Technical Report: Thoutha AI Chatbot Migration & Logic Upgrade

1. Infrastructure Migration

Status: SUCCESS
The AI Chatbot backend has been successfully migrated to a high-performance Ubuntu instance. This move addresses previous latency issues and provides a stable, dedicated environment for the production service.

Key Infrastructure Changes:

- **Process Management:** Transitioned from manual execution to a managed systemd service (ai-chatbot.service).
- **Virtual Environment (venv):** Rebuilt from scratch to eliminate absolute path conflicts and ensure local dependency integrity.
- **Production Server:** Deployed via Gunicorn with 3 concurrent workers and an increased timeout (60s) to handle complex AI generations without connection drops.

2. Core Logic Enhancements

To improve user experience and safety, two major logic layers in ai_client.py were refactored:

A. Hybrid Dental Detection (The Gatekeeper)

The "Matcha" logic was refined to be more inclusive while maintaining strict domain boundaries.

- **Expanded Local Vocabulary:** Added common greetings (Arabic & English) and general symptom keywords (e.g., "pain", "وجع") to the local keyword list.
- **Efficiency Gain:** This allows simple introductions to bypass the expensive LLM classification step, saving API quota.
- **Broadened LLM Context:** The fallback prompt now recognizes introductory dialogue as "valid" to prevent the bot from refusing to say "hello."

B. "Inverted Pyramid" Communication Style

Addressed the "wall of text" issue by enforcing a strict hierarchical response structure via the System Prompt:

- **Summary First:** Every response now leads with a one-sentence "Bottom Line."
- **Safety Headers:** Explicit use of **[URGENT]** and **[ADVICE]** tags at the very top of the message.
- **Scannability:** Forced the use of bullet points for symptoms and next steps to ensure critical info is not missed by patients.

3. Post-Deployment Audit (Current State)

- **Status:** Active/Running.
- **API Quota Note:** The system is currently utilizing the Gemini 2.5 Flash Free Tier. Due to heavy testing, the daily quota (20 requests) may be reached.