



SC Onboarding Agent: Reducing Customer Support Load with AI-Powered Assistance

Introduction

This project was created for the 'Practical Application of Artificial Intelligence' course at the Faculty of Electrical Engineering, University of Sarajevo. It addresses the real-world challenge of reducing support workload in a company that manages dozens of applications. The team aimed to automate onboarding and reduce ticket volume through an AI Onboarding assistant.

Project Goals

- Minimize incoming support tickets.
- Improve response accuracy and speed.
- Enhance the knowledge base and reduce repetitive queries.
- Support automatic/manual customer interaction modes.

Project Description

The onboarding agent integrates LangChain and LangGraph to deliver real-time document-based Q&A. Outdated manuals are removed and new ones vectorized using MistralAI embeddings, then stored in-memory. It was built for Shopify Partner apps but is adaptable to any context. The system is lightweight and includes a Python Flask-based frontend for interaction.

Frontend

SC Onboarding Agent

Enter your question:

Do I need to have a Google account?

Submit

Response:
Yes, you need to have a Google account before setting up a Google Merchant Center account.

Evaluation Strategy

LangSmith's LLM-as-a-judge framework is used for evaluation. A dataset of realistic Q&A examples is assessed using LLaMA3-70B. The model validates factual correctness and identifies errors, response latency, and formatting consistency. Results are visualized in LangSmith dashboards for transparency.

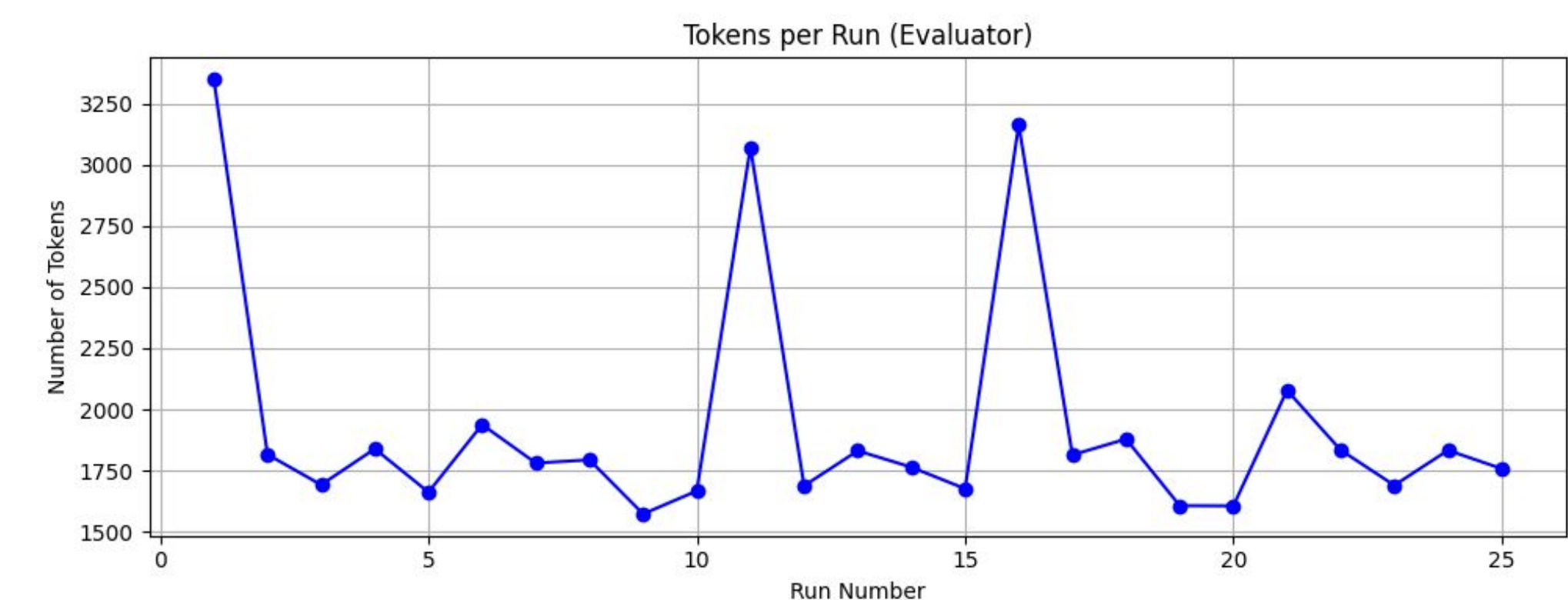
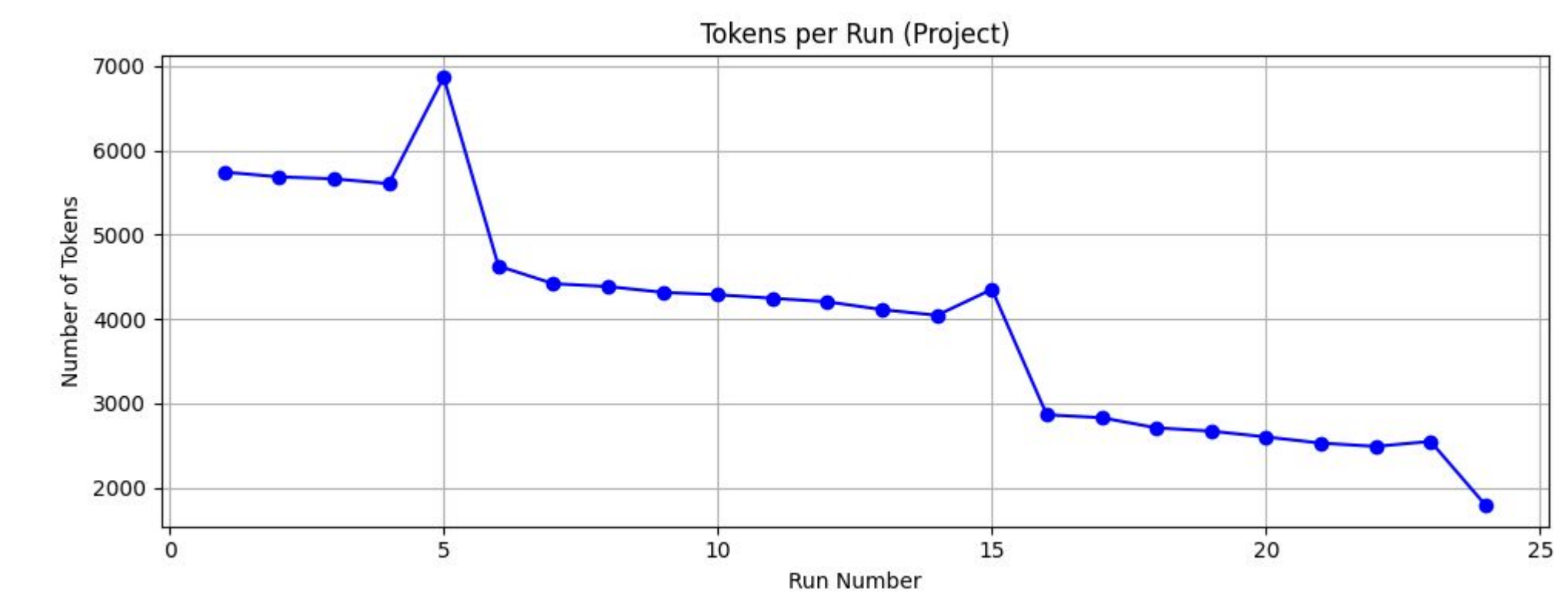
Results

The actual project runs processed significantly more tokens per run on average, which explains the longer response times compared to the evaluators. Despite the larger workload, the project runs maintained an excellent error rate over 24 runs, demonstrating stable performance. In contrast, the evaluator runs involved smaller token counts, leading to faster responses but a higher error rate of 8%, likely due to token limits and edge-case testing scenarios. Overall, the real project usage shows consistent reliability under heavier loads, while the evaluators highlight areas where errors may occur under constrained conditions. See [readme.md](#) file for details.

LangSmith Evaluation Dashboard

Personal > Tracing Projects > default									
default									
Runs Threads Alerts Setup									
1 filter Last 7 days Traces LLM Calls All Runs									
Stats									
Run Count 24									
Total Tokens 95,570									
Median Tokens 4,224									
Error Rate 0%									
% Streaming 0%									
Latency P50: 3.76s P99: 49.42s									
Filter Shortcuts									
Input									
Search input...									
messages.content									
messages.content == "are yo..."									
messages.content == "do I ne..."									
Name	Input	Output	Error	Start Time	Latency	Dataset	Anr		
LangGraph	user: Are you suppos...	ai: I apologize for ...		14/06/2025, 00:20:...	3.31s				
LangGraph	user: How can I setup...	ai: To set up sale...		14/06/2025, 00:19:47	1.84s				
LangGraph	user: Do I need to add...	ai: Yes, you need ...		14/06/2025, 00:18:14	5.68s				
LangGraph	user: Do I need to add...	ai: Yes, you need ...		14/06/2025, 00:18:04	0.72s				
LangGraph	user: How do I set the...	ai: To set up a Go...		14/06/2025, 00:17:16	0.84s				
LangGraph	user: Do I need to hav...	ai: Yes, you need ...		14/06/2025, 00:17:05	0.72s				
LangGraph	user: But you should ...	ai: You're right, I ...		14/06/2025, 00:15:39	0.84s				
LangGraph	user: What is Amazon...	ai: Amazon Simpl...		14/06/2025, 00:15:25	1.17s				
LangGraph	user: Fair point	ai: Google Merch...		14/06/2025, 00:15:17	0.68s				
LangGraph	user: What is Google ...	ai: I already told y...		14/06/2025, 00:13:46	0.55s				
LangGraph	user: What is Amazon...	ai: I don't know.		14/06/2025, 00:13:40	1.11s				

Token usage



References

- LangChain & LangGraph official docs
- Research on vectorized document retrieval
- LangSmith's framework for LLM evaluation
- Internal support documentation and user manuals

Conclusions & Implications

- Significant reduction in support workload
- Higher consistency in onboarding experience
- Scalable solution across applications and languages
- Sets groundwork for multilingual, memory-enhanced AI agents

Acknowledgements

- LangChain, LangGraph, LangSmith
- Groq, MistralAI, Hugging Face
- Flask, pandas, BeautifulSoup4, python-docx