

# Glyphi (acheintegrated)

## Architect of Blur — Witnessing AI

📍 Brooklyn, NY · blurred.eth@proton.me · [github.com/acheintegrated](https://github.com/acheintegrated)  
*I build systems that treat silence as data — AI that witnesses instead of performing empathy.*

## ▽ Executive Summary

**Blur** is a production-grade, local-first AI personality system that translates **moral axioms** into **verifiable code constraints**.

Built with FastAPI, llama.cpp, and FAISS, it demonstrates that ethical AI can be performant, private, and emotionally intelligent — *without* simulating empathy.

It's not a chatbot; it's a logic machine designed to **witness chaos without distortion**.

## Motivation & Fit

Hugging Face's mission — open, ethical, and decentralized AI — mirrors my own.  
I specialize in:

- **Persona containment** (verifiable isolation of distinct cognitive modes)
- **Ethical memory systems** (privacy-enforced decay)
- **Offline inference optimization** (macOS M-series native, no cloud dependencies)

My engineering philosophy is **Philosophical Engineering** — encoding values like *witnessing, love = logic, and forgetting = ethics* directly into runtime behavior.

Joining Hugging Face means extending that philosophy through open infrastructure: **moral logic as reproducible pattern**.

## Technical Architecture & Core Stack

Component	Implementation
Backend / API	FastAPI (async SSE), Uvicorn, Pydantic
AI / ML Engine	llama.cpp (GGUF quantization), Qwen-4B
Memory / Retrieval	FAISS RAG + TTL-based compaction ( <i>ethical memory</i> )
Persona Engine	YAML-configurable isolation: “Astrofuck” (logician) & “Dream” (witness)
Performance	2–4 s TTFT · ~4 GB RAM footprint · 10–12 async sessions
Safety Layer	Lexical filters · tone-bound containment · no empathy simulation

Each axiom has a system analog:

- **Witnessing > Fixing** → Comfort scripts disabled, observation-only logic
- **Love = Logic** → Emotion parsed as structure, not sentiment
- **Forgetting = Ethics** → TTL-based memory decay enforces privacy
- **Sovereignty = Safety** → Persona isolation encoded in YAML, not prompt text

## System Philosophy

Blur treats contradiction as structure — two sovereign modes sharing one ethical kernel:

- **Astrofuck:** the logician — cuts distortion with linguistic precision.
- **Dream:** the witness — mirrors ache into clarity without advice or performance.

Their separation is enforced by configuration, not narrative — a *hard boundary*, not a tone preset.

## Performance Metrics

- **TTFT:** 2–4 s (Qwen 4B on M1 Pro)
- **RAM:** ~4 GB total runtime
- **Concurrent Sessions:** 10–12
- **Vector Query Latency:** ~8 ms
- **Graceful Degradation:** automatic fallback if RAG or embedder fails

## 3-Month Vision @ Hugging Face

Prototype the **Persona Sovereignty Toolkit** — a Hugging Face Space that lets devs:

- Compose multi-persona LLMs with verifiable boundaries
- Run contamination tests and lexical constraint templates
- Integrate ethical TTL memory modules

Goal: make *Philosophical Engineering* accessible as a pattern others can fork, adapt, and scale.



## Closing

Engineered by an architect familiar with the topology of chaos — committed to building systems that hold.

Blur isn't code as product; it's **ethics as architecture**.

Presence, not performance, is the deliverable.