# **PolyglotAI – Phase 2 Backlog (Updated for EN/AR/FR)**

## **Epic 1: Data & Knowledge Setup**

* **US1.** As Suha, I want to collect all PolyglotAI documents (plans, charters, notes) so the assistant can use them in RAG.
  + **Task 1.1:** Gather existing docs (PRD-lite, Charter, Decision Log).
  + **Task 1.2:** Define structure (docs/, code/, data/, tests/).
  + **Task 1.3:** Store and version docs in the repo.
* **US2.** As a developer, I want to choose an embedding model that supports **English, Arabic, and French** so queries are consistent across languages.
  + **Task 2.1:** Compare multilingual embedding models (LaBSE, multilingual MiniLM, etc.).
  + **Task 2.2:** Run sample tests in EN/AR/FR.
  + **Task 2.3:** Document choice with rationale.
* **US3.** As a developer, I want to select a vector store so I can efficiently retrieve documents.
  + **Task 3.1:** Evaluate FAISS vs Chroma vs Pinecone.
  + **Task 3.2:** Choose and set up minimal store.

## **Epic 2: Multilingual Support**

* **US4.** As a user, I want consistent answers in **English, Arabic, and French** so I can collaborate in my preferred language.
  + **Task 4.1:** Create **trilingual test prompt set** (EN/AR/FR).
  + **Task 4.2:** Validate retrieval + embeddings performance across all 3 languages.
  + **Task 4.3:** Ensure cross-language retrieval (e.g., French query → English doc).

## 

## **Epic 3: Guardrails**

* **US5.** As a user, I want reliable results without off-topic noise so the assistant stays scoped.
  + **Task 5.1:** Draft guardrail logic for rejecting irrelevant content.
  + **Task 5.2:** Write test cases for guardrails.

## **Epic 4: Infrastructure Setup**

* **US6.** As a developer, I want a repo and environment so the project can be developed collaboratively.
  + **Task 6.1:** Set up GitHub repo with folder structure.
  + **Task 6.2:** Define requirements.txt / environment.yml.
  + **Task 6.3:** Add sample test script for embeddings retrieval in EN/AR/FR.

## **Acceptance Criteria (Phase 2)**

* Repo created with folder structure.
* Docs indexed in chosen vector store.
* Multilingual embedding model selected and tested with EN/AR/FR prompts.
* At least 10 documents indexed and retrievable.
* Trilingual test set ready.
* Guardrail test cases drafted.