**Multimodal Conversational AI for E-commerce**

This report describes a multimodal conversational AI system designed to understand e-commerce contexts by enabling customers to search for products using both textual queries and visual inputs. The architecture uses CLIP embedding models, efficient vector retrieval mechanisms, and LLM integration.

The system’s architecture begins with a user input layer that accepts both text queries and image uploads through the Dify UI interface. These inputs are processed through a custom FastAPI server that serves as the central routing mechanism, directing requests to appropriate endpoints based on the input modality.

For embedding and vector store component, we employs the OpenAI CLIP-ViT-Base-Patch32 model for generating unified embeddings from both textual and visual inputs, allowing to project images and text into a shared embedding space and thus the comparison between a user’s text query and product images, or between an uploaded image and textual product descriptions. The generated embeddings are stored in a FAISS vector index, which provides highly optimized similarity search capabilities essential for real-time product retrieval. The system contains its product metadata separately in a JSON file for efficient vector operations.

The retrieval component uses FAISS’s efficient k-nearest neighbor search to identify the most relevant products based on embedding similarity. When a query is processed, FAISS searches through the indexed vectors and returns the top-K most similar products. These results are then complemented with corresponding metadata from the JSON storage, creating a comprehensive product information package. The retrieval system formats these results as structured JSON objects containing essential product details such as titles, descriptions, and image URLs, and we achieved 65.3% Recall@1; 71.4% Recall@3; and 73.5% Recall@5.

For LLM we chose Qwen3-8B model. It receives the structured JSON output from the retrieval system and applies templated prompts to generate coherent, contextually appropriate natural language of product matches. Also with CoT, it presents both the product information and reasoning for why certain products were selected, enhancing user understanding and trust in the recommendations.

图片包含 文本

AI 生成的内容可能不正确。We built application frontend directly in Dify, which supports both visual and textual product question-answering.