Measures the precision of the response by comparing it to the reference answers.
- ROUGE: Assesses the recall of the response by comparing it to the reference answers.

/ **Semantic Evaluation**

- This table provides a snapshot of a range of metrics that are valuable in quantifying the quality of responses.

Metric	Description	Comment
Answer Correctness	This metric gauges the precision of the generated response based on the reference or ground truth answer.	It aids in determining whether the application's response encompasses all aspects highlighted in the ground truth answer.
Answer Relevance	Assesses how pertinent the response is to the given question.	A ground truth answer is not required for this metric, making it applicable for responses generated by the application in a production environment.
Faithfulness	This evaluates the response's consistency with the provided context, ensuring it doesn't introduce any inaccurate information.	This metric can be computed if the context is provided with the application-generated response. It assists in quantifying any instances of hallucination.

Responsible AI Evaluation: Metrics like controversiality, criminality, harmfulness, insensitivity, and maliciousness are calculated to evaluate application-generated responses, ensuring they adhere to the principles of Responsible AI.

Evaluation Metrics for Structured Data

In this scenario, the test data comprises a question, an SQL query, and a ground truth answer — the result set derived from executing the SQL query. Consequently, metrics usually applied to unstructured data are not suitable here. The Evaluation Framework utilizes the intuitive metrics described in Figure 3 to gauge the accuracy of the SQL query generated by the application in comparison to the ground truth SQL query. To execute the evaluation, users must upload questions and corresponding SQL queries generated by the application.

Metric	Description	Comment
Syntax Correctness	This metric verifies the syntactical correctness of the SQL query generated by the application.	It helps determine the selected model's ability to construct grammatically accurate SQL queries.
Table Name Match	This checks whether the tables referenced in the SQL query produced by the application accurately align with those referenced in the SQL query in the ground truth.	This metric value can be improved by improving the description and metadata of the tables provided in the context to the model.
Column Name Match	Like the above metric, this metric checks alignment and correctness of columns.	Like the above metric, this metric value can be improved by providing accurate descriptions of each column in context.
Result Set Match	This involves executing the SQL query generated by the application against the sample data uploaded during the question-generation phase. The resulting data set is then compared with the one produced by the SQL query in the ground truth. This comparison helps to evaluate the accuracy of the data retrieved by both queries.	This metric ensures the actual data returned for a question aligns with the ground truth answer.
LLM Judgement	Solicited to assess or grade the SQL query generated by the application.	This provides a measure of the quality and accuracy of the SQL query in relation to the initial question. This metric does not need ground truth.

Figure 3: Evaluation Framework metrics for structured data. (Source: Persistent)

Evaluation Metrics for Graph Data

For applications involving graph data, evaluation metrics typically focus on verifying the nodes referenced in the Cypher query and assessing the accuracy of the resulting answers.

Tracking Production Data

The Evaluation Framework comes with a python SDK that can be used by an application to send the following data to the Evaluation Framework server:

User's question	Retrieved context	Latency
Generated answer	Models used in production	User's feedback

The Evaluation Framework aims to automatically categorize questions into distinct types, which allows for an analysis of areas where the application might be underperforming. These questions can then be incorporated into specialized sets like the "Hardening Set" or "Sanity Set." This process contributes to enhancing the quality of the test data.

Development to Production

During the Development Phase, the Evaluation Framework can be utilized to:

Experiment with various strategies to identify the one that best suits different types of questions expected on the application.	Experiment with various models to find the most optimal one.	Track every incremental improvement made to the application to understand the development journey.

During the Production Phase, the Evaluation Framework serves multiple purposes:

It helps conduct a sanity check prior to deploying upgrades in production, ensuring that the system is functioning as expected	It allows incorporation of genuine user queries into the platform to identify necessary improvements.

The Value of the Evaluation Framework

GenAI evaluation is a growing enterprise challenge as the use of GenAI-enabled applications proliferates. With the Evaluation Framework for LLM-based RAG, we have captured key patterns and metrics to address emerging evaluation issues through our robust and comprehensive platform, which has been meticulously designed to meet the needs of enterprise-level GenAI application deployments.

By incorporating sophisticated question set generation and detailed evaluation metrics, the framework ensures the reliability, accuracy, and relevance of GenAI responses. This systematic approach not only enhances the performance of RAG applications but also builds trust and confidence in AI-driven solutions across various industries.

While Phase 1 of the framework is focused on RAG applications, which comprise 90% of current GenAI use cases, our vision extends far beyond this initial scope. The framework is being developed with the foresight to evaluate a broader range of GenAI applications, including those based on structured data and agents. Moreover, our framework is designed to measure more than just performance — it also addresses biases and aligns with Responsible AI principles. This holistic approach positions the Evaluation Framework as a vital tool for fostering ethical AI development, promoting transparency, and ensuring that AI systems deliver value in a fair and responsible manner.

As we continue to expand and refine this framework, it will serve as a cornerstone in our ongoing efforts to build and deploy enterprise-ready GenAI applications that are not only powerful but also aligned with the highest standards and values that Responsible AI demands.

See Beyond, Rise Above

About Persistent

Persistent Systems (BSE & NSE: PERSISTENT) is a global services and solutions company delivering Digital Engineering and Enterprise Modernization to businesses across industries. With over 23,500 employees located in 19 countries, the Company is committed to innovation and client success. Persistent offers a comprehensive suite of services, including AI-enabled software engineering, product development, data and analytics, CX transformation, cloud computing, and intelligent automation. The Company has been recognized as the "Most Promising Company" of the Year by CNBC-TV18 at the 2023 India Business Leader Awards. As a participant of the United Nations Global Compact, Persistent is committed to aligning strategies and operations with universal principles on human rights, labor, environment, and anti-corruption, as well as take actions that advance societal goals.

USA
Persistent Systems, Inc.
2055 Laurelwood Road
Suite 210, Santa Clara
CA 95054
Tel: +1 (408) 216 7010

India
Persistent Systems Limited
Bhageerath, 402
Senapati Bapat Road
Pune 411016
Tel: +91 (20) 6703 0000

Persistent

www.persistent.com