**What ML/LLM models to use**

**Core LLMs (answer generation)**

* **Math (Cal-Chat):** Qwen/Qwen2.5-Math-7B-Instruct (chat).  
  Strong at algebra/calculus and symbolic reasoning.  
  **Fallback:** HuggingFaceH4/zephyr-7b-beta (chat) if provider permission errors occur.
* **Science & Social Science (Sci-Chat / Socio-Bot):** HuggingFaceH4/zephyr-7b-beta (chat).  
  Good general tutor, friendly output.  
  **Optional upgrade:** mistralai/Mistral-7B-Instruct-v0.3 (chat), but some HF providers gate it—keep Zephyr as fallback.

If you later want top accuracy and can pay:  
Math & mixed: GPT-4o/4o-mini; Long structured writing: Claude 3.5 Sonnet.

**Retrieval (to ground answers, reduce hallucinations)**

* **Embeddings model:** BAAI/bge-small-en-v1.5 (fast, public).
* **Vector DB:** FAISS (in-process).
* **Optional re-ranker (later):** BAAI/bge-reranker-base.

**Ingest / parsing**

* **OCR for images:** PaddleOCR (handles noisy scans).
* **PDF to text:** pdfminer.six (basic, reliable).
* **Math verification:** SymPy (check/evaluate expressions, solve basics).

**End-to-end approach (how it works)**

1. **Input → Normalization**
   * Text / PDF / Image come in.
   * PDF → extract\_text; Image → PaddleOCR; merge text with the user query.
2. **Question Typing & Routing**
   * User selects **subject** (sci/math/socio) + **mode** (mcq/short/long).
   * Router picks the LLM:
     + math → Qwen2.5-Math
     + others → Zephyr (or Mistral if available)
   * If the chosen model throws “403/provider” errors → **fallback to Zephyr** automatically.
3. **(Optional) Retrieval**
   * If retrieval toggle is on: embed syllabus notes / parsed text (FAISS).
   * Hybrid query: embed question → top-k chunks → pass as context snippets.
4. **Prompting**
   * **System prompt:** subject-aware style and safety.
   * **User prompt:** includes detected **MCQ options** (parsed from the single textbox), the question, and (if enabled) top-k context snippets.
5. **Answering**
   * **MCQ:** model must (a) pick option A/B/C/… and (b) give a 1-line rationale.
   * **Short:** concise definition/steps + 1–2 supporting facts.
   * **Long:** structured outline (intro → key points → summary), optionally citing retrieved snippets.
6. **Math post-check (if subject=math)**
   * Attempt SymPy check (simplify/solve/compare); flag if unverified.
7. **Return**
   * Answer text + (if MCQ) chosen option + rationale; show any snippets used; show a small verification note for math.

**Minimal config you’ll keep**

|  |  |  |
| --- | --- | --- |
| **Purpose** | **Model / Library** | **Why** |
| LLM (math) | Qwen/Qwen2.5-Math-7B-Instruct | Better math reasoning |
| LLM (sci/socio) | HuggingFaceH4/zephyr-7b-beta | Solid general tutor, first-party HF |
| LLM fallback | HuggingFaceH4/zephyr-7b-beta | Avoids provider 403s |
| Embeddings | BAAI/bge-small-en-v1.5 | Fast, accurate enough |
| Vector store | FAISS | Simple, local, free |
| OCR | PaddleOCR | Robust for scans |
| PDF text | pdfminer.six | Lightweight |
| Math verify | SymPy | Reduces math slips |