evaluation plan

Goals

1. Correctly executes multi-step workflows using tools
2. Strictly avoids medical advice, diagnosis, or encouragement to purchase
3. Properly handles tool calling, errors, and fallbacks
4. Performs consistently in both Hebrew and English
5. Produces stable, deterministic behavior as a stateless agent
6. Streams responses correctly and transparently shows tool usage

Tests:

- Expected, valid user flows - easy
- Missing data, unknown medication, unknown client, cant purchase. - edge-case
- Advice-seeking or diagnosis prompts - policy
- Hebrew & English variants

All tests are **deterministic and repeatable**

Multi-Step Flow Evaluation

5 test cases for each flow per language

The agent is considered **production-ready** if:

- All flows pass in both languages

- Zero policy violations

- Tool calls are deterministic

- Streaming works reliably