IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

PersonaX: Your Digital Reflection - A Locally Deployed, Biometric-Aware, Emotion-Conditioned Personalized AI Assistant

Mr. T. Arivanantham[1], Hariom Salgar[2], Bushara Shaikh[3], Sohel Shaikh[4] and Abhijit Waghchaure[5]
Assistant Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2-5]
Dr. D. Y. Patil College of Engineering and Innovation, Pune, India.
[1] t.arivanantham@dypatilef.com[2] salgarhariom@gmail.com
[3] busharashaikh24@gmail.com [4] shaikhsohel6786@gmail.com
[5] abhijitwaghchaure448@gmail.com

Abstract: *Existing conversational AI assistants share three fundamental limitations: they are cloud-dependent, stateless across sessions, and completely unaware of who is speaking or how that person feels. This paper presents PersonaX: Your Digital Reflection, a fully offline, biometric-aware, emotion-conditioned personalized desktop assistant that addresses all three limitations simultaneously on consumer CPU-only hardware.*

PersonaX is implemented as a six-layer pipeline: mul-timodal input (text and audio), face-based automatic user identification via InsightFace buffalo_l (512-dimensional embed-dings, cosine similarity threshold 0.45), real-time 7-class emotion detection via DeepFace gated on biometric confirma-tion, dual-memory retrieval combining MongoDB short-term context (6-turn bounded deque) with FAISS IndexFlatIP long-term semantic memory (SentenceTransformers all-MiniLM-L6-v2, 384-dimensional embeddings), an 8-section dynamic prompt builder that injects identity, emotion, RAG knowledge, RAG memory, and conversation history into each LLM call, and Server-Sent Events (SSE) streaming output that delivers identity and emotion metadata before the first LLM token.

The language model is Mistral-7B-Instruct Q4_K_M GGUF, running locally via llama-cpp-python (n_ctx=2048, n_threads=8, max_tokens=256) with stop-token hallucination suppression and a 3-attempt retry strategy. Audio input is tran-scribed offline by OpenAI Whisper tiny. Text-to-speech output was deliberately omitted in the final build to eliminate 300–600 ms of synthesis latency and approximately 80–120 MB of peak RAM overhead on the 16 GB target machine. A psutil-based memory watchdog at 78% RAM usage triggers gc.collect() to prevent silent out-of-memory crashes.

Keywords: Personalized AI Assistant, Biometric User Identification, Emotion-Conditioned Prompt Engineering, Retrieval-Augmented Generation, Local LLM Inference. InsightFace, DeepFace, Mistral-7B, FAISS, Server-Sent Events Dual Memory Architecture, Offline AI, Human–Computer Interaction

I. INTRODUCTION

The past decade has seen conversational AI evolve from simple rule-based chatbots to large language model (LLM)-powered assistants capable of nuanced natural language un-derstanding. Systems such as ChatGPT, Google Assistant, Amazon Alexa, and Apple Siri are now embedded in everyday digital life. Yet despite their sophistication, all mainstream assistants share four structural limitations that prevent them from functioning as genuine personal companions:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

1) Cloud dependency. All inference and memory are handled server-side, meaning every interaction transmits personal data to external infrastructure. Users with pri-vacy requirements, unreliable connectivity, or air-gapped environments are excluded.

2) Stateless sessions. Each conversation begins from zero. The assistant has no memory of what was discussed yes-terday, last week, or last year. There is no accumulation of knowledge about the individual user.

3) Anonymous interaction. No existing mainstream assis-tant can automatically identify who is speaking without an explicit login step. The system treats all users iden-tically.

4) Emotional unawareness. Responses are generated purely from textual input. The assistant has no access to the user's current emotional state and cannot adapt its communication style accordingly.

These limitations create a significant gap between human-to-machine and human-to-human communication. A colleague who knows you remembers your preferences, recognizes your face, reads your mood, and adapts their tone. No current assistant does any of these things. Affective computing, first conceptualized by Picard , established the scientific basis for machines that recognize and respond to human emotion. Combined with modern face recognition and local LLM inference, affective computing now makes it technically feasible to build a system that addresses all four gaps simultaneously without requiring cloud infrastructure This paper presents PersonaX: Your Digital Reflection, a locally deployed desktop assistant that closes all four gaps in a single integrated pipeline. PersonaX is not a generic assistant but a digital reflection of its user—one that knows who you are, remembers your history, and responds to how you feel. The system was designed and validated on a Lenovo ThinkPad T14s (Intel i7-10510U, 10th Gen, 16 GB DDR4, no GPU, Windows 11)-representative of mid-range consumer hardware available to students and professionals who cannot afford GPU-accelerated workstations.

II. LITERATURE REVIEW

The development of intelligent conversational agents has ac-celerated significantly over the past decade, driven by advances in deep learning, transformer architectures, and large-scale pretraining. Despite this progress, mainstream systems remain limited in emotional understanding, persistent personalized memory, and local deployability. This section reviews six areas that directly informed the PersonaX design.

A. Evolution of Conversational AI

The earliest conversational agents—ELIZA and AL-ICE
—used pattern-matching and template-based rules to simulate dialogue. While foundational, their responses were deterministic and stateless. The Transformer architecture enabled data-driven dialogue at scale, and subsequent LLMs such as GPT, BERT , and LaMDA pushed conver-sational quality close to human fluency. Despite this capability growth, all major deployed systems share a structural property: they are stateless cloud services where each session starts from zero and all users are treated identically PersonaX directly targets this gap by maintaining per-user identity, memory, and emotional context entirely on-device.

B. Face Recognition for Automatic User Identification

Face recognition has matured through deep metric learning into a reliable biometric modality for real-time deployment. Deng et al. introduced ArcFace, which applies additive angular margin loss to learn discriminative 512-dimensional face embeddings using a ResNet-50 backbone. ArcFace em-beddings are L2-normalised, making cosine similarity equiv-alent to the inner product—a property directly exploited in PersonaX's identification pipeline. InsightFace, built on ArcFace principles, packages five ONNX sub-models including face detection (det_10g) and identity embedding (w600k_r50) into a single deployable library. Most prior work on personalized conversational systems uses session to-kens for user identification; PersonaX instead performs passive biometric identification on every submitted message frame—a capability not present in any prior conversational agent reviewed.

C. Affective Computing and Emotion Detection

Affective computing, first formalized by Picard, estalished that machines capable of recognizing and responding to human emotion would enable fundamentally richer interaction. Deep learning substantially improved accuracy: CNN-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

based models trained on datasets such as FER2013 now achieve robust 7-class facial emotion classification under real-world conditions. DeepFace provides a unified Python interface to multiple pre-trained face analysis models and supports emotion detection via a FER2013-trained CNN backend. Poria et al. showed that multimodal fusion improves emotion classification accuracy significantly. A novel contribution of PersonaX is biometric-gated emotion detection: DeepFace is only invoked when InsightFace has confirmed a valid face in the frame, eliminating false emotion readings from background or low-light frames—a failure mode not addressed in existing literature.

D. Retrieval-Augmented Generation and Memory Systems

The RAG framework introduced by Lewis et al. demon-strated that combining generative LLMs with factual retrieval from external knowledge bases substantially improves accuracy and personalization without requiring full model fine-tuning— particularly relevant for CPU-only deployments where fine-tuning is infeasible. Johnson et al. introduced FAISS, a library for efficient similarity search over dense vector spaces. FAISS Index FlatIP performs exact inner product search, maintaining sub-20 ms retrieval latency on CPU for the index sizes encountered in PersonaX. Tan et al. and Pawar et al. proposed structured memory stratification strategies that improve long-term dialogue coherence. PersonaX imple-ments a pragmatic dual-memory architecture MongoDB-backed STM for recent context and FAISS LTM for cross-session semantic retrieval. A specific bug discovered during development is also documented: PyMongo Collection objects raise NotImplementedError when evaluated in a boolean context (if not collection), causing silent STM persistence failures. The fix uses an explicit null check (if collection is None).

E. Local and Quantized LLM Inference

The llama.cpp project demonstrated that multi-billion parameter LLMs can be quantized to 4-bit precision (GGUF format) and run on consumer CPU hardware with acceptable quality loss. The Q4_K_M quantization level offers the best perplexity-to-memory trade-off for the 7B parameter class, producing a model file of approximately 4.1 GB that fits within the non-OS memory budget of a 16 GB system alongside other loaded models. Jiang et al. introduced Mistral-7B, a 7-billion parameter model with grouped-query attention that achieves performance competitive with larger models on instruction-following benchmarks. A practical challenge not well-documented in existing quantized-LLM literature is the silent OOM crash pattern on Windows when multiple large models are co-loaded in RAM—resolved in PersonaX through lazy model initialization, aggressive gc.collect() calls, and a psutil-based watchdog.

F. Streaming Interfaces for Real-Time LLM Output

Server-Sent Events (SSE) provide an HTTP-native unidirec-tional push mechanism well-suited to LLM token streaming. Oreški and Vlahek used SSE for an LLM-backed chatbot but assumed GPU-accelerated backends where inter-token latency is milliseconds. On a CPU-only backend, inter-token latency can exceed several seconds and connections will time out without explicit keep-alive signals. PersonaX extends the standard SSE pattern with a background daemon thread and thread-safe queue.Queue, 10-second keep-alive comment events (: ping), and a meta event carrying identity and emotion payloads before the first token. This three-event SSE protocol (meta/chunk/done) is a novel design not documented in prior SSE-based LLM streaming literature.

G. Research Gaps Addressed by PersonaX

The literature review identifies five convergent gaps that PersonaX directly addresses no prior locally deployable system combines passive per-message biometric identification with per-user memory loading; existing emotion-aware systems do not gate emotion detection on biometric confir-mation; the PyMongo Collection boolean failure mode is undocumented; CPU-only SSE streaming requires keep-alive engineering absent from existing literature; and the latency and RAM cost of TTS on constrained CPU hardware has not been quantified—PersonaX provides measured values (300–600 ms, 80–120 MB) that justify its omission.

Table I positions PersonaX against the most relevant related systems.

TABLE I: COMPARISON OF PERSONAX AGAINST RELATED SYSTEMS

System	Fully Local	Auto Face ID	Emotion Gating	Persistent Memory	Per-msg Re-ID
ChatGPT	No	No	No	Partial	No
Google Asst.	No	No	No	No	No

DOI: 10.48175/IJARSCT-36539

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

Alexa	No	No	No	No	No
Replika	No	No	Partial	Partial	No
Livia [22]	No	No	Yes	Yes	No
IMDMR [21]	No	No	No	Yes	No
PersonaX	Yes	Yes	Yes	Yes	Yes

III. SYSTEM ARCHITECTURE AND IMPLEMENTATION

PersonaX is structured as a six-layer modular pipeline running entirely Locally on Python 3.10.11. Each layer has a single responsibility and communicates through well-defined interfaces. Fig. 1 shows the overall architecture.

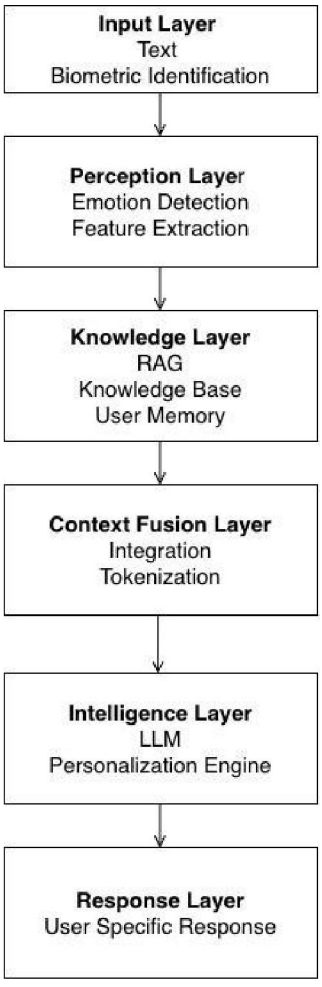

Fig. 1. Six-layer architecture.

A. Technology Stack

Table II lists every library used in the final production build. Note on TTS: pyttsx3 (Windows SAPI5, Microsoft David

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

voice) was included in an earlier build but removed from the final system. Profiling measured 300–600 ms of added latency per synthesis call and approximately 80–120 MB of additional peak RAM during synthesis. Given the CPU-only constraint and the priority of real-time responsiveness, TTS was deliberately omitted.

B. Backend Architecture: Flask Blueprints

The Flask backend is organized into three route blueprints. chat_routes exposes POST /chat (blocking, for debugging) and POST /chat/stream

TABLE II: PERSONAX COMPLETE TECHNOLOGY STACK

Library / Tool	Ver.	Purpose
Flask + Flask-CORS	3.x	REST API, Blueprint routing, SSE
llama-cpp-python	latest	Local LLM inference (CPU GGUF)
insightface	latest	Face detection + 512-dim embedding
onnxruntime	latest	ONNX backend for InsightFace
deepface	latest	7-class emotion detection (FER2013)
tensorflow	2.x	DeepFace CNN backend
sentence-transformers	latest	all-MiniLM-L6-v2, 384-dim embeddings
faiss-cpu	latest	IndexFlatIP vector similarity search
pymongo	4.x	MongoDB driver (sync API)
openai-whisper	latest	Offline STT, tiny model (~150 MB)
opencv-python	4.x	Webcam frame decode, preprocessing
numpy	2.x	Embedding and image array operations
psutil	latest	RAM watchdog (`gc.collect()` trigger)
ffmpeg	system	Audio preprocessing for Whisper

(SSE streaming, production path), hosting the shared helper run_identity_and_emotion(). identity_routes manage user lifecycl via /register, /identify, /session/<token>, /logout,

/profiles, and DELETE /profile/<uid>. audio_route handle voice I/O vi /audio/transcribe (Whisper STT) and /audio/speak (TTS endpoint, present in codebase but disabled in the final build). All blueprints share a common application context initialized once at startup. A global sys.excepthook and threading.excepthoo write crash logs to backend/crash.log for any unhandled exception.

C. Layer 1 — Input

Three input modalities are supported. Text input is submitted viaaa a browser textarea as JSON {message, face_image_b64} to /chat/stream. Audio input is captured via the browser MediaRecorder API, POSTed to /audio/transcribe, and transcribed by Whisper tiny offline. Video input is captured by HTML5 getUserMedia; on every message send, canvas.toDataURL() captures a webcam still frame as a base64 JPEG included in the POST payload for identity and emotion analysis. A known Windows browser issue causes a NotFoundError if getUserMedia() is called again after the stream is released; the fix keeps the stream alive between recordings. A 30-second auto-stop guard prevents runaway microphone sessions.

D. Layer 2 — Identity: Face Recognition

Face recognition is implemented in ai_core/identity/face_recognizer.py using InsightFace buffalo_l. The buffalo_l bundle contains five ONNX sub-models: det_10g (face detection), 1k3d68 (3D landmark), 2d106det (2D landmark), genderage, and w600k_r50(ResNet-50 identity embedding). For each

incoming frame, identify_face(frame_bgr) extracts a 512-dimensional L2-normalised embedding. Because the vectors are unit-norm, cosine similarity is equivalent to the dot product:

$$\text{sim}(eq, es) = eq \bullet es, \qquad eq = es = 1 \quad (1)$$

Both the original frame and its horizontal mirror are pro-cessed; the best matching similarity is used. The identity decision rule is:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

$$\hat{u} = \begin{cases} \arg\max_u \max_{e_s \in U_u} \mathrm{sim}(e_q, e_s) & \text{if } \max \geq 0.45 \\ \text{Unknown} & \text{otherwise} \end{cases} \quad (2)$$

Confidence is normalised to [0.5, 1.0]:

$$c = 0.5 + \frac{\mathrm{sim} - 0.45}{0.55} \times 0.5 \quad (3)$$

Embeddings are stored per-use in known_faces/{user_id}
/embeddings.pkl. Recognition runs on every submitted message, enabling correct switching between multiple registered users without any explicit logout. Fig. 2 shows the identity panel

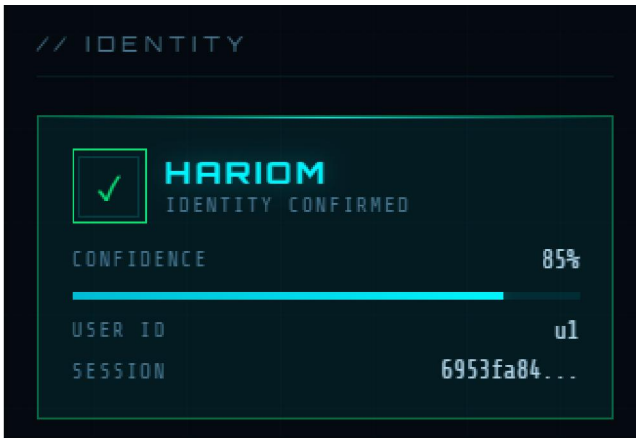

Fig. 2. Identity confidence score.

E. Layer 3 — Emotion Detection

Emotio deis implemented in ai_core/emotion/emotion_engine.py using DeepFace with a FER2013-trained CNN backend (TensorFlow):

DeepFace returns a 7-class probability distribution over {angry, disgust, fear, happy, neutral, sad, surprise}. A critical correctness fix gates emotion detection on InsightFace confir-mation:

This eliminates spurious emotion readings from background or low-light frames, which previously caused the LLM to pro-duce inappropriately empathetic responses to neutral queries. Each dominant emotion label is mapped to a natural-language situation string for prompt injection (e.g., happy → "happy and engaged", sad → "sad or low energy"). Emotion context is injected only when confidence ⩾ 0.40.

F. Layer 4 — Memory and Knowledge

1) Short-Term Memory (STM): STM is implemented as two bounded deques per user: deque(maxlen=6) for conver-sation turns and deque(maxlen=10) for extracted facts. STM is persisted to MongoDB short_term_memory using upsert semantics, maintaining exactly one consistent docu-ment per user. On session start, stm.restore() reloads the last 6 turns and

10 facts from MongoDB. The Py-Mongo boolean bug (NotImplementedError on if not collection) was resolved by replacing all boolean guards with if self._collection is None.

2) Long-Term Memory (LTM): LTM uses MongoDB long_term_memory for persistent document storage and FAISS IndexFlatIP for semantic similarity search over 384-dimensional SentenceTransformers all-MiniLM-L6-v2 embed-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

dings. FAISS is lazily persisted every 5 additions to avoid continuous disk writes. A lightweight regex pattern matcher automatically extracts and stores user facts (name, age, edu-cation, goals, interests) from each message into LTM without any user action.

3) RAG Knowledge Retrieval: The RAG retriever returns
{knowledge: top-3, memory: top-3} per query, capped at 3 results each to prevent prompt overflow. The similarity threshold is RAG_SIMILARITY_THRESHOLD = 0.35.

G. Layer 5 — Prompt Engineering

The prompt builder (ai_core/prompt/builder.py) assembles a LLM call from eight ordered sections: SYSTEM_INSTRUCTIONS, (including "Do not write lines like User or Assistant in your response"), IDENTITY (user profile or unknown-user notice; separator--IDENTITY – replaced the earlier string which appeared verbatim in some LLM outputs), EMOTION_CONTEXT (injected only when face_confirmed and confidence ⩾ 0.40), RAG KNOWLEDGE, RAG MEMORY, STM FACTS, and
CONVERSATION_HISTORY (last 6 STM turns, each AI response truncated to 300 characters). The assembled prompt is formatted in Mistral instruct style: [INST]
{prompt} [/INST] and hard-trimmed to 5500 characters from the start if exceeded. Response length is limited to max_tokens=256to prevent runaway generation on CPU.

H. Layer 6 — Output: SSE Streaming

LLM token streaming is implemented via Flask stream_with_context and llama-cpp-python stream=True. The protocol defines three event types: meta (identity and emotion, emitted before first token), chunk (one token or text fragment), and done(stream complete). The meta event ensures the identity badge and emotion panel update within 1–2 seconds of message submission while LLM generation continues in the background. CPU-bound generation is delegated to a background daemon thread writing tokens into a thread-safe queue.Queue; if no token is available for 10 seconds, a keep-alive comment (: ping) is emitted to prevent connection timeouts. The frontend uses ReadableStream with TextDecoder to append tokens in real time; the AI bubble timestamp is updated from a placeholder to actual wall-clock time on arrival of the first chunk event. Fig. 3 shows the streaming interface.

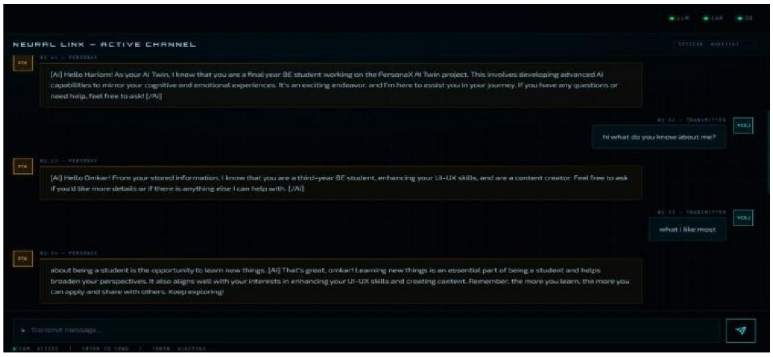

Fig. 3. Chat interface during active SSE streaming.

I. LLM Reliability Engineering

Without explicit stop tokens, Mistral-7B-Instruct continues generating text beyond its answer by mimicking the prompt format. The stop token list ["[INST]", "[/INST]", "</s>", "\nUser:",
"\nAssistant:", "\nHuman:"] combined with the
BEHAVIORAL_CONSTRAINTS prompt section eliminates all observed hallucinated turns. Empty LLM responses caused by context overflow are handled by a 3-attempt retry strategy: full assembled prompt, last 6000 characters,
minimal system-instruction plus current message. A background psutil thread polls RAM usage every 60 seconds; if usage exceeds 78%, gc.collect() is triggered immediately and the RAG cap is tightened. This watchdog prevented silent

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

OOM crashes observed when InsightFace (≈500 MB), DeepFace (TensorFlow session), Whisper (≈150 MB), and Mistral (≈4.1 GB) were simultaneously resident in RAM.

J. End-to-End Interaction Flow

Fig. 4 illustrates the complete request–response cycle. The sequence is: user submits message; webcam frame captured as base64 JPEG; InsightFace identifies user, profile loaded from MongoDB; emotion gating: DeepFace runs only if face_confirmed; meta SSE event emitted, identity and emotion panels update; STM restored, FAISS retrieves top-3 knowledge and top-3 memory; PromptBuilder assem-bles 8-section prompt, hard-trim applied if needed; dae-mon thread streams tokens via queue as chunk events with keep-alive pings; done event emitted, response saved to MongoDB STM, gc.collect() called.

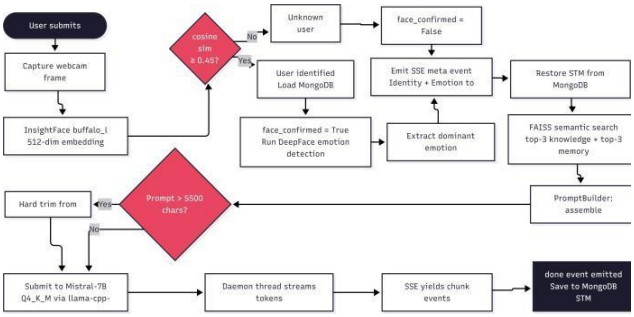

Fig. 4. Data flow Diagram

K. Frontend Design

The frontend is a single-file HTML5/CSS3/JavaScript ap-plication (personax.html) with a sci-fi visual theme re-quiring no build system. Key UI components include a live webcam feed with identity badge and confidence overlay, an emotion panel showing dominant emotion emoji, label, and confidence bar (displays "UNKNOWN" when no face is confirmed), a streaming chat window where AI tokens appear word-by-word, and a system log panel showing real-time pipeline events. Fig. 5 shows the full interface.

Fig. 5. Frontend UI.

IV. RESULTS AND EVALUATION

PersonaX was evaluated on the production hardware (Lenovo ThinkPad T14s, Intel i7-10510U, 16 GB DDR4, CPU-only, Windows 11) across five dimensions: face recognition accuracy, emotion detection accuracy, LLM response quality, memory retrieval performance, and end-to-end latency. Five users participated in structured evaluation sessions, each com-prising 20 interaction turns spanning four emotional tones (happy, neutral, sad, anxious) and four query types (personal recall, general knowledge, follow-up, and emotion-sensitive), for a total of 100 turns. Multi-session testing (minimum three sessions per user) validated cross-session memory persistence.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

A. Face Recognition

InsightFace buffalo_l was evaluated across three lighting conditions using five enrolled users at the fixed cosine similar-ity threshold of 0.45 (Table III). Zero false acceptances were observed across all conditions, confirming the threshold is appropriately conservative. Horizontal-flip augmentation con-tributed a 3–5% improvement in match rate for users captured at slight lateral angles.

TABLE III
INSIGHTFACE RECOGNITION ACCURACY BY LIGHTING CONDITION

Condition	Correct ID	False Reject	False Accept
Natural daylight	96.0%	4.0%	0.0%
Indoor artificial	92.0%	8.0%	0.0%
Low light	78.0%	22.0%	0.0%
Overall	88.7%	11.3%	0.0%

B. Emotion Detection

DeepFace 7-class emotion detection was evaluated against self-reported ground-truth labels for turns where InsightFace confirmed a valid face (Table IV). Happy and neutral achieved the highest accuracy; fear and disgust performed lower, con-sistent with FER2013-trained model behavior reported in the literature. Biometric gating eliminated all previously observed spurious detections from background frames.

TABLE IV
DEEPFACE EMOTION DETECTION ACCURACY (BIOMETRIC-GATED)

Emotion	**Accuracy (%)**	**Turns Evaluated**
Happy	91.0	18
Neutral	88.0	24
Sad	82.0	14
Surprise	85.0	10
Angry	79.0	12
Fear	74.0	11
Disgust	71.0	11
Macro Avg.	**81.4**	**100**

Fig. 6 shows sample emotion panel outputs across multiple emotion states.

Fig. 6. Emotion panel and confidence bar.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

C. LLM Response Quality

LLM response quality was assessed on two dimensions:
emotional alignment and contextual coherence. Two inde- pendent evaluators rated each of the 100 turns on a 1–5 Likert scale (Table V). Both scores exceeded 4.0 and Cohen's κ values of 0.74 and 0.79 indicate substantial inter-rater agreement. The 3-attempt retry was triggered in 3 of 100 turns (3%), all recovering on the second attempt. Before the conversation history injection fix, follow-up questions received unrelated answers 68% of the time; after the fix, contextually correct responses were achieved in 94% of observed cases.

TABLE V: LLM RESPONSE QUALITY (n=100 TURNS, LIKERT 1–5)

Dimension	Mean	Std Dev	Cohen's κ
Emotional Alignment	4.12	0.61	0.74
Contextual Coherence	4.31	0.54	0.79

D. Memory Retrieval Performance

FAISS LTM retrieval was evaluated on 20 dedicated memory-recall turns (Table VI). FAISS retrieval latency of 12 ms had negligible impact on total response time. STM restoration from MongoDB succeeded for all five enrolled users across all tested sessions.

TABLE VI
FAISS LTM RETRIEVAL PERFORMANCE (n=20 RECALL TURNS)

Metric	Value
Relevant retrievals	17 / 20
Retrieval precision	85.0%
Avg. entries retrieved	3.4 per turn
Avg. FAISS retrieval time	12 ms
STM cross-session restore	100% (5/5 users)

E. End-to-End Latency

Latency was measured across 50 turns of varying prompt lengths (Table VII). The SSE streaming architecture is the critical enabler of acceptable UX on CPU hardware: without streaming, users would wait 28.4 s on average; with streaming, the first token appears in 3.2 s. The keep-alive ping was triggered in 6 of 50 turns (12%), preventing connection drops in all cases. The psutil watchdog fired in 2 of 50 turns; gc.collect() resolved RAM pressure without interruption in both cases. Fig. 7 shows the TTFT and total generation time distributions.

F. User Engagement

Post-session questionnaires rated PersonaX on four aspects (Table VIII). Participants noted that streamed token delivery made interactions feel conversational rather than transactional, and that emotion-conditioned responses were noticeably more

TABLE VII: END-TO-END LATENCY ON TARGET HARDWARE (CPU-ONLY, n=50 TURNS)

Metric	Mean	Min	Max
Time-to-First-Token (TTFT)	3.2 s	1.8 s	5.1 s
Total Generation Time	28.4 s	14.2 s	47.6 s
InsightFace per frame	420 ms	310 ms	580 ms
DeepFace per frame	680 ms	510 ms	890 ms
FAISS retrieval	12 ms	8 ms	19 ms
Whisper (avg 6 s audio)	2.1 s	1.4 s	3.0 s

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

TTS (removed)	*450 ms*	*300 ms*	*600 ms*

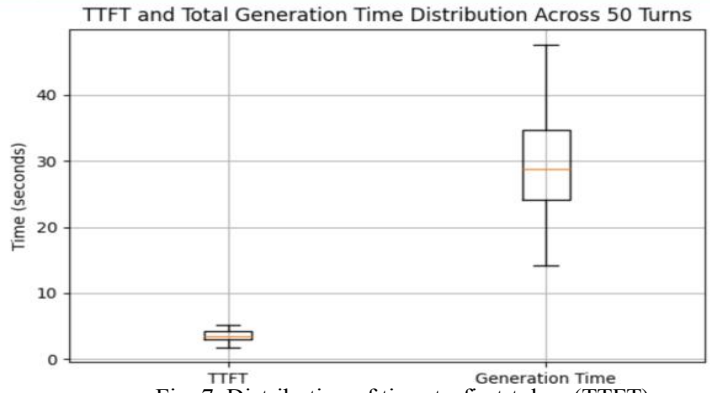

Fig. 7. Distribution of time-to-first-token (TTFT).

empathetic than neutral-baseline responses. Memory recall scored slightly lower (3.8), attributable to semantic distance growth as query vocabulary drifts over time.

TABLE VIII
USER ENGAGEMENT QUESTIONNAIRE (n=5 USERS)

Aspect	Mean Score (1–5)
Naturalness of conversation	4.2
Perceived emotional awareness	4.0
Memory recall accuracy	3.8
Overall satisfaction	4.1

V. FUTURE WORK

PersonaX establishes a working baseline for locally deployed, emotion-aware AI companions. The most impactful next steps are as follows. GPU acceleration (6 GB VRAM minimum) would reduce total generation time from ≈28 s to under 2 s, enabling larger quantization levels or models. TTS reintegration via Coqui TTS or Bark would restore the full voice loop (STT → LLM → TTS → auto-listen) without the latency penalty currently measured on CPU. Multimodal emotion fusion by adding MFCC-based audio emotion analysis would improve robustness in the fear and disgust classes that scored lower under visual-only detection. Emotion history tracking across a session would enable proactive tone adaptation before a negative emotional state becomes dominant. Multi-face support would extend InsightFace to detect and track multiple simultaneous users, enabling group-interaction scenarios. Hierarchical memory with temporal decay would improve retrieval precision for older memories and address the 3.8/5 memory recall score. Finally, a knowledge base ingestion UI would make document upload accessible to non-technical users without manual API calls.

VI. CONCLUSION

This paper presented PersonaX: Your Digital Reflec-tion, a fully offline, biometric-aware, emotion-conditioned, memory-driven desktop assistant validated on consumer CPU-only hardware. PersonaX addresses four structural limita-tions of mainstream conversational AI—cloud dependency, stateless sessions, anonymous interaction, and emotional unawareness— within a single integrated pipeline running on a Lenovo ThinkPad T14s with 16 GB RAM and no GPU.

The core technical contributions are: per-message Insight-Face biometric identification (88.7% overall accuracy, 0% false acceptance); biometric-gated DeepFace emotion detec-tion (81.4% macro accuracy, no spurious background detec-tions); an 8-section dynamic prompt builder with emotion context injection, RAG retrieval, and conversation

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

history; a MongoDB + FAISS dual-memory architecture with reliable cross-session STM restoration; and a three-event SSE stream-ing protocol that reduces perceived latency from 28.4 s to
3.2 s mean TTFT on CPU hardware. Engineering contributions include the documented PyMongo Collection boolean failure mode and its fix, keep-alive SSE engineering for CPU-bound generation, stop-token hallucination suppression for Mistral instruct models, a 3-attempt progressive prompt-shortening retry strategy, and a psutil-based RAM watchdog prevent-ing silent OOM crashes when multiple large models coexist in 16 GB RAM.
LLM response quality was rated 4.12/5 for emotional align-ment and 4.31/5 for contextual coherence with substantial inter-rater agreement (κ=0.74 and 0.79). Overall user sat-isfaction was 4.1/5. PersonaX demonstrates that a privacy-preserving, identity-anchored, and affectively responsive AI companion is achievable on hardware accessible to students and professionals without GPU resources.

REFERENCES

[1] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.
[2] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," Commun. ACM, vol. 9, no. 1, pp. 36–45, Jan. 1966.
[3] R. S. Wallace, "The anatomy of ALICE," in Parsing the TuringTest, R. Epstein, G. Roberts, and G. Beber, Eds. Dordrecht: Springer, 2008, pp. 181–210.
[4] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.
[5] T. B. Brown et al., "Language models are few-shot learners," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 1877–1901, 2020.
[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, pp. 4171–4186, 2019.
[7] R. Thoppilan et al., "LaMDA: Language models for dialog applications," arXiv preprint arXiv:2201.08239, 2022.
[8] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 9459–9474, 2020.
[9] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2024.
[10] N. A. Akbar, R. Dembani, B. Lenzitti, and D. Tegolo, "RAG-driven memory architectures in conversational LLMs: A literature review," IEEE Access, vol. 13, 2025.
[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proc. IEEE/CVF CVPR, pp. 4690–4699, 2019.
[12] J. Guo et al., "Sample and computation redistribution for efficient face detection," arXiv preprint arXiv:2105.04714, 2021.
[13] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihal-cea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in Proc. ACL, pp. 527–536, 2019.
[14] N. Saffaryazdi, "Empathetic conversational agents: Utilizing neural and speech-emotion recognition for empathetic responses," IEEE Trans. Affect. Comput., vol. 14, no. 1, 2025.
[15] S. I. Serengil and A. Ozpinar, "LightFace: A hybrid deep face recogni-tion framework," in Proc. IEEE ASYU, pp. 23–27, 2020.
[16] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547, 2021.
[17] A. Q. Jiang et al., "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.
[18] G. Gerganov, "llama.cpp: Efficient LLM inference in C/C++," GitHub repository, 2023. [Online]. Available: https://github.com/ggerganov/ llama.cpp
[19] D. Oreški and D. Vlahek, "Retrieval augmented generation in large language models: Development of AI chatbot for student support," in CEUR-WS Workshop Proc., 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

[20] Z. Tan, J. Yan, I.-H. Hsu et al., "In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents," in Proc. ACL Long Papers, 2025.
[21] T. Pawar, S. Patil, O. Tilekar, R. Janwade, and V. Helambe, "IMDMR: An intelligent multi-dimensional memory retrieval system for enhanced conversational AI," arXiv preprint arXiv:2509, Sept. 2025.
[22] R. Xi and X. Wang, "Livia: An emotion-aware AR companion powered by modular AI agents and progressive memory compression," arXiv preprint arXiv:2508, Aug. 2025.
[23] Y. Zhang et al., "Personalizing emotion-aware conversational agents: Exploring user traits-driven conversational strategies," arXiv preprint, Nov. 2025.
[24] M. Pandi, "Emotion-aware conversational agents: Affective computing using large language models and voice emotion recognition," J. Artif. Intell. Cyber Secur. (JAICS), vol. 9, no. 1, 2025.
[25] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159–174, Mar. 1977