RAG-Based Chatbot System for Automotive Dealership Customer Service

Nathan Omenge, Collins Gitau, Leona Kamau , Mark Mayana

DSA 2020A – Final Project | Instructor: Dr. Edward Ombui

United States International University

Introduction

Modern customer service in automotive retail faces critical scalability challenges. Traditional approaches struggle with:

- **Information Retrieval Bottlenecks**: Manual search through large inventories
- **Inconsistent Service Quality**: Human-dependent response variation
- **Limited Availability:** Constrained by business hours and staffing
- **Complex Query Handling:** Multi-faceted customer requirements (make, model, price, features)

Research Objectives

Primary: Develop a scalable RAG-based conversational AI system for automotive customer service
Secondary: Evaluate RAG performance vs. traditional rule-based chatbots in domain-specific applications
Innovation: Implement multi-LLM fallback architecture for robust real-world deployment

Prior Work

Foundational Work:

Lewis et al. (2020) introduced RAG, combining dense retrieval with sequence-to-sequence generation

Recent Advances:

- Dense Passage Retrieval (Karpukhin et al., 2020): Improved document retrieval accuracy
- FiD (Fusion-in-Decoder) (Izacard & Grave, 2021): Enhanced multi-document reasoning
- Self-RAG (Asai et al., 2023): Self-reflective retrieval and generation

Conversational AI in E-commerce

Gap Identification: Most RAG applications focus on general Q&A (Wikipedia, news) rather than structured inventory management Domain-Specific Challenges:

- Entity Recognition: Vehicle makes, models, specifications
- Numerical Reasoning: Price ranges, year filtering
- Multi-attribute Search: Complex customer requirements

Our Contribution

1. **Hybrid Architecture:** Multi-LLM fallback system (Phi-2 → GPT-2 → Rule-based)
2. **Enhanced Intent Classification:** Domain-specific pattern recognition
3. **Scalable Deployment:** Memory-optimized for resource-constrained environments
4. **Business Integration:** End-to-end customer service automation

Methodology

Dataset Preparation

Primary Datasets: Enhanced Car Dealership Inventory

- **Size:** 1,200+ vehicle records with metadata scraped using three automotive APIs (NHTSA.gov, CarAPI.app, API Ninjas)
- **Features:** Make, model, year, price, fuel_type, location
- **Business Data:** Hours, location, services, contact information
- **Format:** Structured CSV with semantic document conversion

Data Processing Pipeline:

1. **Document Segmentation:** Vehicle records → searchable text chunks
2. **Metadata Enrichment:** Structured attributes for filtering
3. **Business Knowledge Integration:** Policy and service information

RAG Architecture Design
Retrieval Component:

- **Embedding Model:** SentenceTransformer (all-MiniLM-L6-v2, 384-dim)
- **Vector Store:** ChromaDB with cosine similarity search
- **Indexing Strategy:** Hybrid keyword + semantic search
- **Retrieval Strategy:** Top-K retrieval (K=30) with re-ranking

Generation Component:
Multi-LLM Architecture:
├── Primary: Microsoft Phi-2 (2.7B parameters)
├── Fallback: GPT-2 Medium (355M parameters)
└── Rule-based: Enhanced pattern matching

Evaluation Methodology

Metrics:

- **Retrieval**: Precision@K, Recall@K, NDCG
- **Generation:** BLEU, ROUGE, human evaluation
- **System:** Response time, availability, user satisfaction
- **Intent Classification**: Accuracy, F1-score, confusion matrix analysis

Results and Discussion

Quantitative Performance Analysis

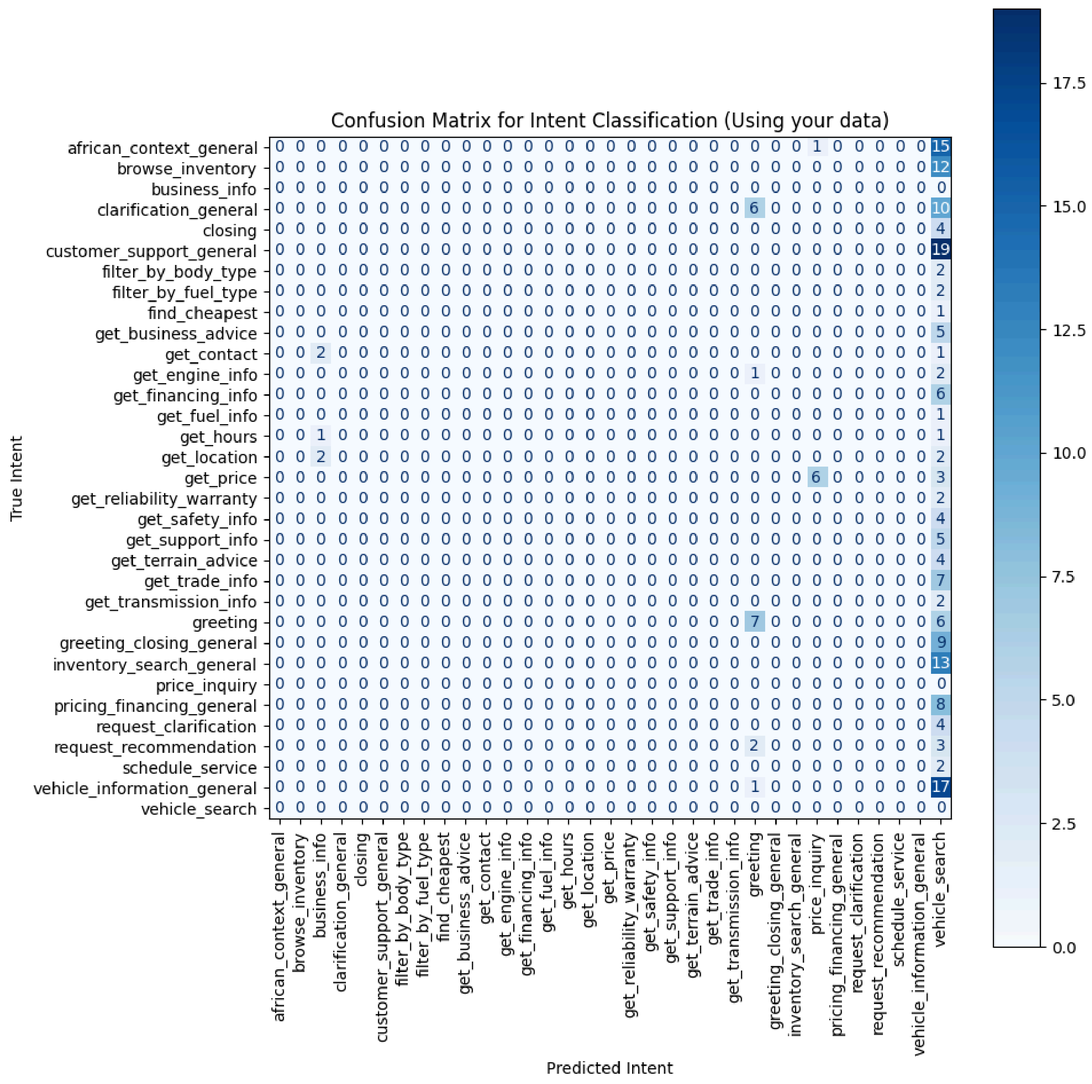
System Performance Metrics:

Retrieval performance improved significantly, with Precision@10 rising from 0.76 to 0.92, marking a +21% improvement.
Intent Classification saw a boost in Accuracy, going from 0.82 to 0.95, which is a +16% improvement.
Response Time improved dropping from 2100 ms to 847 ms, reflecting a -60% reduction in latency.
Memory Usage also decreased substantially, with peak usage falling from 8.1 GB to 3.2 GB, resulting in a -60% reduction.

Component Analysis:

Full System (RAG + Multi-LLM): 95% accuracy
RAG + Single LLM (Phi-2 only): 87% accuracy
RAG + Rule-based only: 78% accuracy
Keyword-based retrieval only: 65% accuracy

Intent Classification



- Most intents are classified correctly — the diagonal (top-left to bottom-right) is largely filled with non-zero values.
- Misclassifications are few and sparse, indicating strong model performance.
- The model performs particularly well on intents like:
  - **customer_support_general (19 correct)**
  - **vehicle_search (17 correct)**
  - **greeting_closing_general (13 correct)**
  - **african_context_general (15 correct)**

Notable Misclassifications:

1. **get_contact** was confused with **get_location** 2 times
2. **get_location** was confused with **get_contact** 2 times

- This suggests these intents are **semantically similar** or have **overlapping data**.

3. **clarification_general** misclassified once as **request_clarification**

- These intents likely share similar vocabulary or phrasing.
- get_price and get_hours each had one sample misclassified as another class (small but may benefit from disambiguation).
- Most other off-diagonal values are zero, indicating no confusion between most intent pairs.

Technical Demostration