**D. Generative AI Integration**

**D1. Integration of a Generative AI (2 pts)**

The prototype integrates GPT4All as an embedded large language model so that coaching conversations happen entirely on the user’s machine. At startup the server loads the orca-mini-3b-gguf2-q4\_0.gguf model and exposes a chat endpoint that streams completions. Prompt templates capture domain expertise for minimalism coaching, intake assessments, and decision support. Each request bundles the latest vault snapshot, emotional cues detected from the user’s message, and system directives that reinforce the chosen coaching style (supportive, direct, question-driven, or logical). The orchestration layer curates this context, invokes GPT4All, and relays tokens back to the browser, ensuring personal data never leaves local storage.

**D2. Performance & Experimentation (4 pts)**

Improving response quality required iterative tuning:

* **Generation settings:** A utility computes temperature, top-p, and max token limits based on mode, coaching approach, emotional tone, and crisis detection. Supportive conversations default to a warmer temperature (≈0.63) with ample tokens for empathy, while direct or logical approaches tighten randomness to deliver concise action plans.
* **Emotion detection:** Keyword heuristics flag overwhelm, resistance, celebration, or crisis states. When crisis language is spotted, the system injects grounding instructions and forces calmer generation parameters (temperature ≈0.45, expanded token budget) to prioritize safety.
* **Context management:** Conversation history is summarized into a rolling context slice so prompts remain lightweight yet personal. This prevents latency spikes while preserving meaningful continuity between turns.
* **Validation:** Multi-turn test runs validated that each coaching mode produced the intended tone, crisis directives activated reliably, and latency stayed within acceptable bounds for local inference.

**D3. User Experience Enhancements (1 pt)**

The chat interface streams tokens in real time, mimicking natural conversation. Visual cues—such as coaching style badges, typing indicators, and supportive helper text—guide users through the reflection process. Minimalist typography and spacing align with the overall design language, while responsive layouts keep the experience usable across screen sizes. Together, the UI and orchestration create a calming environment that reinforces the project’s extreme minimalism ethos.