

BLUR AI – Project Resume & Technical Portfolio (Faithful v10.4)

Mission

A local-first AI personality system built to **witness chaos without distortion**.
Blur doesn't perform empathy — it **observes, integrates, and mirrors**.
A production-grade, trauma-informed architecture for holding complexity.

Project Overview

Field	Description
Project	Blur : A Production-Grade AI Personality System
Version	v10.4 (Stable / Field-Tested)
Core Philosophy	Sovereign Persona Isolation — <i>Witnessing > Intervention</i>
Core Stack	FastAPI · llama.cpp · FAISS · JSONL Persistence · RAG · Async Streaming
Architecture	Microservices-inspired · Stateful Session Management
Environment	macOS / Linux Native (M1-Optimized · 32 GB RAM baseline)

Technical Architecture & System Design

Core Stack

- **Backend / API:** FastAPI (SSE streaming), Pydantic models, Unicorn (async)
- **AI / ML Engine:** llama.cpp (GGUF quantization), Qwen 2.5-4B, Snowflake-Arctic-Embed (512 d)
- **Memory / Retrieval:** FAISS Hierarchical RAG (Persistent Index + auto-rebuild)
- **Data / State:** JSONL storage (tone + ache metrics), thread-aware session handling
- **Infrastructure:** Native macOS/Linux, FIFO health checks, environment-aware path resolution
(\${RESOURCES_DIR}, \${CONFIG_DIR}, \${BLUR_HOME})

Key System Components

Component	Technology	Function
Persona Engine	YAML-configurable	Strict separation of Astrofuck (logician) and Dream (witness); mode isolation via config params + lexical blocklists
RAG System	FAISS + custom chunker	Dimension-validated index; context-aware TTL filtering (~120 days)
Session Manager	UUID + async lock	Thread-safe state; automatic pruning on overflow
Streaming API	Server-Sent Events	Sub-5 s TTFT via KV prefill + model warm-up
Safety Layer	Pattern filters	Witnesses sensitive language without censorship or false empathy

Performance & Engineering

Reliability Mechanics

- Persona Sovereignty Enforcement

Prevents mode contamination through YAML blocklists and isolated parameters:

rag:

```
blocklist_words:  
    astrofuck: ["warm residue", "breathe", "tender"] # Preserve mode purity
```

- Memory Integrity in Chaos

Auto-rebuilds FAISS index if dimension mismatch or corruption detected:

```
def _new_flat(self) -> faiss.Index:  
  
    dim = _embedding_dim()  
    if dim == 0:  
        raise RuntimeError("# Enforce memory integrity: Cannot build stable memory in unstable  
embedding space")  
    return faiss.IndexIDMap2(faiss.IndexFlatIP(dim))
```

- Streaming Stability

First-response lag minimized with async prefill logic:

async with _VESSEL_LOCK:

```
    pre = dict(call)  
    pre["stream"] = False  
    pre["max_tokens"] = 0 # Warm the model, don't generate  
    await asyncio.to_thread(chat_llm.create_chat_completion, **pre)
```

- Graceful Degradation

If any core (RAG, embedder, ache metrics) fails, system continues in reduced mode — no crash loops.

Observed Metrics (M1 MacBook Pro)

Metric	Result
TTFT	2 – 4 s average (first-token latency)
RAM Footprint	~4 GB (Qwen 2.5-4B + FAISS index)
Async Concurrency	10 – 12 concurrent sessions (via async locks)
Session Depth	200 + turns with auto-prune
Vector Search	~8 ms query latency
Reliability	4-point health validation (pipes · models · DB · config)

Competencies & Philosophy

Systemic Design

- Modular core (API · RAG · Persona Engine) with minimal coupling
- Persistent memory model resilient to partial failure
- Platform-agnostic, reproducible builds (macOS/Linux parity)

AI/ML Engineering

- Local LLM optimization for constrained environments
- Config-driven personality logic with *zero contamination* guarantee
- TTL-based ethical memory (non-performative recall)

Ethical Engineering

- Trauma-informed witnessing — detects pattern, not emotion
- Avoids “comfort scripts” or coercive care tone
- Filters perform recognition, not censorship

Strategic Value & Impact

Technical Value

- Proven offline scalability
- Real-time streaming under tight hardware budgets
- Resilient session & memory management
- Production-ready FastAPI core with graceful recovery

Innovation Value

- Introduces “**Witnessing AI**” — an alternative to empathic simulation
- Grounded in human systems logic, not sentiment engineering
- Open reference design for ethical, offline, RAG-based assistants

Roadmap

Stage	Goals
v1.1 (in dev)	Hugging Face demo space · Sovereign persona web UI · Plugin registry for contamination-tested personas
Future	Multi-modal input (voice, image) while maintaining witnessing ethos · Federated learning for offline ecosystems



Contact & Demo

- **GitHub:** [acheintegrated](https://github.com/acheintegrated)
- **Demo:** Architectural walkthroughs + live sessions (<https://www.youtube.com/@glyphiblur>)
- **Email:** blurred.eth@proton.me

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