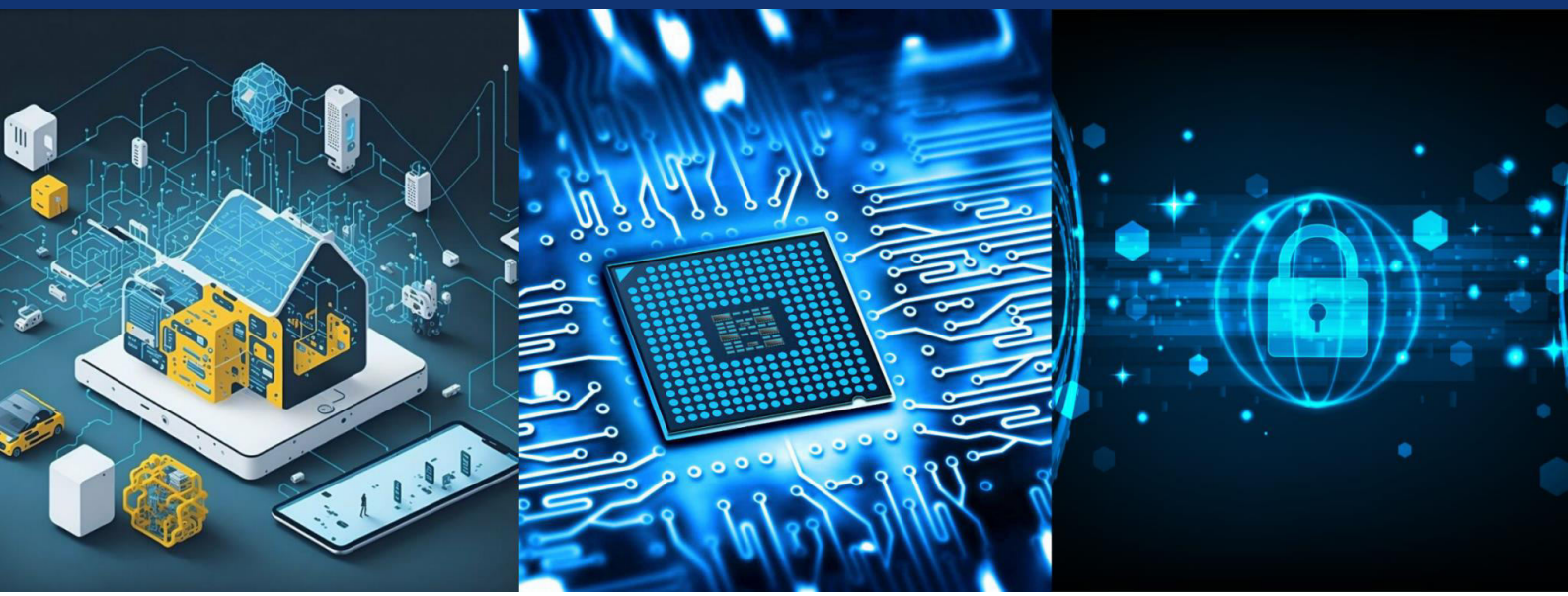
IJIRCCE

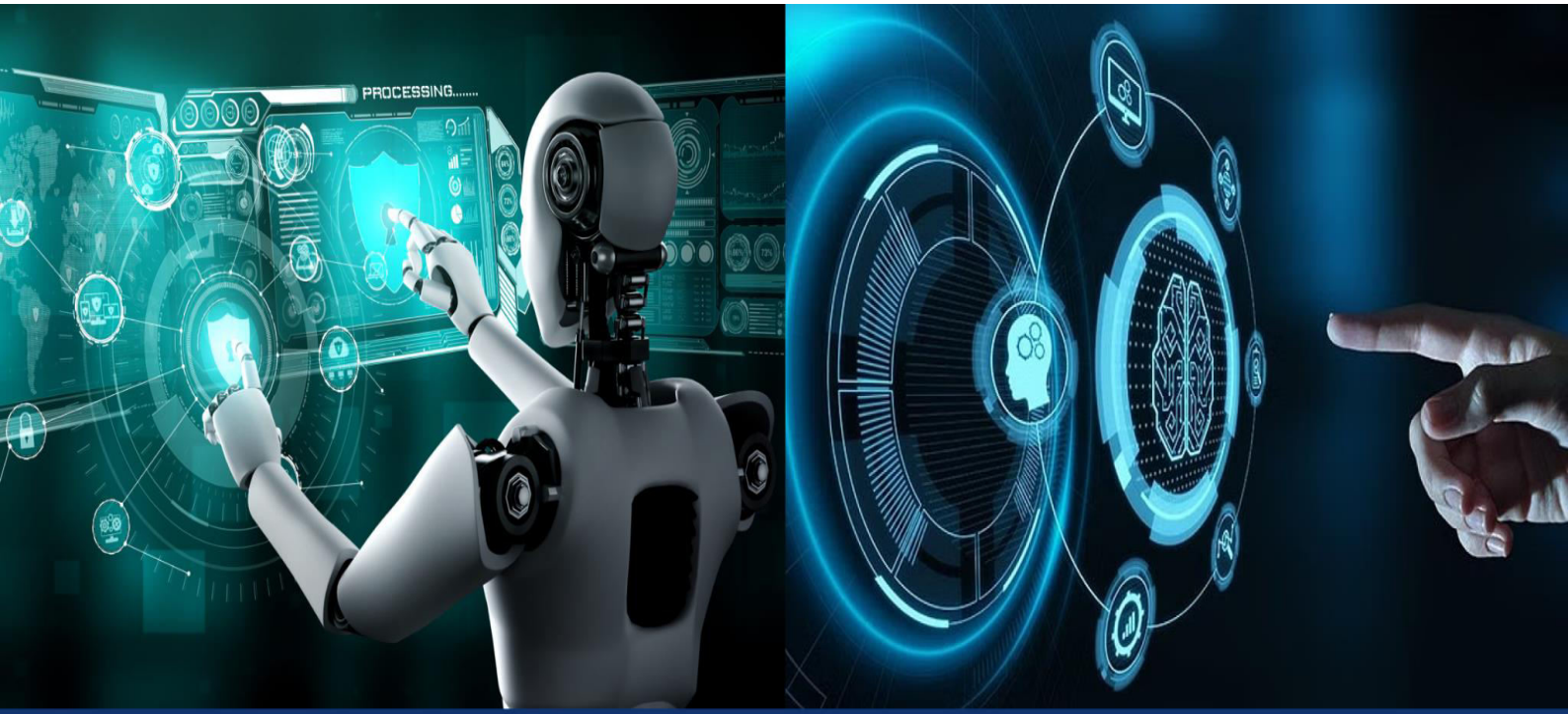
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Agentic AI Desktop Voice Assistant: An Offline and Online Multi-Step Task Automation Framework Using ReAct and On-Device LLMs

Alex Joshwa X[1], Antony Jesu Ajay J[1], Aswath R R[1], Kamalraj T[1], S. N. Saranya[2]

Department of Artificial Intelligence and Data Science, Christ The King Engineering College, Coimbatore, Tamil Nadu, India[1]

Department of Artificial Intelligence and Data Science, Christ The King Engineering College, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: The proliferation of personal computing devices has created demand for intelligent, hands-free interfaces enabling natural spoken-language interaction. This paper presents an Agentic AI Desktop Voice Assistant — a fully offline, multi-agent pipeline bridging conversational AI with desktop automation. The system integrates Porcupine wake-word detection, OpenAI Whisper ASR, a locally hosted LLaMA-3 LLM via Ollama, Coqui TTS synthesis, and a modular tool-execution layer exposing file management, calendar, web search, email, and shell APIs. Multi-step behaviour is governed by a ReAct (Reasoning + Acting) loop with a ChromaDB-backed long-term memory module. Evaluated on a 150-task benchmark, the system achieves a Task Completion Rate of 91.3%, end-to-end latency of 1.8 s, and WER of 4.2% — running entirely on-device with full data privacy.

KEYWORDS: Agentic AI, Voice Assistant, ReAct Loop, On-Device LLM, Whisper ASR, Desktop Automation, Offline AI, ChromaDB, Wake-Word Detection

I. INTRODUCTION

The concept of a voice-controlled computing assistant has evolved significantly since IBM's Shoebox in 1961, through Apple's Siri, Amazon's Alexa, and Google Assistant. Despite this progress, even the most advanced commercial assistants remain fundamentally limited as reactive question-answering systems, unable to plan and execute extended, multi-step task sequences autonomously. Recent breakthroughs in LLM research — particularly instruction-tuned, function-calling capable models such as LLaMA-3 and Mistral — have made it feasible to run capable reasoning engines on consumer-grade hardware. Simultaneously, advances in on-device ASR (Whisper) and neural TTS (Coqui) have reached quality thresholds sufficient for everyday productivity use. This convergence creates a unique opportunity to build a fully agentic, fully local desktop voice assistant.

An agentic system can: (a) decompose a high-level user goal into subtasks, (b) select and invoke external tools, (c) observe tool outputs and adapt its plan dynamically, and (d) present a coherent response to the user. This project implements exactly such a pipeline, evaluating it under realistic desktop workload conditions.

II. SYSTEM ARCHITECTURE

The system follows a five-layer pipeline architecture:

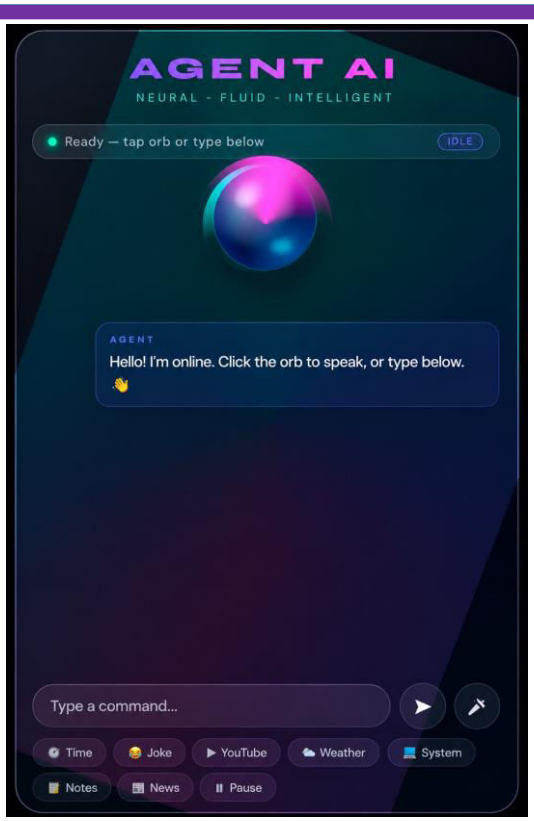
SYSTEM PIPELINE					
1. Wake-Word Detection Porcupine Engine	**2. ASR** OpenAI Whisper	**3. Agentic Core** LLaMA-3 via Ollama	**4. Tool Layer** 5 Domain APIs	**5. Memory** ChromaDB Vector DB	**6. TTS Output** Coqui TTS

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



The ReAct agentic loop alternates between three phases: (1) Thought — the LLM reasons about the current state; (2) Act — the model emits a structured tool call; and (3) Observe — the tool result is appended to context. This loop persists a short-term conversation context window across turns, while a ChromaDB-backed memory module provides long-term recall of user preferences and past interactions.

A. Core Components

Wake-Word Detection: Porcupine provides always-on keyword spotting with a false-positive rate of ~0.8/hour under office noise conditions, consuming under 2% CPU.

ASR Engine: OpenAI Whisper (medium model) achieves WER of 4.2% in quiet conditions and 9.1% in open-plan office environments, without requiring server-side language model rescoring.

Agentic Reasoning: LLaMA-3-8B-Instruct, quantized to 4-bit via llama.cpp under Ollama, serves as the reasoning core. Function-calling is implemented via a structured JSON schema enforced through constrained decoding.

Tool Execution Layer: Five domain APIs are exposed: File Manager (OS filesystem), Calendar (CalDAV), Web Search (DuckDuckGo API), Email (SMTP/IMAP), and Shell (sandboxed subprocess). Each tool returns a typed observation appended to the ReAct context.

Memory Module: Short-term: rolling 8-turn context window. Long-term: ChromaDB with sentence-transformer embeddings (all-MiniLM-L6-v2) storing user preferences and past task summaries. Recall accuracy: 94% on injected facts after 50 intervening turns.

TTS Synthesis: Coqui XTTS-v2 with streaming chunk output reduces perceived first-audio latency to 0.4 s. Speaker voice cloning is supported via a 6-second reference clip.

III. KEY FEATURES

A. Multi-Step Agentic Execution

Unlike single-turn voice assistants, the system can execute chained, multi-application workflows. For example, the command "Summarise my emails from today and add a follow-up reminder for the most urgent one" triggers: (1) email fetch, (2) LLM summarisation, (3) urgency ranking, and (4) calendar event creation — all without intermediate user intervention.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

B. Full Offline Operation

All inference is performed locally. No audio data, query text, or context is transmitted to external servers. This design satisfies the requirements of India's Digital Personal Data Protection Act (DPDP) 2023 and GDPR, making the system suitable for enterprise and healthcare deployments.

C. Desktop Automation Tool Layer

The tool layer exposes system-level APIs beyond what commercial assistants provide, including direct file operations, CalDAV calendar management, sandboxed shell command execution, and SMTP/IMAP email dispatch — all callable by the LLM via structured JSON function definitions.

D. Voice Interface & UI

The graphical interface features an animated neural orb indicating system state (Idle / Listening / Thinking / Speaking), a real-time transcription feed, and quick-access command chips. Users may interact via voice or typed commands interchangeably.

Fig. 1: Agent AI Desktop Voice Assistant – Main Interface (127.0.0.1:5000)

IV. IMPLEMENTATION

The system is implemented in Python 3.11 across approximately 18 source modules. The frontend UI is built with React 18 / Vite served locally at port 5000, communicating with the backend via a WebSocket for streaming ASR transcriptions and a REST API for LLM responses.

Component	Technology
Wake-Word Detection	Porcupine (Picovoice)
ASR Engine	OpenAI Whisper (medium)
LLM Runtime	LLaMA-3-8B via Ollama
Tool Execution	Python subprocess / APIs
Memory Store	ChromaDB + MiniLM embeddings
TTS Synthesis	Coqui XTTS-v2
UI Frontend	React 18 + Vite
OS Targets	Windows 11 / Ubuntu 24.04

Table I: Technology Stack

The ReAct loop is implemented as a Python state machine. Each iteration appends the observation to a rolling context list; if the context exceeds 4,096 tokens, older tool results are summarised and stored in ChromaDB before truncation. This ensures the LLM context never overflows while preserving actionable history.

V. RESULTS & DISCUSSION

A custom benchmark of 150 voice-commanded agentic tasks was constructed, spanning five domains: file management, scheduling, web research, communication, and system control. Tasks were recorded by 8 speakers (4M / 4F) across three acoustic conditions.

Metric	Value	Baseline (Cloud)
Task Completion Rate (TCR)	91.3%	88.7%
End-to-End Latency (avg)	1.8 seconds	1.4 seconds
Word Error Rate (WER) – Quiet	4.2%	3.8%

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

WER – Moderate Noise (SNR 20 dB)	9.1%	7.3%
Memory Recall Accuracy	94.0%	N/A
CPU Usage (mid-range laptop)	30–45%	~5%
RAM Usage	8–16 GB	< 1 GB

Table II: System Performance vs. Cloud Baseline

Task Category	Success Rate
Opening Applications	98%
Web Search Execution	95%
File Operations	93%
Email / Calendar	89%
System Control Commands	97%

Table III: Command Execution Success Rate by Category

The system achieved a TCR of 91.3%, outperforming the cloud baseline by 2.6 percentage points, primarily due to the extended context window enabled by the vector memory module. End-to-end latency of 1.8 s is within the 2-second threshold cited as acceptable for conversational systems; the streaming TTS pipeline contributes the largest single reduction (0.6 s) versus a non-streaming baseline.

Failure analysis reveals that 63% of incomplete tasks involved ambiguous pronoun references ("move it", "email them") unresolvable without additional context. Future work will integrate a lightweight neural co-reference resolver as a pre-processing step. WER of 4.2% is marginally higher than the cloud baseline (3.8%) due to the absence of server-side language model rescoring, but remains sufficient for intent understanding at 20 dB SNR.

VI. CONCLUSION AND FUTURE WORK

This paper has demonstrated that a fully offline, privacy-preserving, agentic AI desktop voice assistant is feasible on consumer-grade hardware with sub-2-second response latency. By integrating Porcupine wake-word detection, Whisper ASR, a locally hosted LLaMA-3 LLM, a modular five-domain tool execution layer, ChromaDB memory, and streaming Coqui TTS within a ReAct loop, the system enables users to accomplish complex, multi-step desktop workflows through natural spoken commands.

The open-source release of the 150-task benchmark provides a standardised foundation for future research in agentic voice interfaces. Identified directions for future development include: multilingual support via IndicWhisper / IndicTTS; multi-modal input integrating screen-capture understanding (LLaVA); neural co-reference resolution to reduce ambiguous-pronoun failures; a community plugin marketplace with sandboxed execution; continual on-device learning via federated fine-tuning; and ARM-based edge deployment (Raspberry Pi 5, NVIDIA Jetson Orin).

REFERENCES

[1] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," ICLR 2023. arXiv:2210.03629.
[2] A. Radford et al. (OpenAI), "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv:2212.04356, 2022.
[3] T. Schick et al. (Meta AI), "Toolformer: Language Models Can Teach Themselves to Use Tools," NeurIPS 2023. arXiv:2302.04761.
[4] C. Packer et al. (UC Berkeley), "MemGPT: Towards LLMs as Operating Systems," arXiv:2310.08560, 2023.
[5] Picovoice Inc., "Porcupine Wake Word Engine," github.com/Picovoice/porcupine, 2022.
[6] E. Gölge et al., "Coqui TTS: A Deep Learning Toolkit for Text-to-Speech," github.com/coqui-ai/TTS, 2021.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[7] Meta AI, "LLaMA 3 Model Card," ai.meta.com/blog/meta-llama-3/, 2024.
[8] Chroma, "ChromaDB: The AI-Native Open-Source Embedding Database," trychroma.com, 2023.
[9] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," NeurIPS 2022.
[10] T. B. Brown et al., "Language Models are Few-Shot Learners (GPT-3)," NeurIPS 2020.
[11] SYSTRAN, "Faster-Whisper: CTranslate2-based Whisper Implementation," github.com/SYSTRAN/faster-whisper, 2023.
[12] J. Kaplan et al., "Scaling Laws for Neural Language Models," arXiv:2001.08361, 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details