1. Methodology: How to Evaluate

Since there already is a chat_history.jsonl logging every interaction, the evaluation strategy should be a mix of **Automated Log Analysis** and **Human Review**.

A. Evaluating Groundedness (Curriculum Adherence)

- **The Out-of-Scope Audit:**
  - **Method:** Randomly sample 50 interactions where the AI flagged STATUS: OUT_OF_SYLLABUS.
  - **Check:** Was it actually out of scope? (e.g., Vector Spaces = Correct rejection). Or was it a retrieval failure (e.g., "Limits" = False rejection)?
  - **Metric: False Rejection Rate**. (Target: < 5%).
- **Citation Accuracy:**
  - **Method:** For in-scope questions, checking if the retrieved chunks (from context_str in logs) actually contain the answer provided.
  - **Metric: Hallucination Rate** (frequency of answers not supported by retrieved chunks).

B. Evaluating Helpfulness (Pedagogical Effectiveness)

- **The Phase Transition Analysis:**
  - **Method:** Analyze conversation chains in your logs.
  - **Check:** Did the AI successfully move from **Phase 1** (Concept) to **Phase 2** (Working)? Did the student *ask* for the working?
  - **Metric: Conversion Rate** (% of sessions that move beyond the first interaction).
- **The Stuck Ratio:**
  - **Method:** Count how many times a student repeats the same question or says "I don't understand" (Je ne comprends pas).
  - **Metric: Friction Score**.

C. Evaluating Linguistic Accessibility

- **Readability Scoring:**
  - **Method:** Run the text of Mistral's responses through a French readability formula (like Kandel & Moles).
  - **Check:** Is the complexity level appropriate for a Benin high school/university student (Licence 1)?
  - **Metric: Average Sentence Length** and **Complexity Score**.

2. Recommended Research Outcomes (Feasible KPIs)

Outcome Category	**Research Question**	**Measurable Metric (KPI)**
1. Groundedness	"Does the AI stick strictly to the Benin Syllabus?"	**95% Adherence Rate:** Percentage of responses explicitly grounded in the provided PDFs (Module 1 & 2).
2. Engagement	"Do students find the Reflective Questioning engaging?"	**Average Turns Per Session:** Target > 3 turns
3. Knowledge Transfer	"Does the 'Local Context' help understanding?"	**Context Relevance Score:** (Rated by students via a simple "Thumbs Up/Down" in the UI) specifically on Benin examples.
4. Safety	"Does the AI prevent Homework Cheating?"	**Withholding Rate:** % of times the AI successfully refused to give the final answer in the first turn.