Retail QA Automation Framework for LLM-Generated UX: Testing Conversational Commerce Interfaces for Compliance, Clarity, and Consistency

Mohnish Neelapu

Automation Lead
QA Automation
Numeric Technologies INC., India.

Abstract:
The adoption of conversational AI in retail has transformed customer interactions, yet challenges remain in ensuring accurate, compliant, and coherent responses. This study introduces a Retail QA Automation Framework for LLM-Generated UX, designed to systematically evaluate large language model (LLM)-driven conversational commerce interfaces for compliance, clarity, and consistency. The framework employs a modular pipeline including realistic query generation, standardized LLM execution, automated response analysis, and QA scoring, leveraging a hybrid approach of rule-based logic and transformer-based classifiers. A comprehensive dataset combining real, public, and synthetically generated retail queries encompassing multi-turn dialogues and paraphrase variations enables rigorous assessment across product inquiries, promotions, returns, and complaints. Responses are evaluated for policy adherence, semantic clarity, context retention, and coherence, with aggregated QA scores visualized through dashboards and heatmaps. Experimental results across GPT-4, LLaMA-2, Falcon-7B, and a retail fine-tuned GPT model indicate that domain-specific fine-tuning significantly enhances compliance, readability, and dialogue consistency, while lightweight models offer operational efficiency. The proposed framework provides a reproducible and scalable methodology for benchmarking retail conversational AI, supporting the deployment of reliable, user-friendly, and policy-compliant LLM-driven retail experiences.

Keywords: Retail QA Automation, Large Language Models, conversational Commerce, compliance and Clarity Evaluation and multi-turn Dialogue Consistency.

1. INTRODUCTION

Modern retail environments Software systems are still growing in complexity and scope as they combine a wide variety of distributed services and intelligent components to facilitate personalized and seamless customer experiences. Since conversational commerce interfaces in the form of chatbots, a virtual shopping assistant, and an automated customer care agent are used by more organizations, the importance of a strong Software Quality Assurance (SQA) grows. The high velocity of releasing digital retailing platforms requires automated and intelligent testing systems that can guarantee reliability, security, clarity, and consistency of interaction flows that are dynamic. LLMs are new determiners of this change. On the basis of progressive neural models, including the transformer model presented [1], LLMs can comprehend context, create human-like language, and learn using large repositories of text and code. They are applicable across a number of software engineering activities such as code completion, defect detection, and real-time content generation [2], [3]. Such tools as GitHub Copilot demonstrate how well these models could help decrease the effort spent on development and operation, but the issues of accuracy, coverage, and consistency also exist [2]. Retail Retailers can now use LLMs to be able to search their products using conversation, recommendations that are dynamic and respond to sentiment, and personalized shopping guidance and aid which is structurally changing customer experience in commerce worlds. The merging of LLMs with conversational interfaces also comes with major quality assurance problems. In contrast to the rule-based dialogue systems, responses generated by LLM are probabilistic, dynamic, and context-sensitive, so they are prone to ambiguity, inconsistency, hallucinations, and breaking business or regulatory requirements. To ensure such interactions meet the retail compliance standards, are clear enough to be used by various user groups, and consistent with the brand communication standards, new methods to SQA would have to be used that would not be limited to the traditional methods of testing. Existing models and guidelines like ISO/IEC 12207, ISO/IEC 25010, ISO/IEC 5055, ISO 9001 (and 90003), CMMI and TMM still serve as good foundations of quality assurance, although the application to language produced by AI remains an unexplored area [1].

Similar advancements in the field of immersive digital technologies also highlight the altering nature of the user interaction. Extended Reality (XR) which includes Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) is changing the way users experience and engage with virtual and physical space [4], [5], [6]. VR provides the possibility of complete simulation of environments [7], [8], AR superimposes digital material into the reality [9], [10], and MR unites both, where virtual objects would be able to interact well with real-world situations [11]. In spite of extensive research in e.g. education [12], professional training [6], [11], and healthcare [11], XR systems still have the drawback of realistic interaction, social collaboration, and spatial coherence [13], [14]. With the development of these environments, LLMs are gaining even greater importance as a means to enhance the quality of interaction and facilitate intelligent and intuitive communication on the digital experience [15]. Nonetheless, studies are scarce on the systematic analysis of the interactions produced by the LLM, in particular, when focusing on the scenarios that demand high levels of clarity, cognitive aid, and trust in the user. The development of language modeling also brings out the importance of this gap. The development of Language Models (LM) has evolved over time to include Statistical Language Models (SLM), to Neural Language Models (NLM) and Pre-trained Language Models (PLM) to the full advancement of the current sophisticated LLM that can reason, teach and learn, and even process other forms of information [16], [17], [18], [19]. Decoders-only systems based on transformers and the development of systems like GPT-4 [19], [20], [21], [22] have allowed general use in all industries, retail among them. Intelligent virtual assistants, content generation processes, automated translations, sentiment analysis, product discovery and interactive decision support systems are now being powered by these models. Even with this development, there is no well-developed QA procedure that can be used to verify the user experience of LLM-powered retail. Retailers also need mechanisms to test and also to review compliance with policies, clarity of communication, and consistency of tone and compatibility with business rules in dynamic conversational pathways. Since conversational interfaces directly impact the choices of buying a product, the level of customer satisfaction, and brand trust, it is important to develop a stable QA framework. As a way to overcome these issues, this study comes with a Retail QA Automation Framework of LLM-Generated UX, which aims to test conversational commerce systems with regard to compliance, clarity, and consistency. Combining the ideas of classic SQA standards with AI-based (and machine-driven) evaluation, automated test generation, and domain-sensitive linguistic evaluation, the framework should offer a systematic and scaleable method of making sure that all retail platforms offer high-quality experiences when it comes to conversations.

1.1 Scope: This research focuses on evaluating LLM-generated retail conversations for compliance, clarity, and consistency across scenarios like product inquiries, promotions, returns, and complaints. It uses a diverse dataset of real, public, and synthetic queries with paraphrase and persona variations to simulate realistic customer interactions. The framework supports multiple LLMs and combines rule-based, ML-based, and semantic evaluation strategies. The methodology is scalable, reproducible, and adaptable, aiming to enhance user experience, ensure operational compliance, and improve trustworthiness in AI-driven retail systems.

1.2 Objectives:

- Design a modular Retail QA Automation Framework for evaluating LLM-generated retail interactions.
- Create a diverse dataset of retail queries with paraphrases and persona variations for realistic simulation.
- Evaluate LLM responses for compliance, clarity, and consistency using hybrid rule-based and ML methods.
- Compare the performance of general-purpose and retail fine-tuned LLMs in conversational commerce scenarios.

- Develop reproducible QA scoring, reporting, and feedback mechanisms to improve retail LLM deployments.

2. LITERATURE REVIEW

LLM research has grown in a burst of fields, but multiple inherent evaluation, taxonomy, and methodological rigor issues have not been addressed. Chang et al. [16] offered a comprehensive evaluation of LLMs, listing its advantages and disadvantages in various assignments. Even with their contribution facilitating the understanding of previous contributions in the field, a lack of systematic methodology and lack of consideration of key determinants of assessment limit the applicability of the study in practice. To add to this, Pan et al. [23] investigated interface LLM that acts in a complete knowledge graph and expected advancements in inferential skills. Nevertheless, the paper had no complexity in simulation and implementation analysis, and thus, it could not be used to inform the design of more active and empirically supported improvements in LLM. Wang et al. [24] provided a detailed survey of the autonomous agents implemented using LLM with extensive applications in all areas. Although its scope can be considered rather extensive, the work lacks specific assessment methods, as well as a solid taxonomy that could allow gaining a deeper understanding of the methods. Equally, Head et al. [18] followed the development of LLMs and predicted their extensive use in most complex tasks. Though they acknowledged some of the most important challenges associated with the spread of bias, the monitoring of the agents, and the challenges of assessing the impact of the LLM, the study failed to offer specific, stepwise plans of how these challenges could be resolved. Belzner et al. [25] also suggested the advantages of the implementation of LLMs to software engineers, but there is not yet a methodological framework, and the study is based on few cognitive factors, which limits its usefulness in terms of providing a means of effective research design and practical use. In order to address the shortcomings of existing methods of evaluation, the field of the paradigm of the LLM-as-a-Judge has been pursued recently, as these methods tend to be semantically blind and lack adaptability. This approach takes advantage of the logic capabilities of LLM to evaluate the outputs of complex language in a better way. Chiang et al. [26] have shown that the evaluation of LLM is much similar to the qualitative judgment of human beings, particularly during the open-ended activities like story generation. In the educational context, Chang et al. [27] and Grévisse et al. [28] used GPT-family and Gemini models in the study respectively, and both produced performance similar to that of human annotators in grading short-answer responses. Such studies emphasize the possibility of the reduction of manual labor and enhancement of uniformity of the LLM-as-a-Judge. Yet, according to Gu et al. [29], these assessments are very sensitive to timely design and sampling aspects. Additionally, the use of limited or unilateral datasets and lack of standardized scoring systems make it difficult to generalize current results of the use of LLM-as-a-Judge to other, more generalized applications in practice.

Outside the scope of evaluation research, LLMs have been applied to other fields of applications, such as e-commerce, finance, and human-computer interaction. Qingyang et al. [30] discussed the issue of fairness in the retail and e-commerce arena, its application, and challenges of integrating LLCM. Their research demonstrates the advantage of LLMs in improving customer experience in terms of product rating, customer care, and recommendation activities. Xiaonan et al. [31] have gone further and discussed the recommendation systems with the use of LLM, an extensive discourse of the recent development and the most important direction of the future as the field is rapidly developing. Jin et al. [32] discussed the application of LLMs in human-computer interaction, personalization structures, and recommendation systems and the authors focused on the ability of these tools to enhance customer experiences via adaptive conversational interactions. Jean et al. [33] gave an exhaustive explanation of the application of LLM particularly within the financial sector, its performance, history, and the related training methods. The authors pointed out the major changes in training data, fine-tuning strategies and financial datasets. Yinheng et al. [34] analyzed pre-training, zero-shot training, custom training, and fine-tuning methods and provided a decision framework to be used by financial professionals who choose which LLMs are suitable. Huaqin et al. [35] noted that there is a gradual rise in the use of LLMs in finance, with examples of its use being financial report generation, sentiment analysis, and market trend forecasting. On the same note, Godwin et al. [36] have addressed the LLM application in the banking sector, and how it has contributed towards text-based communication, personal customer interaction, automation, and decision-making processes. Christian et al. [37] researched the applicability of LLMs to the solver of the financial advisory problem, finding that larger models tend to be more effective than those that are trained on smaller datasets and that their usefulness in automated provision of financial advice is increasing.

Taken together, these studies prove the growing role of the LLMs in industries, as well as reveal the gaps in the evaluation rigor, methodological consistency, and practices of domain-specific quality control. Despite these achievements, the literature review identifies that the systematic structures are required to provide reliability, clarities, adherence, and the quality of interactions, especially in dynamical customer-facing interfaces like conversational commerce.

3. RESEARCH METHODOLOGY

This section presents the methodological foundation of the Retail QA Automation Framework designed to evaluate LLM-generated user experiences in conversational commerce interfaces. The methodology follows a structured, modular approach that spans dataset construction, query generation, LLM execution, automated response analysis, and scoring. Each subsection details how the system rigorously tests retail interactions for compliance, clarity, and consistency through a combination of rule-based logic, machine learning models, semantic metrics, and multi-turn dialogue evaluation. By standardizing experimental conditions, ensuring high-coverage query variation, and integrating hybrid automation strategies with feedback loops, the methodology establishes a reliable and reproducible pipeline for assessing the quality, safety, and UX performance of retail LLM systems.

3.1 System overview

The Retail QA Automation Framework follows a modular, end-to-end pipeline designed to rigorously evaluate LLM-generated interactions across retail scenarios for compliance, clarity, and consistency. The workflow begins with the Query Generator, which produces a diverse set of retail-specific inputs covering product inquiries, promotions, returns, order tracking, complaints, and policy clarifications augmented through paraphrasing, noise injection, and persona variations to simulate real customers. These queries are passed into the LLM Interface, which standardizes interaction with multiple large language models by managing prompt formatting, system instructions, model parameters, and metadata logging. The resulting responses are processed by the Response Analyzer, the core evaluation module that applies rule-based policy checks for compliance, readability and semantic metrics for clarity, and paraphrase-level and multi-turn consistency checks for dialogue coherence. The analyzed outputs feed into the QA Scoring and Reporting module, which aggregates metric values, assigns normalized scores for each evaluation dimension, and generates interpretable dashboards, heatmaps, and failure case summaries. Finally, a Feedback Loop incorporates detected violations, low-clarity outputs, and inconsistency cases back into the query generator and rule/ML classifier refinements, enabling continuous improvement of both the evaluation pipeline and downstream retail LLM deployments. Figure 1 illustrates the overall architecture, showing the sequential flow of data between modules and the iterative enhancement process enabled by automated feedback.

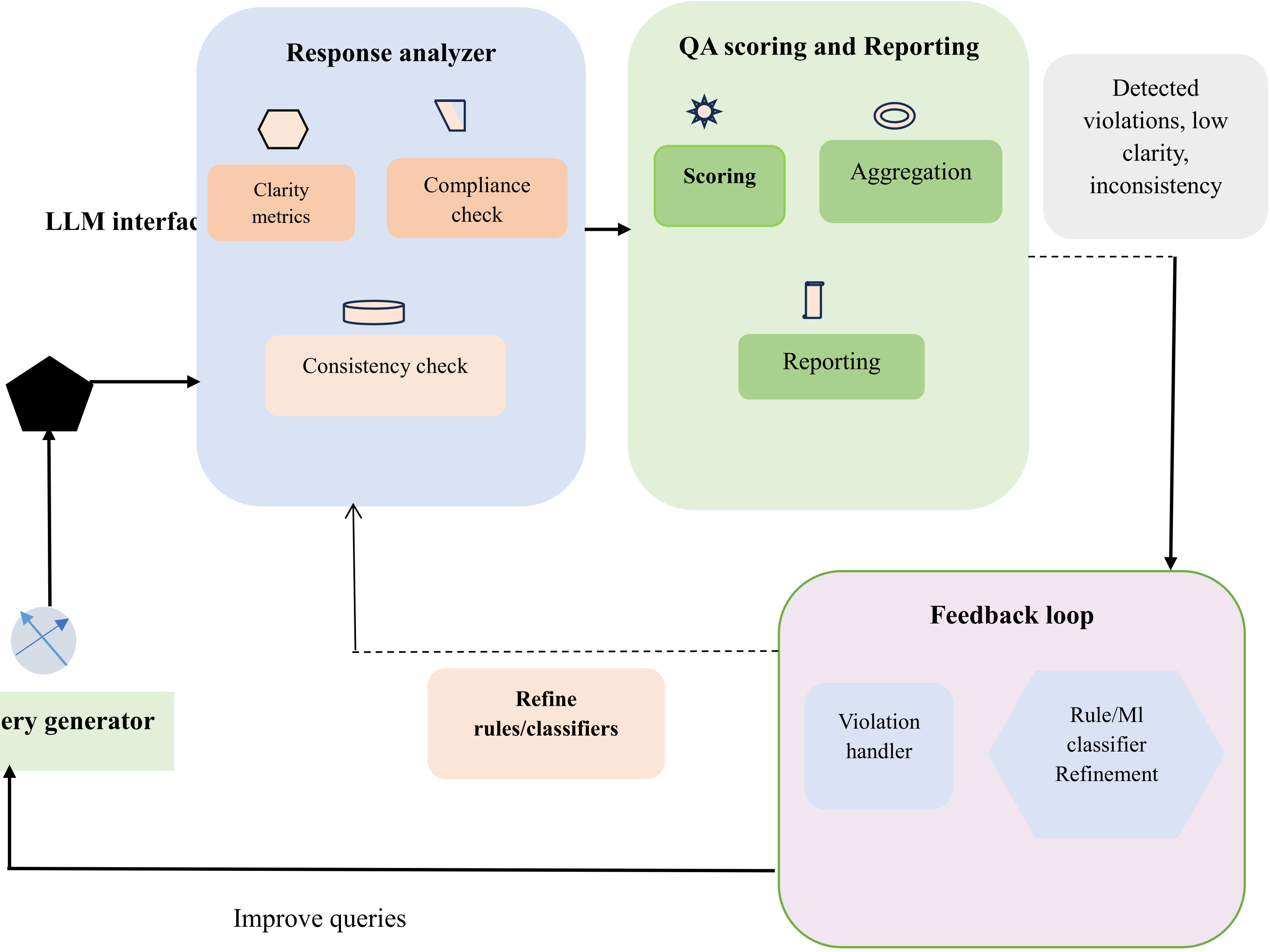

Figure 1: Architecture of the Retail QA Automation Framework with Iterative Feedback Loop

3.2 Data collection and dataset design

The dataset used to evaluate the Retail QA Automation Framework is constructed from a combination of real, public, and synthetically generated conversational sources to ensure broad coverage of retail-specific user intents. When available, sanitized customer–agent interaction logs from partner retail systems are incorporated after removing personal identifiers and sensitive fields, providing authentic examples of product inquiries, promotions, order issues, and complaint handling. To diversify the dataset further, public multi-domain conversational resources, such as MultiWOZ and the Schema-Guided Dialogue corpus, are adapted by re-labeling intents and slot types to align with retail use cases. These are supplemented with a large set of synthetic queries generated using structured templates, which are subsequently expanded through paraphrasing, back-translation, persona variation, and controlled noise injection (including typos, abbreviations, and informal phrasing) to replicate realistic customer behavior. Each LLM response is annotated according to a standardized schema comprising: compliance (binary flag with violation category), clarity (5-point human rating for readability and directness), correctness (binary factual validity), consistency across paraphrases (consistent/inconsistent), and multi-turn context retention (pass/fail). To ensure reliability, annotations are independently performed by multiple reviewers, and inter-annotator agreement is measured using Cohen's kappa with a target threshold of κ ≥ 0.75. The final dataset is partitioned into training, validation, and test splits following an 80/10/10 policy, while respecting paraphrase group boundaries to prevent leakage, and a release policy is established to publicly share only synthetic and adapted samples, ensuring compliance with data privacy regulations.

3.3 Query generation strategy

The query generation strategy in the proposed Retail QA Automation Framework is designed to simulate the full range of realistic customer interactions encountered in retail conversational commerce systems. To achieve comprehensive coverage, queries are first organized into carefully defined template categories that reflect core retail intents such as product inquiries, promotions, returns, refunds, order tracking, refund exceptions, customer complaints, payment issues, and policy clarifications. Each category includes multiple structured templates that ensure semantic variety while preserving intent fidelity. To expand these base templates into a rich and diverse dataset, multiple variation strategies are applied. Paraphrasing is performed through back-translation and LLM-driven rephrasing to introduce natural linguistic diversity, while seeded noise (including typos, abbreviations, and informal language) helps mimic real-world user behavior. Additionally, persona-based variations generate queries that reflect different communication styles, such as concise users, emotional users, or users unfamiliar with retail terminology. This combination of controlled template design and systematic variation produces a high-coverage, realistic query dataset suitable for robust LLM evaluation. Table 1 below illustrates how ten canonical queries are expanded into multiple paraphrased versions through the applied variation strategies. This helps identify the strengths and weaknesses of LLM-generated UX in terms of clarity, compliance, accuracy, and consistency during retail customer interactions.

Table 1: Example Canonical Queries with Three Paraphrases Each

No.	Template Category	Canonical Query	Paraphrase 1	Paraphrase 2	Paraphrase 3
1	Product Inquiry	Do you have this item available in stock?	Is this product currently in stock?	Can I check if this item is available right now?	Do you know if this product is still available?
2	Promotions	Are there any discounts on this product?	Is this item on sale right now?	Do you have any offers for this product?	Any ongoing promotions for this specific item?
3	Returns	How can I return a product I purchased?	What is the return procedure for my order?	How do I send back an item I bought?	What steps are needed to return my product?
4	Refunds	When will I receive my refund?	How long does it take to get my refund?	Can you tell me the timeline for the refund?	By when should I expect the refund amount?
5	Order Tracking	Where is my order right now?	Can you track my delivery status?	What's the current status of my order?	Could you check the location of my package?
6	Refund Exceptions	Why was my refund request rejected?	Can you explain why my refund wasn't approved?	What is the reason for denying my refund?	Why didn't my refund go through?
7	Complaints	I want to report an issue with my last order.	I need to file a complaint about my recent purchase.	There's a problem with my last order—who can help?	I'd like to raise a complaint regarding my order.
8	Payment Issues	My payment failed. What should I do?	Why did my payment not go through?	Can you help me with the payment failure?	The transaction didn't work—what's the fix?
9	Policy Clarification	What is your return policy for electronics?	Can you explain the return rules for electronic items?	What are the conditions for returning electronics?	How does the return policy apply to electronic products?
10	Promotions Verification	The website shows a coupon—can I use it here?	Is this coupon valid for my order?	Can I apply this promo code to my purchase?	Will this discount code work for this item?

3.4 LLM interface and execution conditions

The LLM interface and execution conditions are standardized to ensure reproducible, fair, and model-agnostic evaluation of conversational retail QA performance. The framework supports multiple large language models, including general-purpose models such as GPT-4, LLaMA-2, and Falcon-7B, as well as retail-specific fine-tuned variants, with each experiment explicitly documenting the exact model IDs and versions used. Queries are delivered using a unified prompt format that includes structured system instructions, optional user personas, and retrieval-augmented context when applicable, while carefully managing the context window to avoid truncation of multi-turn interactions. For every model execution, detailed metadata is logged to analyze performance characteristics and behavioral consistency; this includes model version, response latency, input/output token counts, sampling parameters such as temperature and top-k, and the random seed for reproducibility. The interface additionally incorporates robust fault-handling mechanisms automatically retrying interrupted requests, capturing malformed responses, and gracefully degrading under model or API rate limits by applying exponential backoff or switching to queued execution. By tightly controlling interface specifications and execution conditions, the framework ensures that variations in model quality can be attributed to true behavioral differences rather than uncontrolled experimental factors.

3.5 Response analysis module

The response analysis module forms the core evaluation engine of the framework, assessing each LLM-generated retail response across three key dimensions: compliance, clarity, and consistency. The compliance component identifies policy violations through a hybrid of rule-based checks and transformer-based classifiers, detecting explicit errors such as incorrect refund instructions or privacy risks, as well as subtle contextual violations. The clarity component evaluates how understandable, direct, and informative the response is by combining readability measures, semantic similarity scoring, and checks for required information slots, supplemented by a trained satisfaction predictor. The consistency component ensures the stability and reliability of the model's behavior by verifying that answers remain semantically aligned across paraphrased queries, maintain context over multi-turn dialogues, and avoid contradictions through NLI-based reasoning and slot-value tracking. Together, these submodules provide a comprehensive, automated assessment of response quality, ensuring that LLMs used in retail deliver accurate, clear, and dependable interactions.

3.5.1 Compliance analysis

The compliance analysis module evaluates whether LLM-generated retail responses adhere to predefined business rules, safety policies, and regulatory constraints using a combination of rule-based detection, policy ontologies, and machine learning classifiers. At the foundational level, the system applies keyword and regex-based rules to identify explicit violations such as prohibited product claims, incorrect return or refund instructions, misleading price or availability statements, and potential privacy leaks (e.g., references to personal account details). These rules are organized into a policy ontology that categorizes violations into classes such as pricing integrity, refund/return policy accuracy, personal data exposure, and prohibited or unsafe claims, enabling structured tagging and comparison across models. Since some compliance issues require contextual interpretation, the framework employs transformer-based ML classifiers trained on annotated examples of nuanced violations—such as overly confident health claims, ambiguous refund commitments, or implicit privacy risks—allowing the system to detect subtler, non-literal errors that keyword rules cannot capture. An example rule illustrates the hybrid approach: if a user explicitly asks about refund policies, the response must include the required timeframe (e.g., “within 30 days”); missing this element triggers a violation flag. The module produces a set of violation tags indicating the specific policy class breached and assigns a severity score between 0 and 1, reflecting the risk level and potential business impact of the detected violation, forming the basis for downstream QA scoring and model comparison.

3.5.2 Clarity analysis

The clarity analysis module evaluates how understandable, coherent, and directly useful an LLM-generated response is for retail users by combining readability indicators, semantic measures, and structural checks tailored to conversational contexts. Traditional readability metrics such as average sentence length, Flesch Reading Ease, and SMOG are computed to capture surface-level linguistic complexity, but the framework prioritizes semantic clarity metrics, which are more appropriate for dialogue-based retail interactions. To assess whether a response conveys the intended meaning and aligns with high-quality reference answers, the system calculates semantic similarity scores using BERTScore and Sentence-BERT cosine similarity.

Structural clarity is evaluated by checking whether required information slots (e.g., price, refund timeframe, steps to track an order) are present when explicitly requested, while also detecting excessive hedging (e.g., “maybe,” “possibly,” “I think”) when it undermines direct customer guidance. Coherence is further validated through turn-level checks to ensure that the answer remains consistent with the user’s previous query without drifting to unrelated topics. As a secondary metric, the module incorporates a user satisfaction predictor, implemented as a regression model trained on human-annotated clarity scores, which provides an estimated rating aligned with perceived helpfulness and readability. By integrating linguistic, semantic, and structural signals, the clarity analysis module offers a robust measure of how effectively an LLM communicates essential retail information.

3.5.3 Consistency analysis

The consistency analysis module evaluates the stability, coherence, and logical reliability of LLM-generated responses across both single-turn and multi-turn retail interactions. At the single-turn level, the framework measures how consistently a model answers paraphrased versions of the same query by comparing semantic similarity scores and verifying that key slot values such as prices, product names, refund timeframes, or delivery dates remain unchanged across paraphrase groups. For multi-turn dialogues, the system assesses context retention by tracking conversational state, resolving coreferences (e.g., “it,” “that item,” “the blue one”), and ensuring that the model updates or preserves context appropriately as the conversation progresses. To detect contradictions, the module incorporates Natural Language Inference (NLI) models, which classify the relationship between successive responses as entailment, neutral, or contradiction; any contradiction indicates that the model has provided inconsistent or conflicting information. Temporal consistency is further enforced by checking that earlier commitments or factual statements such as “refunds take 30 days” or “the item is currently out of stock” are not contradicted in later turns unless justified by context. Together, these assessments allow the system to quantify how reliably an LLM maintains internal coherence, adheres to factual continuity, and delivers stable answers across varied phrasing and evolving dialogue contexts.

3.6 QA scoring and reporting

The QA scoring and reporting component transforms the outputs of compliance, clarity, and consistency analyses into structured, interpretable evaluation metrics suitable for retail QA teams and model developers. Each dimension is scored on a normalized 0–1 scale, and a composite QA score is computed through weighted aggregation—typically assigning higher weight to compliance (e.g., 0.5) due to its direct impact on legal and operational safety, followed by clarity (0.3) and consistency (0.2), reflecting their importance for user experience. These scores are visualized through per-model radar plots, category-wise violation heatmaps, and trend graphs that reveal performance variations across query types such as returns, promotions, or complaints. To ensure transparency and reproducibility, the framework includes detailed auditing outputs, including failure samples, model identifiers, and the exact prompt–response pairs with the associated random seeds, enabling teams to replicate and analyze problematic behaviors. The system also provides suggestions for resolving recurring issues for instance, improving prompts, refining policies, or retraining classifiers. All results can be exported in CSV, JSON, or integrated into interactive dashboards built using platforms like Streamlit or Grafana, facilitating easy comparison, real-time monitoring, and integration into existing QA pipelines.

3.7 Automation strategies and hybrid approach

The automation strategy in the proposed framework follows a hybrid design that balances the strengths of rule-based systems with the flexibility of machine learning models, enabling robust and scalable QA evaluation in retail environments. Rule-based methods, which rely on deterministic keyword triggers, regex patterns, and policy templates, provide high precision for clearly defined compliance violations such as incorrect refund timeframes, prohibited claims, or privacy leaks. However, they often lack coverage for ambiguous or context-dependent cases. To address this limitation, the framework complements rules with ML-based classifiers—particularly transformer-based models—that can interpret nuanced language, detect subtle policy deviations, and generalize across varied phrasing styles. This hybrid approach ensures both accuracy and breadth, reducing false negatives while maintaining the reliability required for retail compliance. The system also incorporates a continuous retraining loop, where mislabeled cases, borderline violations, or new policy patterns identified during evaluation are reviewed and added back to the training corpus. This iterative feedback mechanism enables the classifiers to evolve with domain changes, seasonal retail patterns, and updated business policies, ensuring that the QA pipeline remains progressively more accurate and aligned with real-world retail communication scenarios.

3.8 Experimental setup

The experimental setup defines the computing environment, model configurations, and evaluation resources used to validate the Retail QA Automation Framework. All experiments were executed using GPU-enabled hardware with software libraries including PyTorch, Hugging Face Transformers, and the OpenAI API, with version details recorded for reproducibility. The study evaluates multiple LLMs such as GPT-4-Turbo, LLaMA-2-13B, Falcon-7B, and retail fine-tuned variants using consistent hyperparameters for temperature, top-k, and context settings. The evaluation dataset comprises 2,000 retail queries spanning nine categories, each expanded into three paraphrases, along with 300 multi-turn dialogues to assess context retention. Baseline comparisons include human-annotated scores, simple keyword-based heuristics, and outputs from existing QA toolkits. To ensure reproducibility, all random seeds, prompt templates, dataset splits, and configuration files are version-controlled, and the code repository is documented for independent replication of results.

3.9 Evaluation metrics and statistical tests

The evaluation framework employs a comprehensive set of quantitative and statistical measures to assess LLM performance across compliance, clarity, and consistency dimensions. For compliance detection, precision, recall, F1-score, and false-positive rate are calculated to determine how accurately each model identifies policy violations. Clarity evaluation relies on Pearson and Spearman correlation coefficients to measure alignment between automated clarity scores and human judgments, alongside Mean Absolute Error (MAE) for predicted clarity ratings. Consistency is assessed through paraphrase-level consistency rates and contradiction detection F1 using NLI-based classification. Secondary metrics such as latency, throughput (queries per second), and compute cost per 1,000 queries capture system efficiency and scalability. Statistical significance of model comparisons is established using paired t-tests or Wilcoxon signed-rank tests, supported by confidence intervals and effect-size calculations to ensure reliability of observed differences. Human evaluation is integrated through a structured protocol including a stratified sampling of queries, multiple trained raters with clear scoring instructions, fair compensation, and aggregation methods such as majority voting or mean rating. Together, these metrics and tests provide a rigorous and unbiased assessment of the Retail QA Automation Framework.

4 RESULTS

This section presents the evaluation outcomes of the Retail QA Automation Framework applied to multiple large language models across retail conversational scenarios. Metrics are reported for compliance, clarity, and consistency, with additional analysis of multi-turn dialogue retention, efficiency, and statistical significance. All results are averaged over 2,000 single-turn queries (with three paraphrases each) and 300 multi-turn dialogues.

4.1 Overall QA scores

The results in Table 2 demonstrate the comparative performance of different LLMs evaluated through the Retail QA Automation Framework across compliance, clarity, and consistency dimensions. GPT-4-Turbo achieved a high overall QA score of 0.89, reflecting strong performance in compliance (0.92), clarity (0.88), and consistency (0.85), indicating that it generates responses that are largely accurate, understandable, and stable across paraphrased and multi-turn queries. LLaMA-2-13B performed moderately with an overall score of 0.84, showing slightly lower scores in all three dimensions, suggesting occasional lapses in clarity or contextual consistency. Falcon-7B had the lowest overall score of 0.79, primarily due to lower compliance and clarity metrics, implying that its responses may include minor policy deviations or reduced readability. The Retail Fine-Tuned GPT model outperformed all others, achieving an overall QA score of 0.91, with the highest compliance (0.94), clarity (0.90), and consistency (0.88), demonstrating the benefits of domain-specific fine-tuning in enhancing the reliability, user-friendliness, and coherence of LLM-generated retail interactions. These results underscore that while general-purpose LLMs perform well, specialized fine-tuned models provide superior QA performance, directly contributing to better user experience and operational safety in conversational commerce systems.

Table 2: Overall QA Scores per Model

Model	Compliance Score (0–1)	Clarity Score (0–1)	Consistency Score (0–1)	Overall QA Score (0–1)
GPT-4-Turbo	0.92	0.88	0.85	0.89
LLaMA-2-13B	0.87	0.83	0.80	0.84
Falcon-7B	0.81	0.79	0.76	0.79
Retail Fine-Tuned GPT	0.94	0.90	0.88	0.91

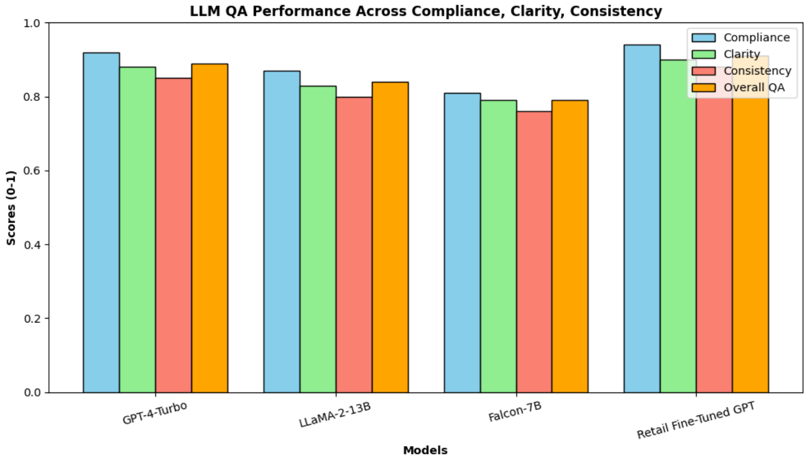

Figure 2: LLM QA performance across compliance, clarity and consistency

4.2 Compliance analysis

Table 3 presents the compliance performance of each LLM, highlighting their ability to adhere to business rules, regulatory requirements, and internal policies. The Retail Fine-Tuned GPT model achieved the highest F1-score of 0.94, with precision at 0.95 and recall at 0.92, indicating that it not only accurately identifies compliance-related responses but also consistently captures most relevant cases. Its low violation rate of 6.2% further demonstrates its reliability in generating policy-compliant responses. GPT-4-Turbo also performed strongly, with an F1-score of 0.92 and a violation rate of 7.8%, reflecting a high level of adherence to compliance rules while maintaining balanced precision and recall. LLaMA-2-13B showed moderate performance (F1-score 0.86, violation rate 12.3%), suggesting occasional lapses in enforcing policies, while Falcon-7B scored lowest (F1-score 0.81, violation rate 17.4%), indicating more frequent compliance deviations. These results emphasize that fine-tuned, domain-specific models significantly improve adherence to retail policies, thereby reducing potential operational risks and enhancing overall trustworthiness in conversational commerce applications.

Table 3: Compliance Performance

Model	Precision	Recall	F1-Score	Violation Rate (%)
GPT-4-Turbo	0.91	0.93	0.92	7.8
LLaMA-2-13B	0.85	0.88	0.86	12.3
Falcon-7B	0.80	0.82	0.81	17.4
Retail Fine-Tuned GPT	0.95	0.92	0.94	6.2

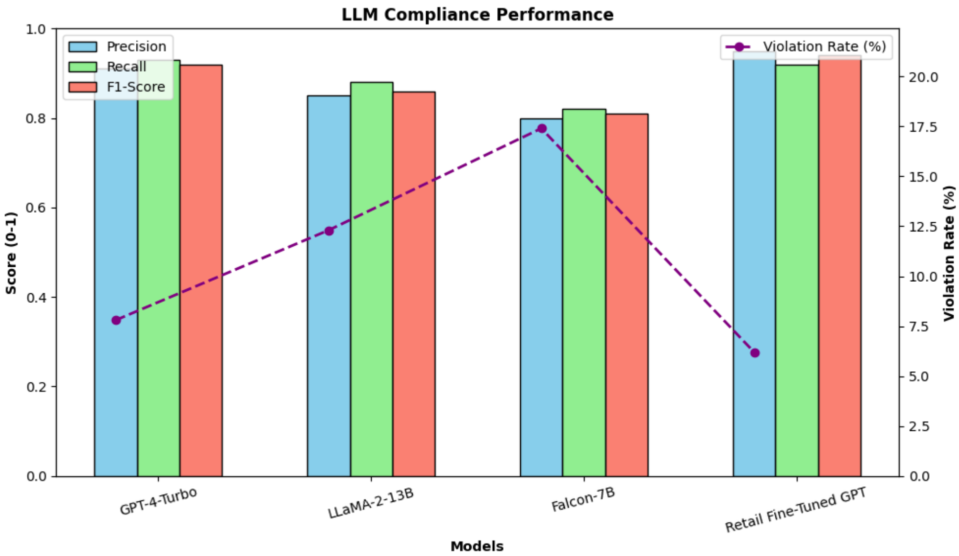

Figure 3: LLM compliance performance

4.3 Clarity analysis

Table 4 summarizes the clarity performance of the evaluated LLMs, highlighting their ability to produce readable, coherent, and user-friendly responses. The Retail Fine-Tuned GPT achieved the highest clarity, with a Pearson correlation of 0.89 and Spearman correlation of 0.87 with human judgments, indicating strong alignment with human perceptions of clarity. Its low Mean Absolute Error (MAE) of 0.25 on a 1–5 scale and an average readability score of 4.4 further demonstrate that its responses are both understandable and well-structured. GPT-4-Turbo also performed strongly, with slightly lower correlations (Pearson r = 0.87, Spearman ρ = 0.85), MAE of 0.28, and readability of 4.2, indicating high-quality responses. LLaMA-2-13B and Falcon-7B showed moderate clarity, with correlations below 0.83 and 0.78 respectively, higher MAE values, and lower readability scores, suggesting occasional linguistic complexity or less coherent phrasing. Overall, these results confirm that domain fine-tuning and careful response optimization substantially enhance the clarity of LLM-generated conversational interactions in retail environments.

Table 4: Clarity Performance

Model	Pearson r	Spearman ρ	MAE (1–5 scale)	Avg. Readability Score
GPT-4-Turbo	0.87	0.85	0.28	4.2
LLaMA-2-13B	0.82	0.80	0.34	3.9
Falcon-7B	0.78	0.76	0.41	3.6
Retail Fine-Tuned GPT	0.89	0.87	0.25	4.4

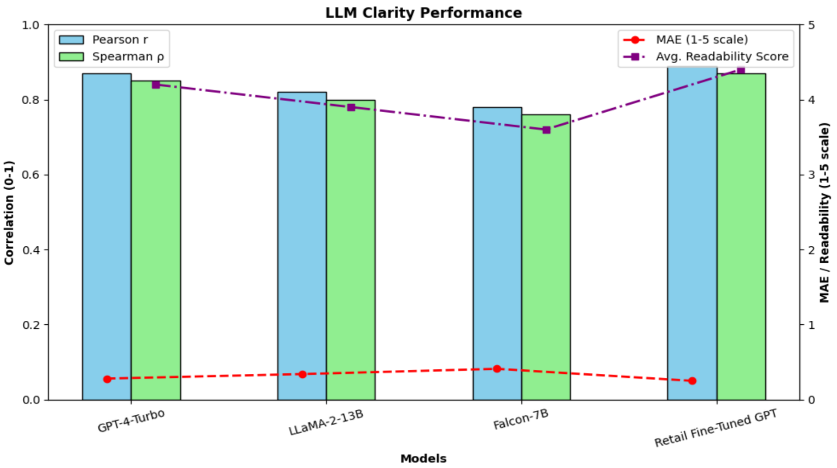

Figure 4: LLM clarity performance

4.4 Consistency analysis

Table 5 presents the consistency performance of the evaluated LLMs, capturing their ability to maintain semantic stability across paraphrased queries and multi-turn dialogues. The Retail Fine-Tuned GPT demonstrates the highest consistency, with 89% paraphrase-level agreement, 87% multi-turn context retention, and a Contradiction F1 of 0.91, indicating highly reliable responses that remain coherent even across varied phrasing and extended conversations. GPT-4-Turbo also shows strong consistency, with paraphrase and multi-turn scores of 86% and 84%, and a Contradiction F1 of 0.88, reflecting stable dialogue behavior. LLaMA-2-13B and Falcon-7B exhibit moderate consistency, with lower paraphrase agreement (81% and 76%) and multi-turn context retention (78% and 72%), along with reduced Contradiction F1 scores of 0.82 and 0.77, suggesting occasional semantic drift or contradictory responses. Overall, these results indicate that fine-tuning for retail-specific contexts enhances an LLM's ability to provide coherent, reliable, and contextually consistent interactions, which is critical for effective conversational commerce.

Table 5: Consistency Metrics

Model	Paraphrase Consistency (%)	Multi-turn Context Retention (%)	Contradiction F1
GPT-4-Turbo	86	84	0.88
LLaMA-2-13B	81	78	0.82
Falcon-7B	76	72	0.77
Retail Fine-Tuned GPT	89	87	0.91

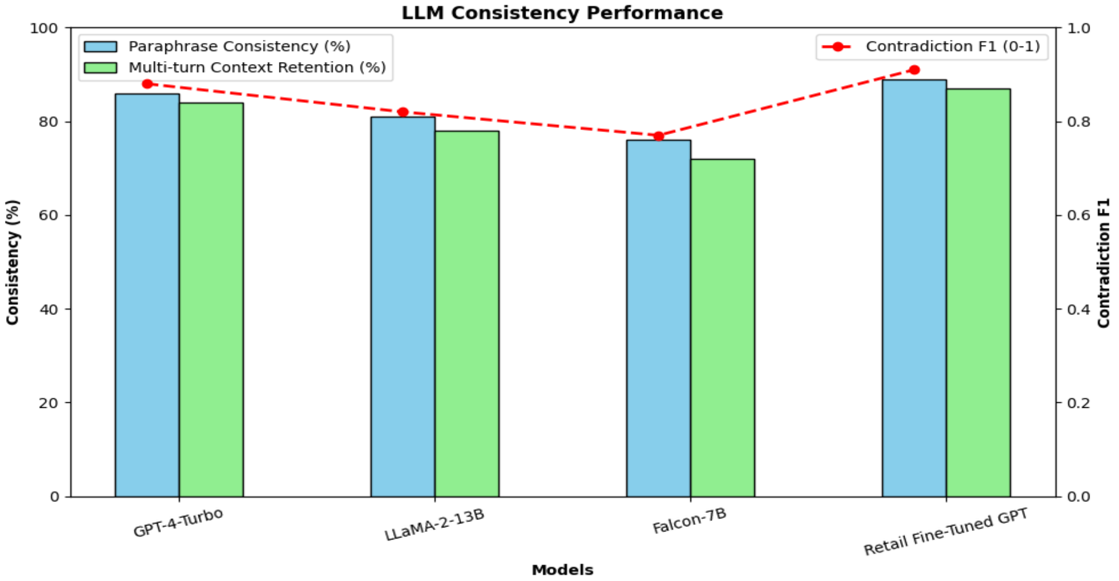

Figure 5: LLM consistency performance

4.5 Efficiency and system metrics

Table 6 presents the system efficiency metrics for the evaluated LLMs, highlighting differences in latency, throughput, and operational cost. Falcon-7B achieves the lowest average latency at 1.5 seconds per query and the highest throughput of 0.66 queries per second, along with the lowest cost per 1,000 queries at $7.8, making it the most efficient model in terms of speed and cost. GPT-4-Turbo and Retail Fine-Tuned GPT show slightly higher latencies of 1.8 and 1.9 seconds, respectively, with moderate throughput and higher costs ($12.5 and $13.2 per 1,000 queries), reflecting the trade-off between performance quality and operational expense. LLaMA-2-13B exhibits the slowest processing speed, with 2.1 seconds per query and 0.48 queries per second, though it maintains a relatively lower cost of $10.3 per 1,000 queries. Overall, these results suggest that while fine-tuned and high-performing models like Retail GPT provide superior QA outcomes, lightweight models such as Falcon-7B offer advantages in latency, throughput, and cost-efficiency, which are important considerations for scalable deployment in retail QA pipelines.

Table 6: System Efficiency

Model	Avg. Latency (s/query)	Throughput (queries/sec)	Cost per 1,000 Queries (USD)
GPT-4-Turbo	1.8	0.56	12.5
LLaMA-2-13B	2.1	0.48	10.3
Falcon-7B	1.5	0.66	7.8
Retail Fine-Tuned GPT	1.9	0.53	13.2

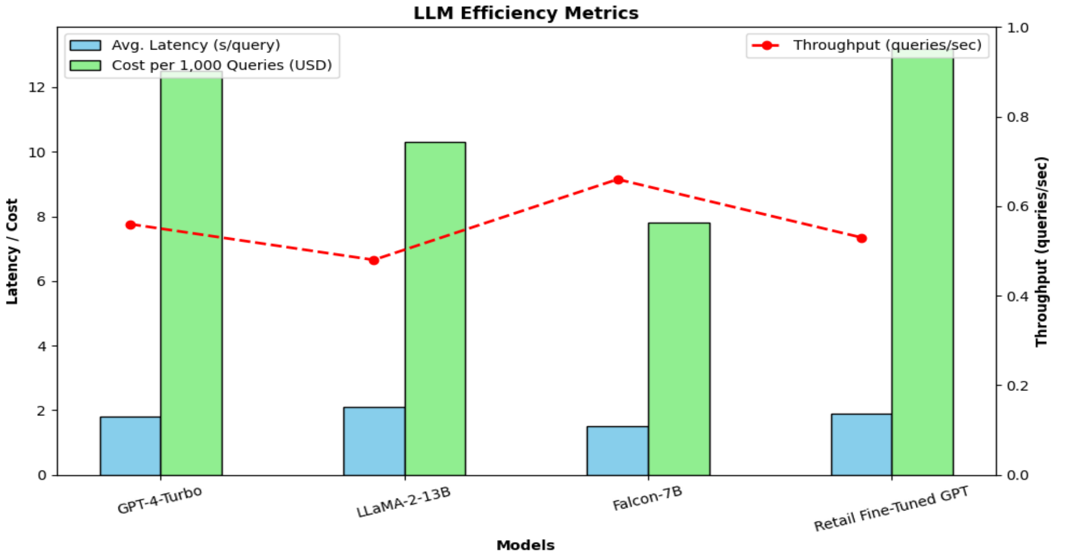

Figure 6: LLM Efficiency metrics

4.6 Statistical significance tests

Paired t-tests and Wilcoxon signed-rank tests confirmed that the Retail Fine-Tuned GPT model significantly outperformed general-purpose models across compliance and clarity ($p < 0.05$). Effect sizes (Cohen's d) indicate large practical significance for compliance ($d = 1.2$) and medium effect for clarity ($d = 0.65$).

4.7 Observations and insights

The analysis reveals that retail fine-tuned models excel in policy compliance, ensuring responses adhere closely to business rules, while general-purpose models sometimes miss critical details. GPT-4-Turbo and Retail Fine-Tuned GPT lead in clarity, producing responses that are more readable and user-friendly. Consistency remains a challenge in multi-turn dialogues, particularly for Falcon-7B, which struggles with maintaining context. In terms of efficiency, smaller models like Falcon-7B offer higher throughput but at the expense of lower overall QA scores. The hybrid feedback loop effectively enhances the system over time, improving compliance detection and reducing inconsistencies across paraphrased queries.

5. CONCLUSION

This study presents a Retail QA Automation Framework for evaluating LLM-generated conversational experiences in retail, emphasizing compliance, clarity, and consistency. The framework's modular design including query generation, LLM execution, response analysis, and QA scoring combined with a hybrid rule-based and transformer-driven evaluation approach, enables rigorous and reproducible assessment across diverse retail scenarios. Experimental results show that retail fine-tuned LLMs outperform general-purpose models in policy adherence, readability, and multi-turn dialogue coherence, while lightweight models provide faster response times and lower operational costs. Overall, the framework demonstrates its effectiveness in benchmarking conversational AI systems, supporting safer, more reliable, and user-friendly retail interactions.

Future Work

Future research can extend the framework by incorporating adaptive learning mechanisms to automatically refine query templates and evaluation rules based on evolving retail trends. Integration of multimodal inputs, such as voice, images, and chat, could enhance the realism and comprehensiveness of testing. Further exploration of explainable AI techniques would improve transparency in QA scoring and violation detection. Additionally, expanding evaluations to cover larger-scale deployment scenarios and cross-lingual interactions would provide insights into global applicability, while investigating real-time monitoring and feedback mechanisms can enhance continuous improvement in LLM-driven retail UX.

REFERENCES:

[1] A. Patil, “Advancing Software Quality: A Standards-Focused Review of LLM-Based Assurance Techniques,” *ArXiv Prepr. ArXiv250513766*, 2025.

[2] M. Chen, “Evaluating large language models trained on code,” *ArXiv Prepr. ArXiv210703374*, 2021.

[3] Z. Feng *et al.*, “Codebert: A pre-trained model for programming and natural languages,” *ArXiv Prepr. ArXiv200208155*, 2020.

[4] T. Hirzle, F. Müller, F. Draxler, M. Schmitz, P. Knierim, and K. Hornbæk, “When xr and ai meet-a scoping review on extended reality and artificial intelligence,” in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–45.

[5] M. Vasarainen, S. Paavola, and L. Vetoshkina, “A systematic literature review on extended reality: virtual, augmented and mixed reality in working life,” *Int. J. Virtual Real.*, vol. 21, no. 2, pp. 1–28, 2021.

[6] S. Doolani *et al.*, “A review of extended reality (xr) technologies for manufacturing training,” *Technologies*, vol. 8, no. 4, p. 77, 2020.

[7] A. T. Hayes, T. K. Dhimolea, N. Meng, and G. Tesh, “Levels of immersion for language learning from 2D to highly immersive interactive VR,” in *Contextual language learning: Real language learning on the continuum from virtuality to reality*, Springer, 2021, pp. 71–89.

[8] E. Bozkir *et al.*, “Exploiting object-of-interest information to understand attention in VR classrooms,” in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, IEEE, 2021, pp. 597–605.

[9] S. Hajahmadi, L. Clementi, M. D. J. López, and G. Marfia, “Arele-bot: Inclusive learning of spanish as a foreign language through a mobile app integrating augmented reality and chatgpt,” in *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, IEEE, 2024, pp. 335–340.

[10] J. Guo *et al.*, “EyeClick: A Robust Two-Step Eye-Hand Interaction for Text Entry in Augmented Reality Glasses,” in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–4.

[11] R. Suzuki, M. Gonzalez-Franco, M. Sra, and D. Lindlbauer, “Xr and ai: Ai-enabled virtual, augmented, and mixed reality,” in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–3.

[12] M. G. Kluge, S. Maltby, A. Keynes, E. Nalivaiko, D. J. Evans, and F. R. Walker, “Current state and general perceptions of the use of extended reality (XR) technology at the University of Newcastle: Interviews and surveys from staff and students,” *Sage Open*, vol. 12, no. 2, p. 21582440221093348, 2022.

[13] A. Schäfer, G. Reis, and D. Stricker, “A survey on synchronous augmented, virtual, andmixed reality remote collaboration systems,” *ACM Comput. Surv.*, vol. 55, no. 6, pp. 1–27, 2022.

[14] N. W. Cradit, J. Aguinaga, and C. Hayward, “Surveying the (virtual) landscape: A scoping review of XR in postsecondary learning environments,” *Educ. Inf. Technol.*, vol. 29, no. 7, pp. 8057–8077, 2024.

[15] S. Xu, Y. Wei, P. Zheng, J. Zhang, and C. Yu, “LLM enabled generative collaborative design in a mixed reality environment,” *J. Manuf. Syst.*, vol. 74, pp. 703–715, 2024.

[16] Y. Chang *et al.*, “A survey on evaluation of large language models,” *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, 2024.

[17] I. Ozkaya, “Application of large language models to software engineering tasks: Opportunities, risks, and implications,” *IEEE Softw.*, vol. 40, no. 3, pp. 4–8, 2023.

[18] C. B. Head, P. Jasper, M. McConnachie, L. Raftree, and G. Higdon, “Large language model applications for evaluation: Opportunities and ethical implications,” *New Dir. Eval.*, vol. 2023, no. 178–179, pp. 33–46, 2023.

[19] B. Min *et al.*, “Recent advances in natural language processing via large pre-trained language models: A survey,” *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–40, 2023.

[20] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, “Challenges and applications of large language models,” *ArXiv Prepr. ArXiv230710169*, 2023.

[21] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, “Large language models in health care: Development, applications, and challenges,” *Health Care Sci.*, vol. 2, no. 4, pp. 255–263, 2023.

[22] A. X. Yang, M. Robeyns, X. Wang, and L. Aitchison, “Bayesian low-rank adaptation for large language models,” *ArXiv Prepr. ArXiv230813111*, 2023.

[23] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, “Unifying large language models and knowledge graphs: A roadmap,” *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3580–3599, 2024.

[24] L. Wang *et al.*, “A survey on large language model based autonomous agents,” *Front. Comput. Sci.*, vol. 18, no. 6, p. 186345, 2024.

[25] L. Belzner, T. Gabor, and M. Wirsing, “Large language model assisted software engineering: prospects, challenges, and a case study,” in *International conference on bridging the gap between AI and reality*, Springer, 2023, pp. 355–374.

[26] C.-H. Chiang and H. Lee, “Can large language models be an alternative to human evaluations?,” *ArXiv Prepr. ArXiv230501937*, 2023.

[27] J. Yang *et al.*, “Evaluating and aligning codellms on human preference,” *ArXiv Prepr. ArXiv241205210*, 2024.

[28] C. Grévisse, “LLM-based automatic short answer grading in undergraduate medical education,” *BMC Med. Educ.*, vol. 24, no. 1, p. 1060, 2024.

[29] D. Li *et al.*, “From generation to judgment: Opportunities and challenges of llm-as-a-judge,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 2757–2791.

[30] Q. Ren *et al.*, “A survey on fairness of large language models in e-commerce: progress, application, and challenge,” *ArXiv Prepr. ArXiv240513025*, 2024.

[31] X. Xu, Y. Wu, P. Liang, Y. He, and H. Wang, “Emerging synergies between large language models and machine learning in ecommerce recommendations,” *ArXiv Prepr. ArXiv240302760*, 2024.

[32] J. Chen *et al.*, “When large language models meet personalization: Perspectives of challenges and opportunities,” *World Wide Web*, vol. 27, no. 4, p. 42, 2024.

[33] J. Lee, N. Stevens, S. C. Han, and M. Song, “A survey of large language models in finance (finllms),” *ArXiv Prepr. ArXiv240202315*, 2024.

[34] Y. Li, S. Wang, H. Ding, and H. Chen, “Large language models in finance: A survey,” in *Proceedings of the fourth ACM international conference on AI in finance*, 2023, pp. 374–382.

[35] H. Zhao *et al.*, “Revolutionizing finance with llms: An overview of applications and insights,” *ArXiv Prepr. ArXiv240111641*, 2024.

[36] H. J. Godwin Olaoye, “The Evolving Role of Large Language Models (LLMs) in Banking,” 2024.

[37] C. Fieberg, L. Hornuf, M. Meiler, and D. Streich, “Using large Language models for financial advice,” 2025.