

# BUILDING A CHATBOT FOR TRAVEL AGENCIES

## OPTIMIZING ACCURACY AND CONTEXT-AWARENESS IN DOMAIN-SPECIFIC AI ASSISTANTS

AUTHOR: DAVID ŠVANCER

SUPERVISOR: ING. MARTIN KOSTELNÍK

### MOTIVATION

Standard LLMs and "naive" RAG systems often hallucinate critical travel details or lose track of complex conversations. In the travel industry, precision is non-negotiable: a wrong price or a missing amenity leads to immediate customer distrust.

This project implements an Advanced RAG pipeline to ensure strict factual grounding, seamlessly handle conversational context, and provide high-quality, reliable travel recommendations.

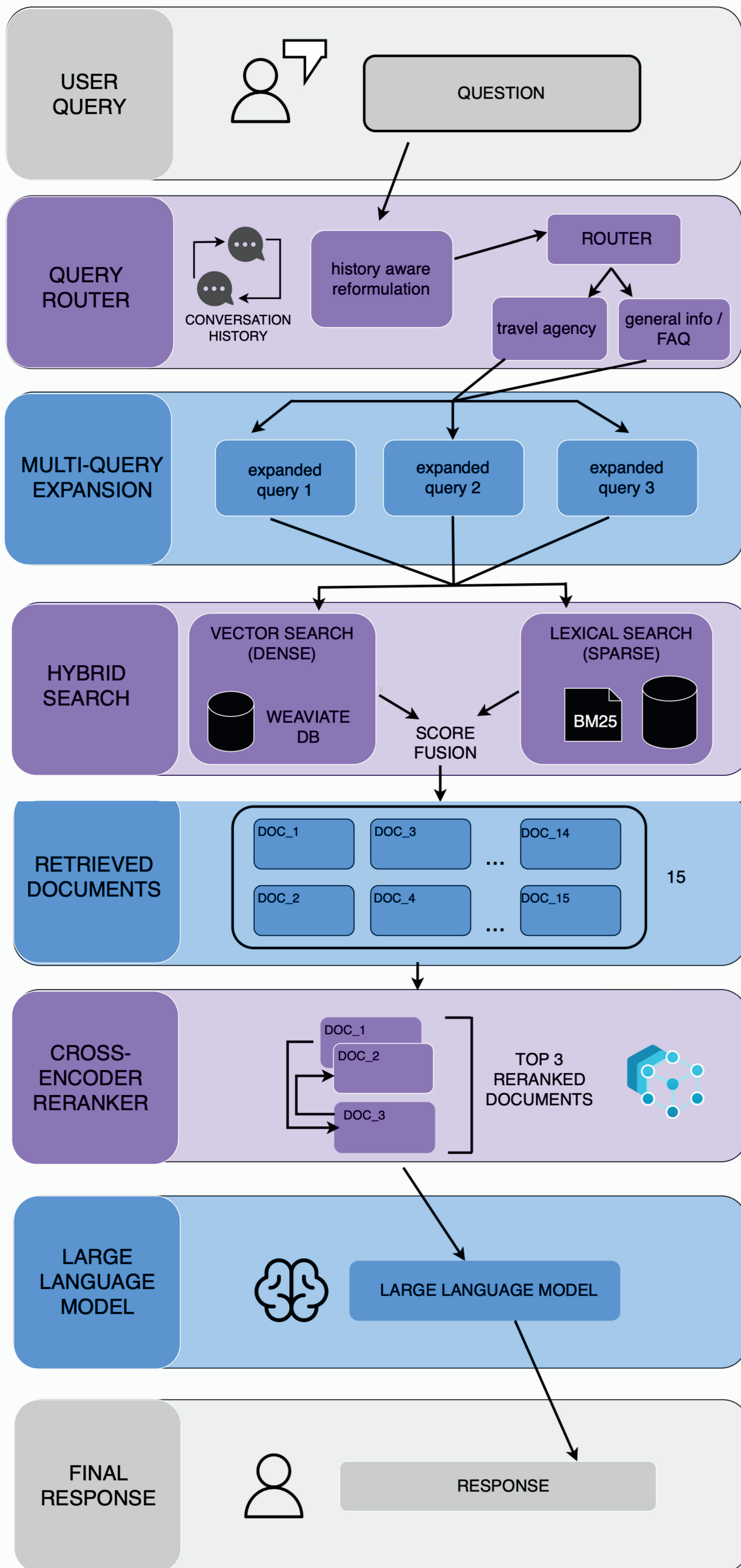


Figure 1: Advanced RAG system architecture and data flow

### CHAT EXAMPLE

**You:** I'm looking for a quiet apartment near Washington Park for a business trip. Short answer

**AI:** Based on the context, I would recommend the "Modern / Downtown Albany 1 BR APT" apartment. It's located near Washington Park and has a private backyard that's fully fenced, which should provide a quiet and peaceful environment for your business trip. The apartment also has a smart lock, carbon monoxide alarm, and exterior security cameras on property for added security. The price range for this apartment is not specified in the context, but it's a great option to consider for a quiet and comfortable stay near Washington Park.

**You:** That sounds good. Is there high-speed Wifi there for my video calls?

**AI:** According to the reviews, the "Modern / Downtown Albany 1 BR APT" apartment has WiFi, but the speed is not specified. One reviewer mentioned that the WiFi was "spotty", so it's possible that the connection might not be as reliable as you need for video calls. If high-speed WiFi is a must-have for your business trip, you might want to consider another option.

### EVALUATION

#### 1. Foundation: Choosing the Right Embedding Model

Comparison of Embedding Models on Baseline Architecture

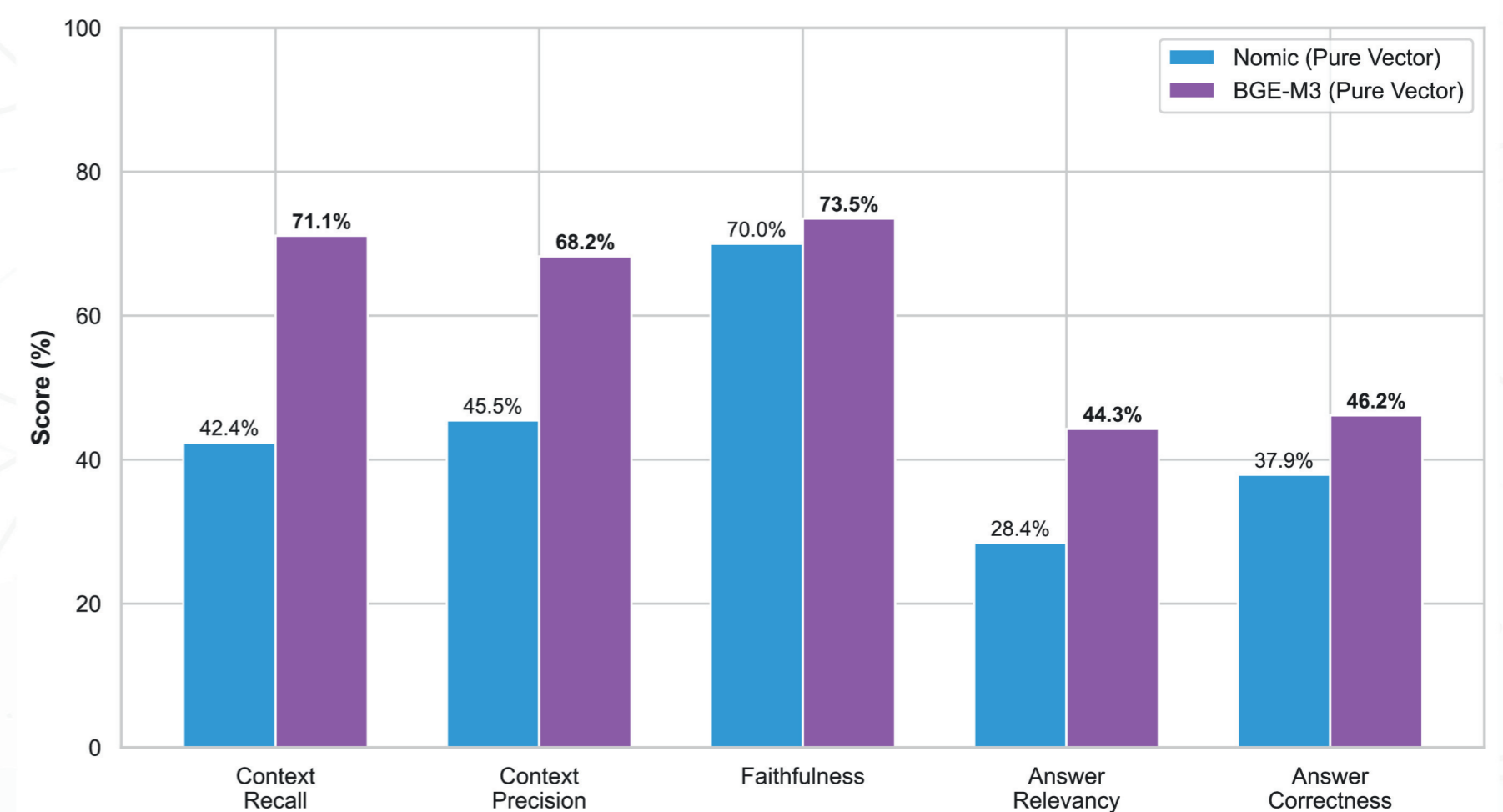


Figure 2: Comparison of embedding models

#### 2. Architecture Evolution & The "RAG Trade-off"

Detailed Evolution of Key Metrics Across Architectures

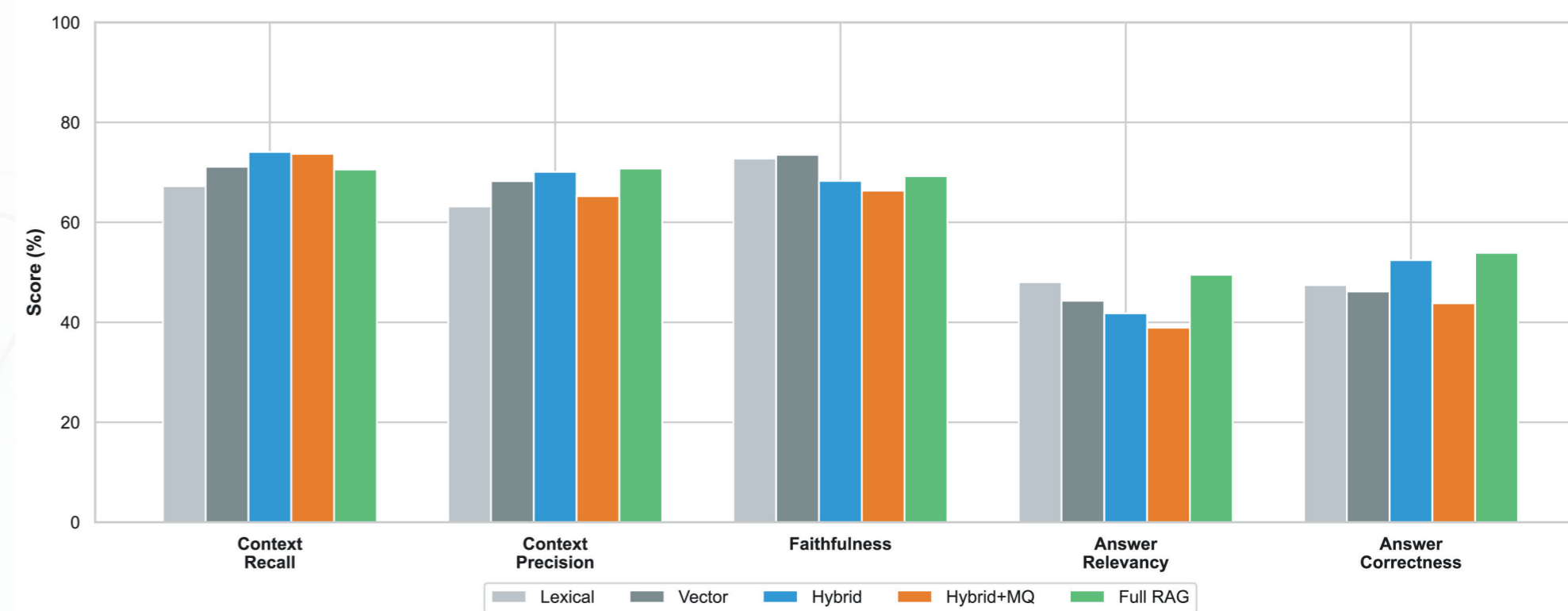


Figure 3: Comparison of LLM evaluations metrics

#### 3. Stress-Testing the Pipeline Across User Behaviors

Comprehensive Stress Test: Precision Analysis Across All Query Types

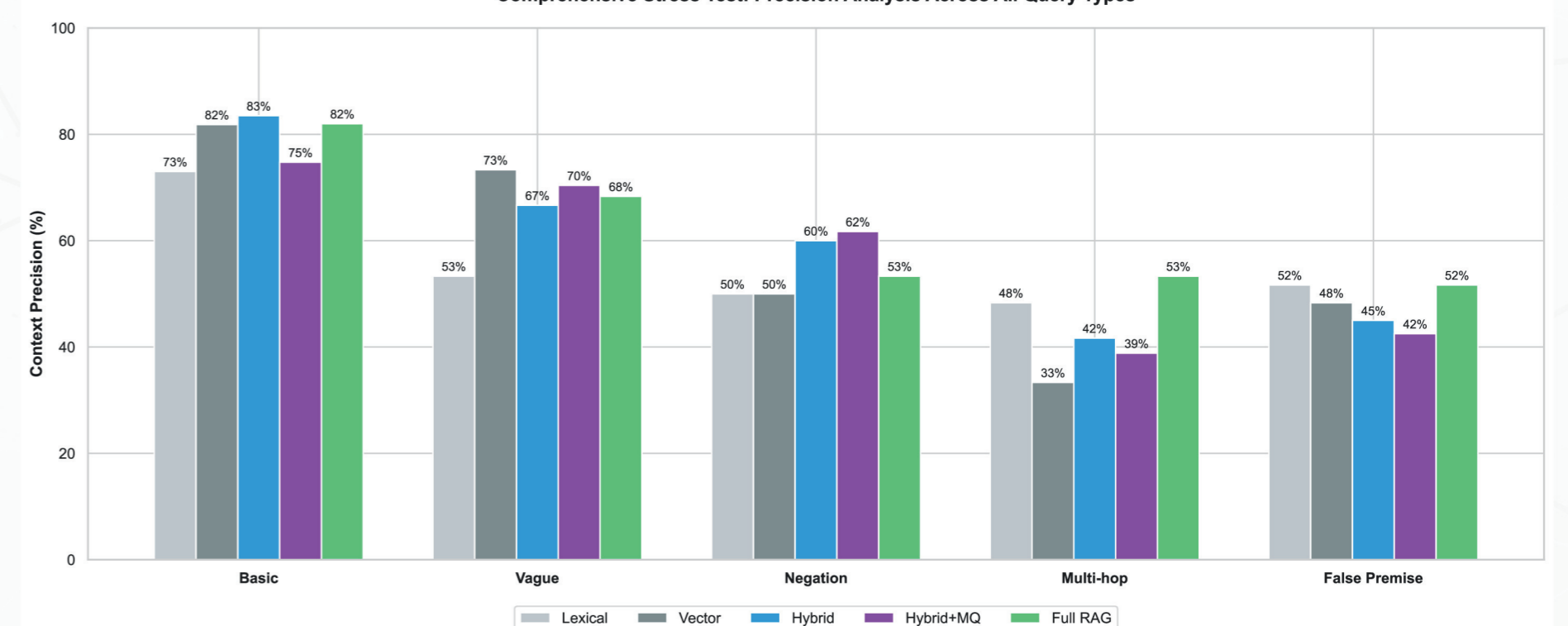


Figure 4: Comparison of LLM architectures results of trap messages