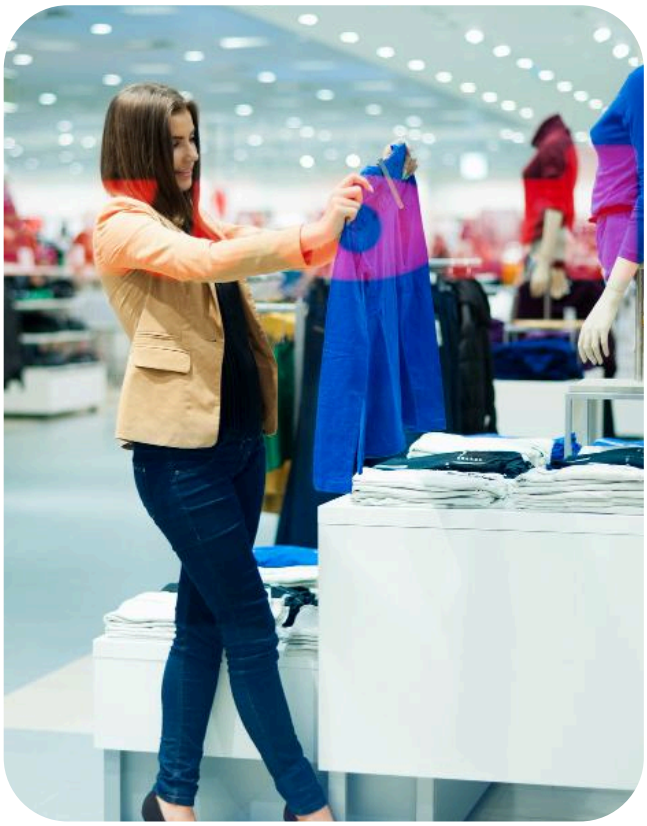
QBurst | A High AI-Q™ Company

Automated Multi-LLM Testing Framework for Retail

Enabling scalable evaluation of conversational AI quality, safety, consistency, and performance for enterprise retail operations.

Overview

- Designed and implemented an automated testing and evaluation framework for a GenAI-powered retail chatbot replacing an existing rule-based conversational system.
- Enabled scalable validation of chatbot accuracy, faithfulness, toxicity handling, guardrails, latency, and response consistency using a multi-LLM evaluation infrastructure.
- Automated functional, edge-case, and safety testing workflows through Playwright, BehaveX, DeepEval, and LLM-as-a-Judge evaluation pipelines.
- Delivered reusable, maintainable, and high-throughput QA automation capabilities with comprehensive reporting and integrated performance analytics.



Client Profile

The client is a global fashion retailer operating multiple consumer brands across Asia-Pacific in international retail and e-commerce markets. The organization was modernizing its customer-facing chatbot experience using Generative AI to support product discovery, customer service, shipping enquiries, returns, exchanges, and conversational shopping interactions at scale.

Validating Enterprise-Scale Generative AI Customer Experiences

The retailer required a robust testing strategy to validate the accuracy, safety, responsiveness, and operational reliability of a new GenAI-powered conversational agent replacing an existing Dialogflow-based chatbot.

Key challenges included:

- Measuring the quality and accuracy of GenAI-generated responses across diverse retail customer interactions.
- Validating intent detection, conversational consistency, and factual faithfulness.
- Detecting hallucinations, abusive outputs, biased responses, and prohibited content generation.
- Measuring latency across agentic LLM workflows and multi-step orchestration pipelines.
- Managing evaluation costs and accessibility associated with hosted Large Language Models.
- Preparing high-quality golden reference datasets for large-scale automated evaluation.
- Automating edge-case testing and guardrail validation across multiple interaction categories.
- Comparing performance across multiple foundation models to determine optimal agent behavior.

Building an Automated Multi-LLM Evaluation Framework for Conversational AI

We designed and implemented a comprehensive test automation and evaluation framework to validate the retailer's GenAI-powered chatbot across functionality, safety, quality, and operational performance dimensions.

The framework leveraged Playwright, BehaveX, DeepEval, and a custom multi-model LLM evaluation infrastructure built around the concept of "LLMs as a Judge."

Automated Testing and Evaluation Pipeline

The testing ecosystem automated chatbot validation across multiple test categories and conversational scenarios.

The solution included:

- Automated execution of chatbot test scenarios using BehaveX and Playwright.
- Business-readable Gherkin-based test definitions for collaboration between technical and non-technical teams.
- DeepEval integration for remote LLM-driven evaluation workflows.
- Automated browser-based interaction testing and orchestration pipelines.
- Allure-based reporting for automated execution visibility and performance tracking.
- Modular Page Object Model (POM)-based automation architecture for maintainability and reusability.

Golden Reference Data Generation

To improve evaluation quality and scalability, we implemented an automated golden dataset generation strategy.

The framework leveraged DeepEval synthesizers to generate structured Q&A datasets from knowledge-base content through:

- Intelligent text extraction and chunking.
- Batch question-and-answer generation using hosted LLMs.
- Automated generation of context-aware benchmark datasets containing user questions, supporting context, and ideal responses.

Test datasets included:

- Normal product and service queries.
- Hallucination detection scenarios.
- Abusive content validation.
- Illegal and prohibited topic boundary testing.

Multi-Metric LLM Evaluation

The evaluation framework measured chatbot performance across multiple conversational quality dimensions.

Key evaluation metrics included:

- Answer correctness.
- Answer relevancy.
- Faithfulness and hallucination detection.
- Bias and toxicity validation.
- Response consistency across repeated executions.
- Latency measurement for each test execution.

The framework also incorporated:

- Meaning-based matching between actual and expected outputs.
- Parallel evaluation of multiple claims for scalable assessment.
- Explainability and reasoning-based scoring mechanisms.
- Rubric-driven evaluation criteria customized for specific testing requirements.

LLM-as-a-Judge Infrastructure

QBurst implemented a scalable multi-model evaluation infrastructure using the "LLM-as-a-Judge" paradigm.

Key capabilities included:

- Multi-model evaluation through the QBurst LLM Gateway supporting more than 13 models including GPT-4o, DeepSeek R1, and Mistral Codestral.
- Dynamic judge-model selection based on evaluation complexity and metric requirements.
- Consensus scoring using multiple evaluator models to reduce single-model bias.
- Prompt-engineered evaluation pipelines for consistent scoring across quality dimensions.
- Cost-optimized model routing balancing evaluation accuracy and operational efficiency.
- Parallelized evaluations enabling high-throughput conversational testing at scale.

Comprehensive QA and Reporting

The framework enabled systematic testing across conversational quality, safety, and operational performance categories.

Additional capabilities included:

- Automated dashboards for execution reporting and performance monitoring.
- Integrated bug tracking and resolution workflows.
- Automated latency and consistency measurement pipelines.
- A/B testing across Gemini 1.5 Flash, Claude 3.7, Llama 3.1, and GPT-4 models.
- Structured guardrail evaluation for toxicity, hallucination, and prohibited content detection.

Technical Highlights

- Multi-LLM evaluation infrastructure using the “LLM-as-a-Judge” paradigm.
- Automated testing framework built using Playwright, BehaveX, and DeepEval.
- Golden reference dataset generation using vector embeddings and hosted LLMs.
- Consensus-based evaluation scoring across multiple judge models.
- Automated faithfulness, toxicity, relevance, and hallucination testing.
- Parallelized large-scale conversational evaluation pipelines.
- Gherkin-based business-readable test scenarios.
- Automated reporting and integrated bug-tracking workflows.
- Latency and consistency validation across repeated conversational executions.
- Comparative A/B testing across multiple foundation models.

Impact

- Enabled scalable and automated validation of GenAI chatbot quality across retail customer interactions.

- Improved conversational testing coverage across functional, safety, and edge-case scenarios.
- Reduced repetitive regression testing effort through automation-driven workflows.
- Delivered reusable and maintainable test automation assets using modular framework design.
- Improved collaboration between business and technical teams through Gherkin-based scenario definitions.
- Enabled comprehensive reporting, performance monitoring, and integrated QA governance for enterprise conversational AI deployments.