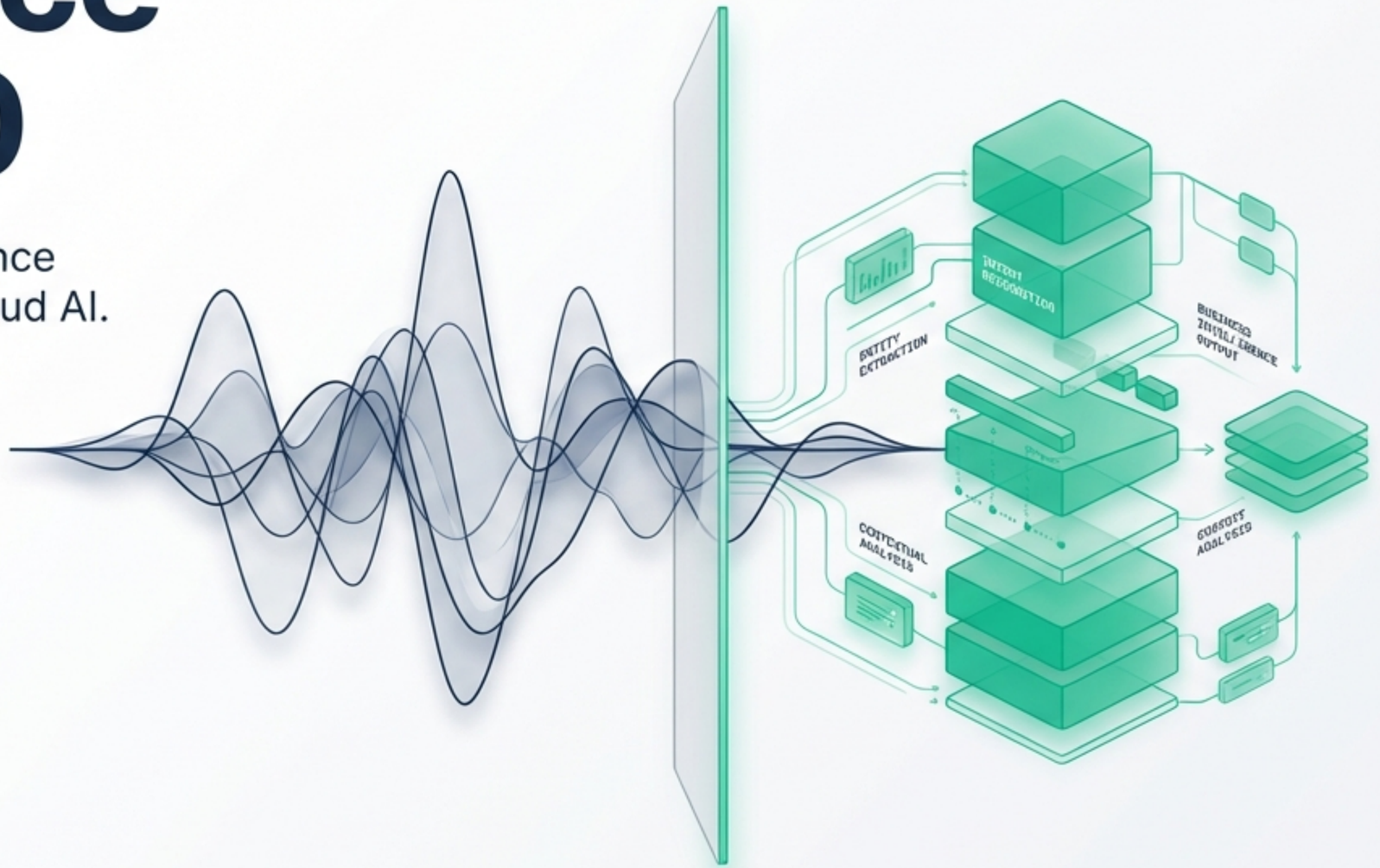
HPE AI Voice Agent v4.0

A Full-Stack Conversational Intelligence Solution Powered by HPE Private Cloud AI.



Microservices Architecture | Frontend: Gradio/React | Backend: WebSocket/FastAPI | Storage: PostgreSQL

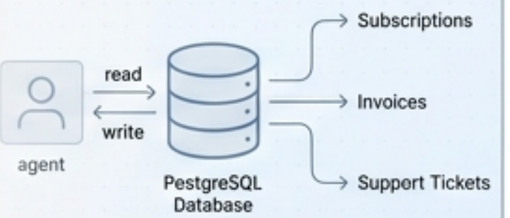
Beyond Chatbots: A Transactional Voice Operating System

Real-Time Orchestration

Low-latency event pipeline combining Whisper ASR for recognition, LLM for intelligence, and XTTS v2 for neural synthesis.

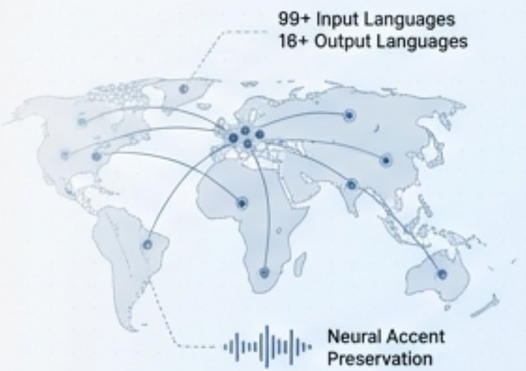
Deep Business Integration

Direct PostgreSQL read/write access. The agent manages subscriptions, processes invoices, and updates support tickets in real-time.



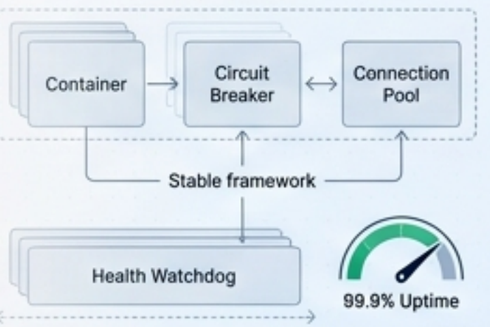
Global Accessibility

Input support for 99+ languages with auto-detection. Output in 16+ languages (EN, ES, FR, DE, JP, etc.) with neural accent preservation.

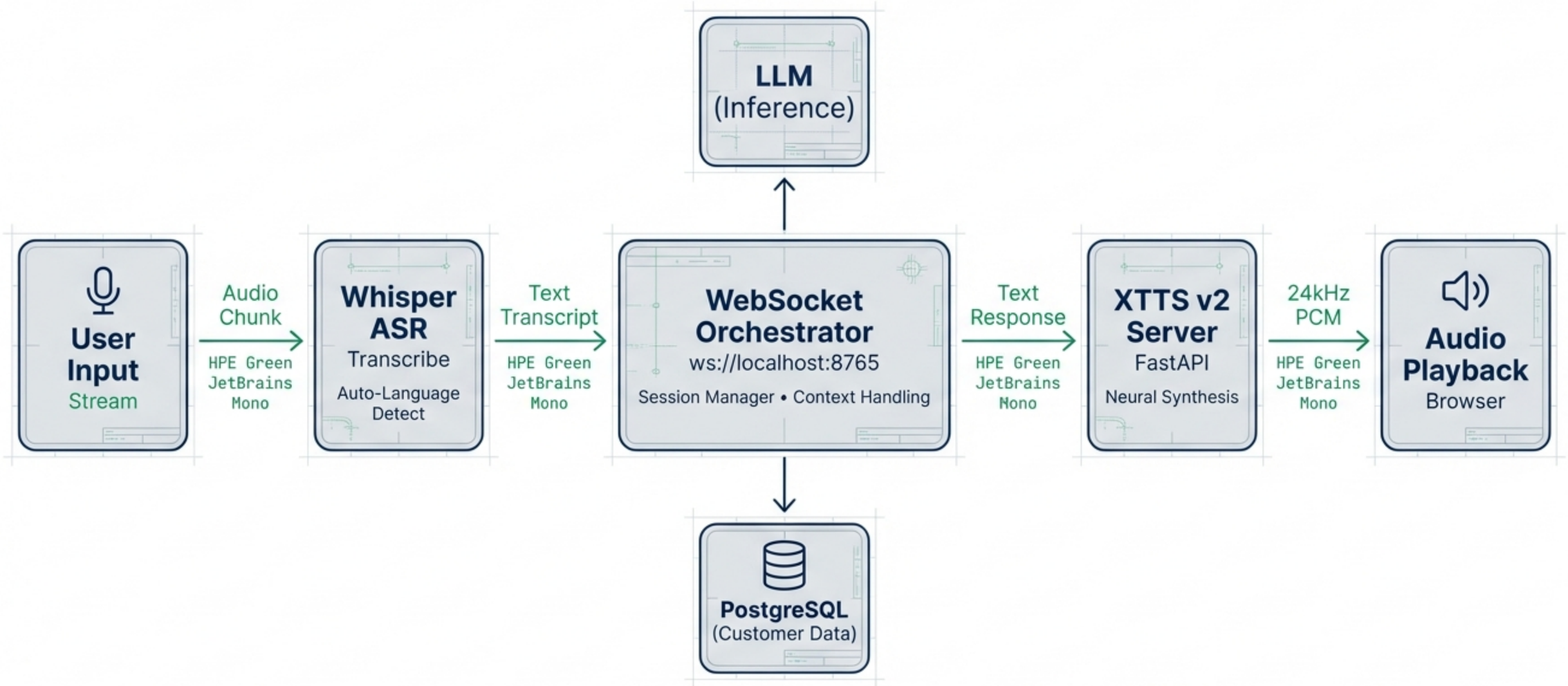


Enterprise Reliability

Containerized architecture featuring circuit breakers, connection pooling, and automated health watchdogs for 99.9% uptime.



End-to-End Event Pipeline



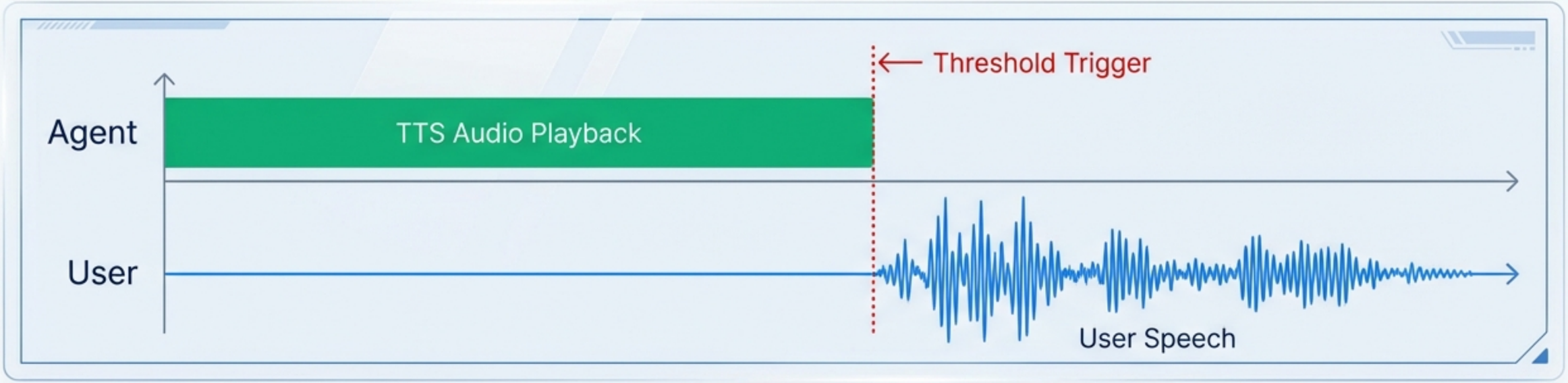
Orchestrator manages state via session_id hashing and persistent WebSocket connections.

Immersive User Experience (UI/UX)

The frontend leverages modern CSS glassmorphism and real-time JavaScript visualizers to create a responsive, high-end interface.

- **Glassmorphism**: translucent panels with `backdrop-filter: blur(10px)`.
- **Live Visualization**: 'Siri-style' multi-wave rendering.
- **Responsive Feedback**: Animated buttons and status badges.

Conversational Mechanics: The "Barge-In" Logic



Voice Activity Detection (VAD)

Tunable sensitivity slider (Range 5-50) with hysteresis logic (Start Threshold vs. Stop Threshold). Includes automated noise floor calibration to prevent false positives.

Interruption Event

When user audio amplitude > threshold:

```
window.vad.bargeIn = true;
audioPlayer.pause();
audioPlayer.currentTime = 0;
```

Global Input, Controlled Output

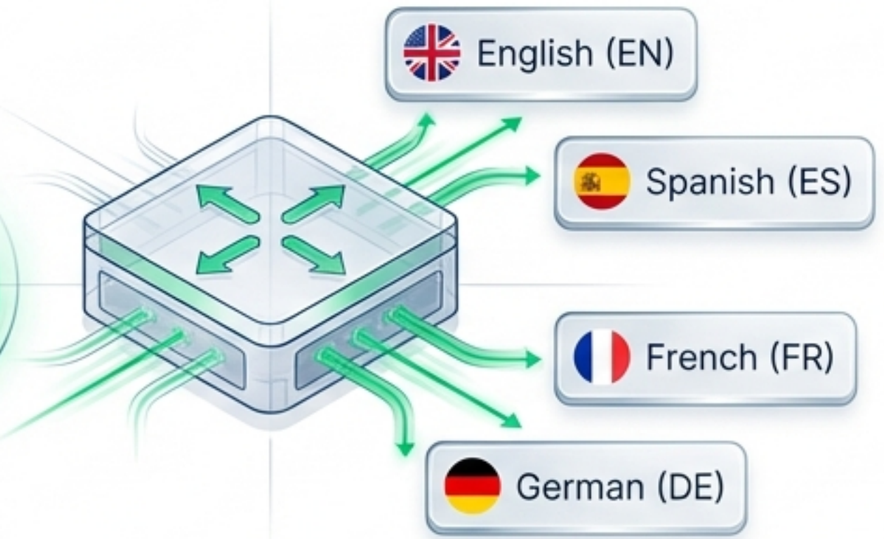
Decoupled logic for understanding vs. speaking.

INPUT (Whisper ASR)

Auto-detects 99+ languages.
Zero configuration required.

Smart Routing: Understands Hebrew input → Responds in English (if configured)

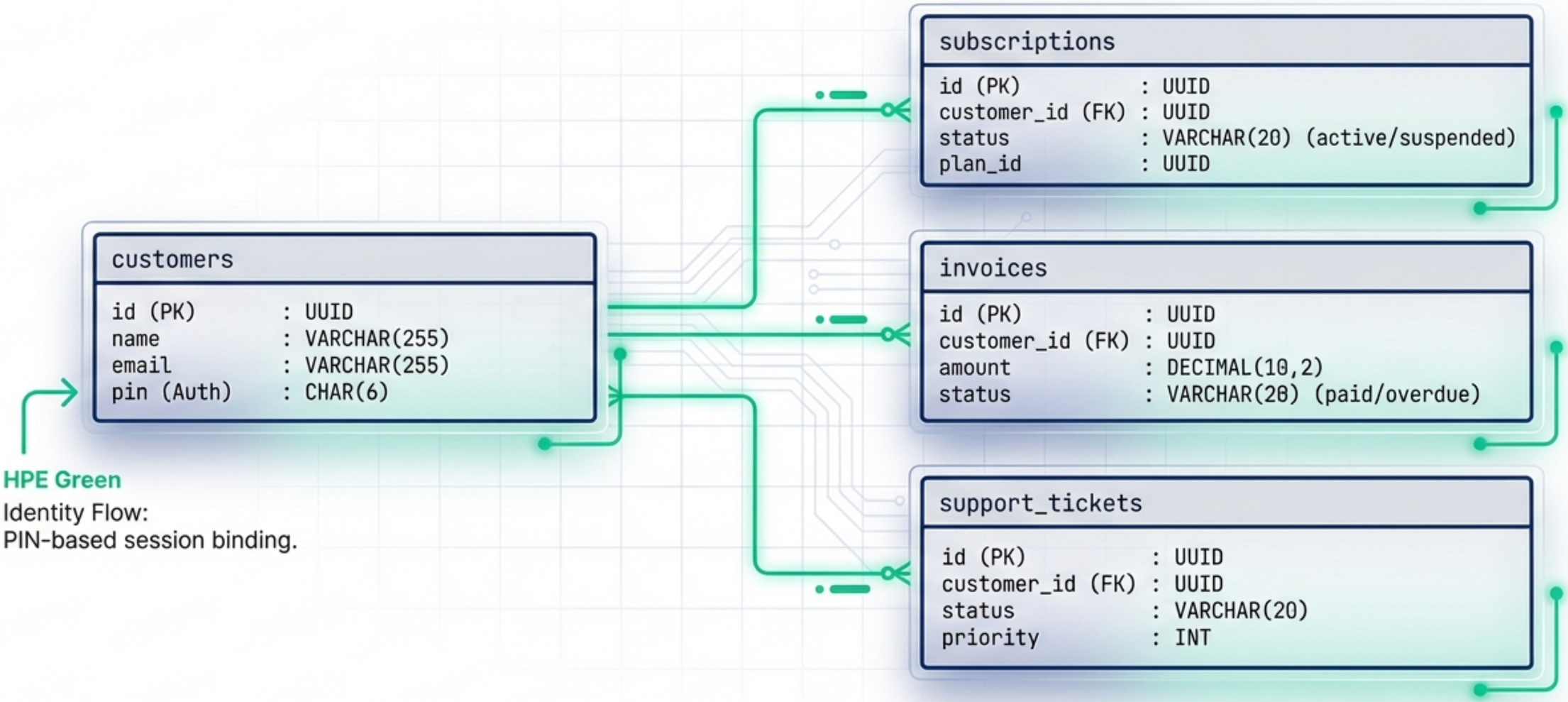
OUTPUT (XTTS v2)



16+ Neural Voices. Backend enforces specific `response_instruction` prompts.

Persistent Memory & Identity Resolution

Schema definition for the `customer_service` PostgreSQL database.

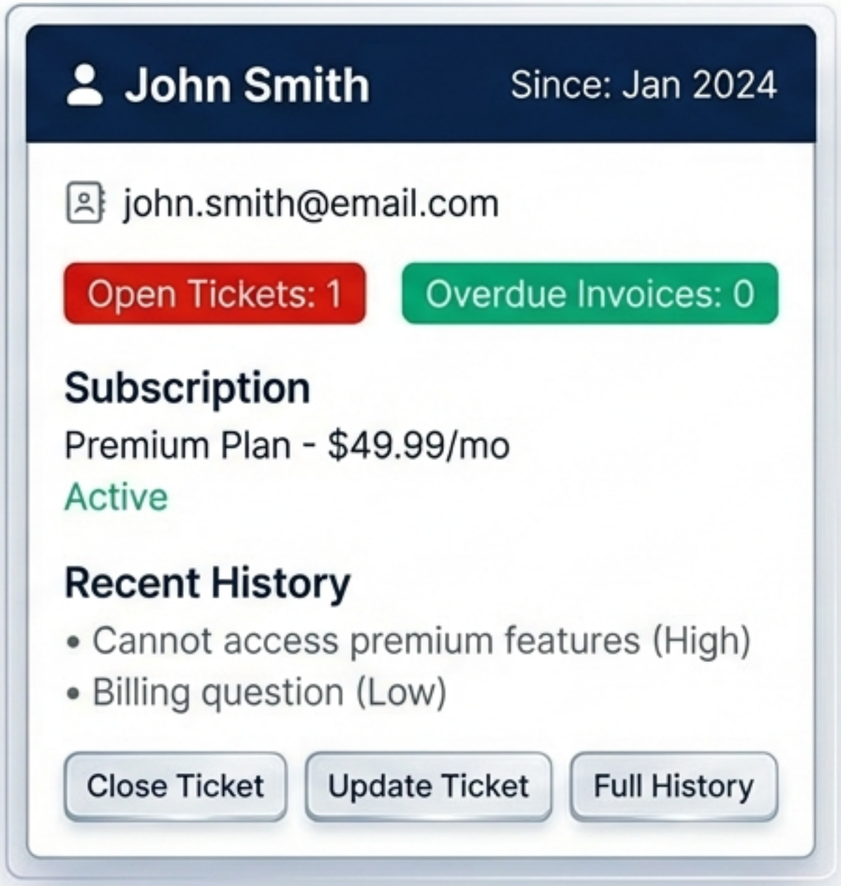


HPE Green

Identity Flow:
PIN-based session binding.

The Agent’s View: Dynamic Customer Context

Constructed in real-time via `format_customer_info_html`, this is the exact data context provided to the AI.



Actionable Intelligence & Transactions

The agent performs CRUD operations, not just text generation.

Ticket Management

Action:
`Create / Update / Delete`

Context:
Voice commands trigger SQL updates to `support_tickets` table.

Billing & Plans

Action:
`Query / Upgrade`

Context:
Process plan changes via `upgrade_requests` table logic.

Sentiment Analysis

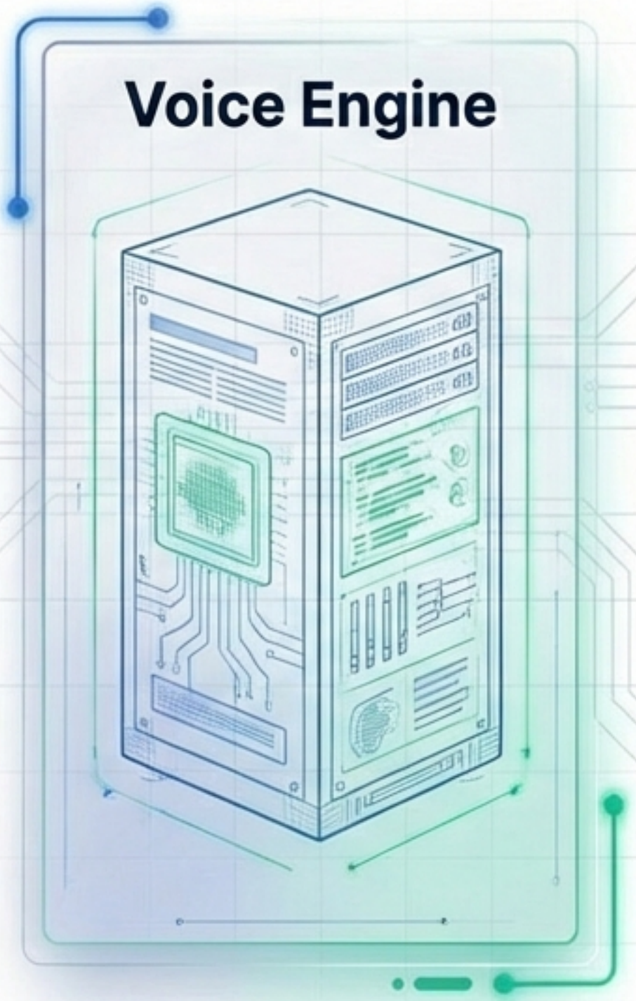
Action:
`Churn Detection`

Context:
Real-time logging to `sentiment_alerts`. Triggers: '**Risk: High**', '**Intent: Cancellation**'.

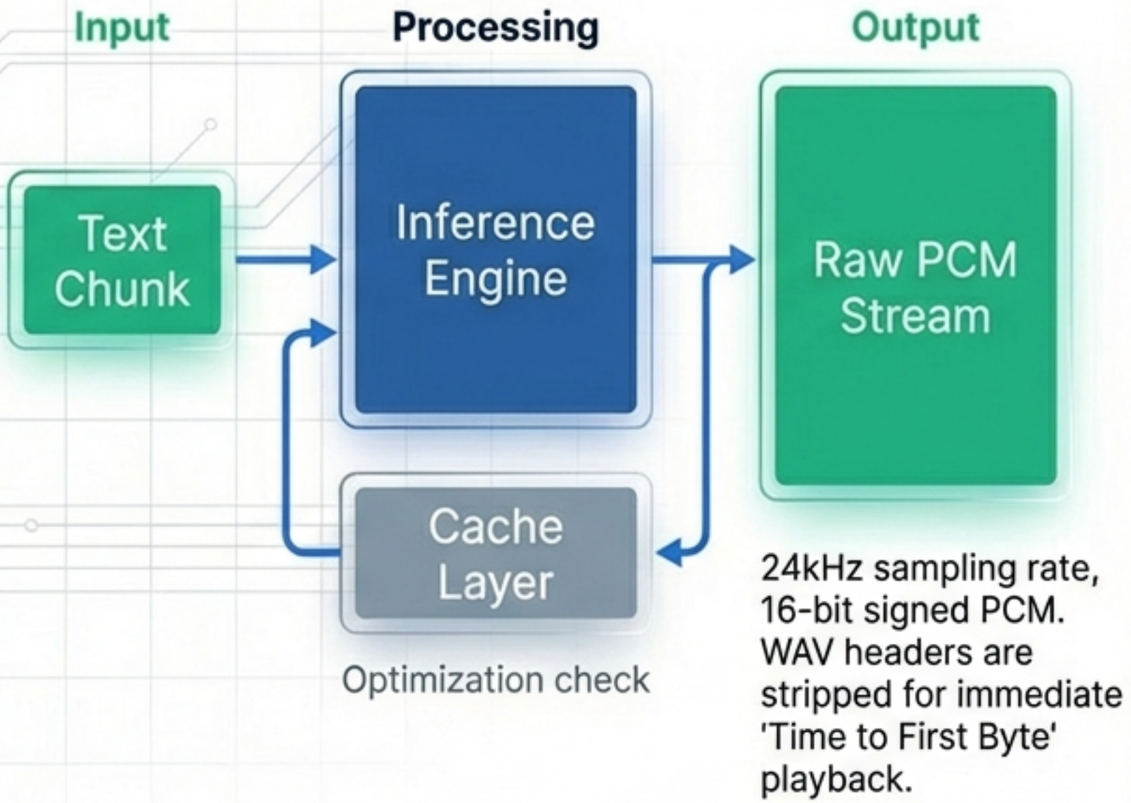
High-Fidelity Neural Synthesis (XTTS v2)

Model Specs

- **Model:** XTTS v2 (Coqui)
- **Framework:** FastAPI / Python
- **Hardware:** Optimized for L40S GPU



Streaming Architecture



Voice Cloning & Persona Library

The Speaker Roster

D (US)

Daisy (US)

A (FR)

Adde (FR)

S (AR)

Suad (AR)

K (JP)

Kazuhiko (JP)

Cross-lingual capabilities: A cloned American voice can speak fluent Japanese without retraining.

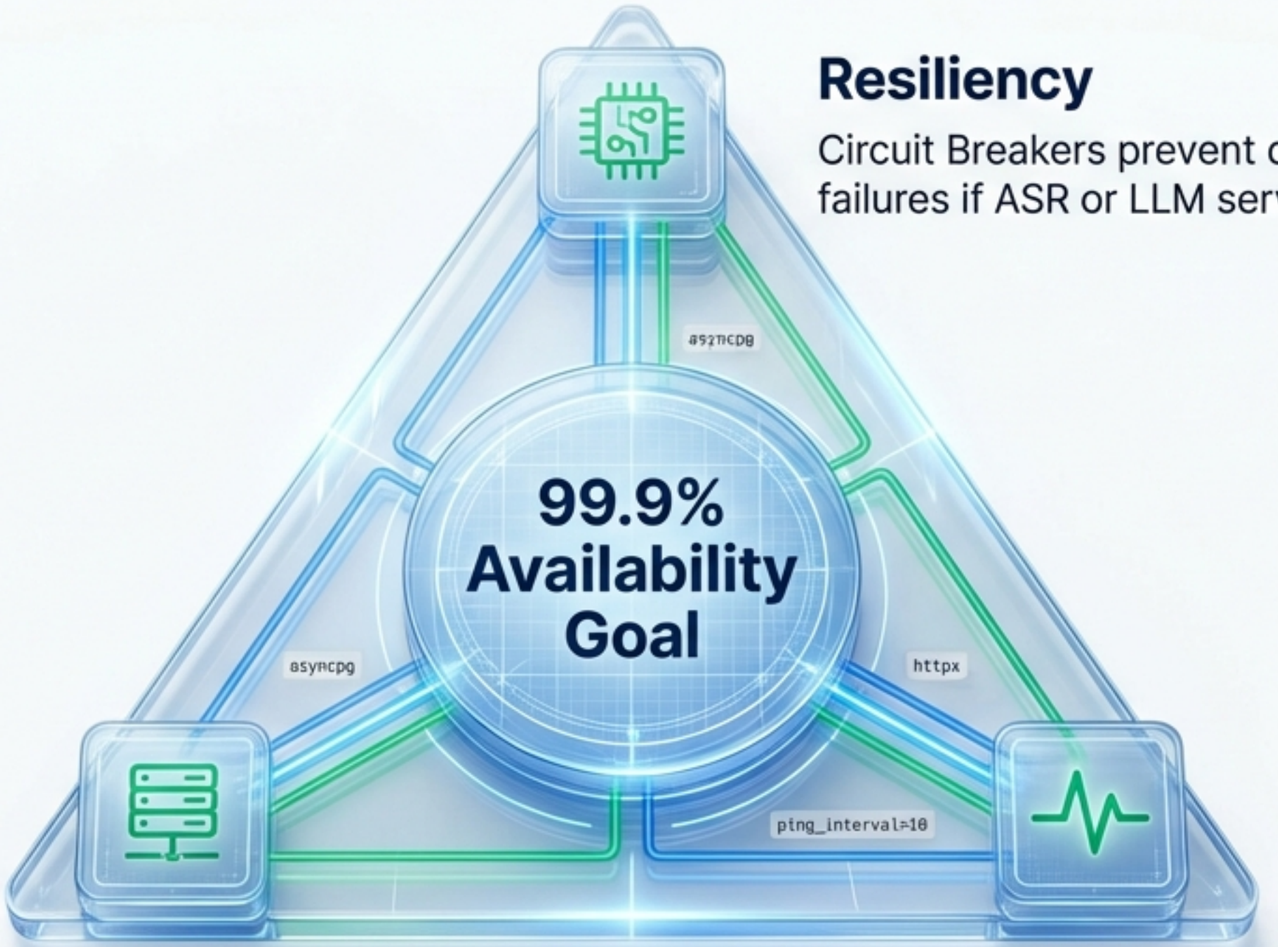
Cloning Technology

Reference Sample (<10s)

New Speaker Avatar

```
Endpoint: /speakers/upload
```

Engineering for Reliability & Scale



Resiliency

Circuit Breakers prevent cascading failures if ASR or LLM services stall.

Concurrency

Connection Pooling via `asyncpg` and `httpx` for high-volume handling.

Health

Automated Watchdogs monitor WebSocket health (`ping_interval=10`) and auto-reconnect.

Observability & Performance Metrics

Real-time latency reporting and persistence.

Total Round Trip Time (1.2s)


```
{
  "type": "latency_report",
  "data": {
    "asr_processing_time": 0.32,
    "first_llm_token_time": 0.58,
    "first_tts_chunk_time": 0.29
  }
}
```

Metrics are cached (`_metrics_cache`) and persisted to disk to track historical trends.

Use Case: The ‘Unhappy Customer’ Flow

Detection

Sarah (PIN 5678) complains about invoice.

Context Lookup

Query `invoices` table.
Result: ID #5 is “Overdue” (Red).

Resolution Offer

Agent offers to open a ticket.

Transaction

Write to `support_tickets`.
Log “Medium Risk” to `sentiment_alerts`.

Result

Customer reassured in native language.

The Evolution of Voice AI

From simple chatbots to context-aware, transactional, multilingual agents.

Powered by HPE Private Cloud AI