Eli Lilly BoB Platform – Design Debrief

# 1. Executive Summary

The Eli Lilly BoB (Bot-of-Bots) Platform is a scalable, intelligent orchestration layer designed to route user queries to the most relevant AI agent based on topic and intent. Built using AWS Bedrock and Streamlit, it combines multi-agent architecture with real-time scoring and evaluation to streamline healthcare interactions for patients and internal users.

# 2. Problem Statement

With the proliferation of AI agents across different healthcare domains, there was a need for a centralized solution to orchestrate and optimize their usage. Key challenges included agent selection, response quality evaluation, and latency management.

# 3. Design Highlights

- Modular agent registry supports plug-and-play architecture.  
- Pre-routing logic for low-latency responses to known ticket-type queries.  
- Confidence and relevance scoring using Amazon Titan embeddings.  
- LLM-based judgment to select the most appropriate agent output.  
- Real-time feedback through Streamlit UI with performance metrics.

# 4. Key Architectural Decisions

- Chose AWS Bedrock to standardize access to LLMs and agents.  
- Used ThreadPoolExecutor for parallel agent invocation.  
- Cached embeddings to reduce duplicate inference costs.  
- Used Streamlit for rapid prototyping and UI delivery.

# 5. Business Impact

- Reduced response time for ticket-related queries by ~70%.  
- Improved user trust through transparency of evaluation metrics.  
- Enabled modular agent management without code duplication.  
- Created a scalable foundation for broader enterprise AI routing.

# 6. Handoff Notes

- Credentials should be moved to AWS Secrets Manager or environment variables.  
- New agents can be added by updating the AGENTS dictionary.  
- Extend the reference\_answers dictionary and re-embed when required.  
- Consider persisting session data to a DB for audit or insights.