AI-Powered Customer Support Chatbot with Multimodal Capabilities - Technical Specification

Technical Specification Document

Project Title: AI-Powered Customer Support Chatbot with Multimodal Capabilities

Version: 1.0

Date: May 20, 2025

Prepared by: AI Systems Design Team

Table of Contents

AI-Powered Customer Support Chatbot with Multimodal Capabilities - Technical Specification

---

1. Introduction

This document outlines the technical specifications for the development and deployment of an AI-powered customer support chatbot. The system is designed to handle multimodal input (text, image) and provide intelligent, context-aware responses in real-time across various customer service scenarios.

2. Project Overview

The chatbot leverages Natural Language Processing (NLP), Computer Vision (CV), and Retrieval-Augmented Generation (RAG) techniques. It supports interactions via website chat, mobile app, and social media platforms. Key use cases include resolving customer queries, identifying issues from product images, and providing knowledge base answers.

3. Objectives

- Automate 80% of customer service queries.
- Support multimodal inputs (text and image).
- Integrate with existing CRM and ticketing systems.
- Reduce average response time to under 2 seconds.

4. System Architecture

- Frontend: React.js-based web widget, mobile SDK.
- Backend: Python FastAPI microservices.
- Model Layer:
  - Text: OpenAI GPT-4 Turbo via API.
  - Vision: Azure Custom Vision or CLIP model.
  - RAG: LangChain + FAISS Vector DB.

- Database: PostgreSQL for structured data, Azure Blob for media.

- DevOps: Docker, Kubernetes, GitHub Actions, Azure DevOps.

5. Functional Requirements

- FR1: Accept user queries via chat interface.

- FR2: Upload and process images.

- FR3: Use RAG to retrieve relevant context.

- FR4: Generate response using AI model.

- FR5: Log interactions in CRM.

- FR6: Escalate complex cases to human agents.

6. Non-Functional Requirements

- NFR1: Response latency < 2 seconds.

- NFR2: Uptime 99.9%.

- NFR3: Scalability for 10,000+ concurrent users.

- NFR4: GDPR and CCPA compliance.

7. Data Requirements

- Historical chat logs for fine-tuning.

- Product manuals and documentation.

- Annotated image dataset (faulty vs. normal).

- Customer profiles and ticket metadata.

8. Model Specifications

- Text Model: GPT-4 Turbo (OpenAI API).

- Vision Model: OpenAI GPT-4V or CLIP, with fine-tuned layers.

- RAG: LangChain + FAISS + Azure Document Intelligence.

- Pipeline:

  - Query -> Context Retrieval -> Prompt Construction -> Response Generation.

9. APIs and Integrations

- CRM Integration: Salesforce API.

- Ticketing: Zendesk API.

- Cloud Vision: Azure Custom Vision.

- Auth: OAuth 2.0, JWT-based sessions.

- Webhooks: For alerts and logging.

10. Deployment Strategy

- Staging and Production clusters.

- CI/CD via GitHub Actions.

- Canary deployments for model updates.

- Azure Kubernetes Service (AKS) for hosting.

11. Security and Compliance

- End-to-end encryption (TLS 1.3).

- Role-based access control (RBAC).

- Data anonymization for training.

- Compliance: ISO 27001, GDPR, SOC2.

12. Monitoring and Maintenance

- Real-time monitoring: Prometheus + Grafana.

- Logging: ELK stack.

- Alerting: PagerDuty integration.

- Weekly model evaluation.

13. Timeline and Milestones

Milestone	Target Date
Requirement Finalization	June 10, 2025
MVP Development	July 15, 2025
Model Integration	August 5, 2025
Beta Testing	August 20, 2025
Production Launch	September 10, 2025

14. Appendix

- Glossary of Terms

- Sample Queries and Responses

- API Reference (link)

- Model Evaluation Metrics

End of Document