THE PILLAR MEN

AGENTIC AI & AUTONOMOUS SYSTEMS

TITLE - **NEXT-GEN CUSTOMER SUPPORT USING AGENTIC AI ARCHITECTURE**

Team Members- Lokesh Somaiya
Advait Daware
Kshitij Siya

PROBLEM STATEMENT

Traditional customer support bots and method

Context Blindness

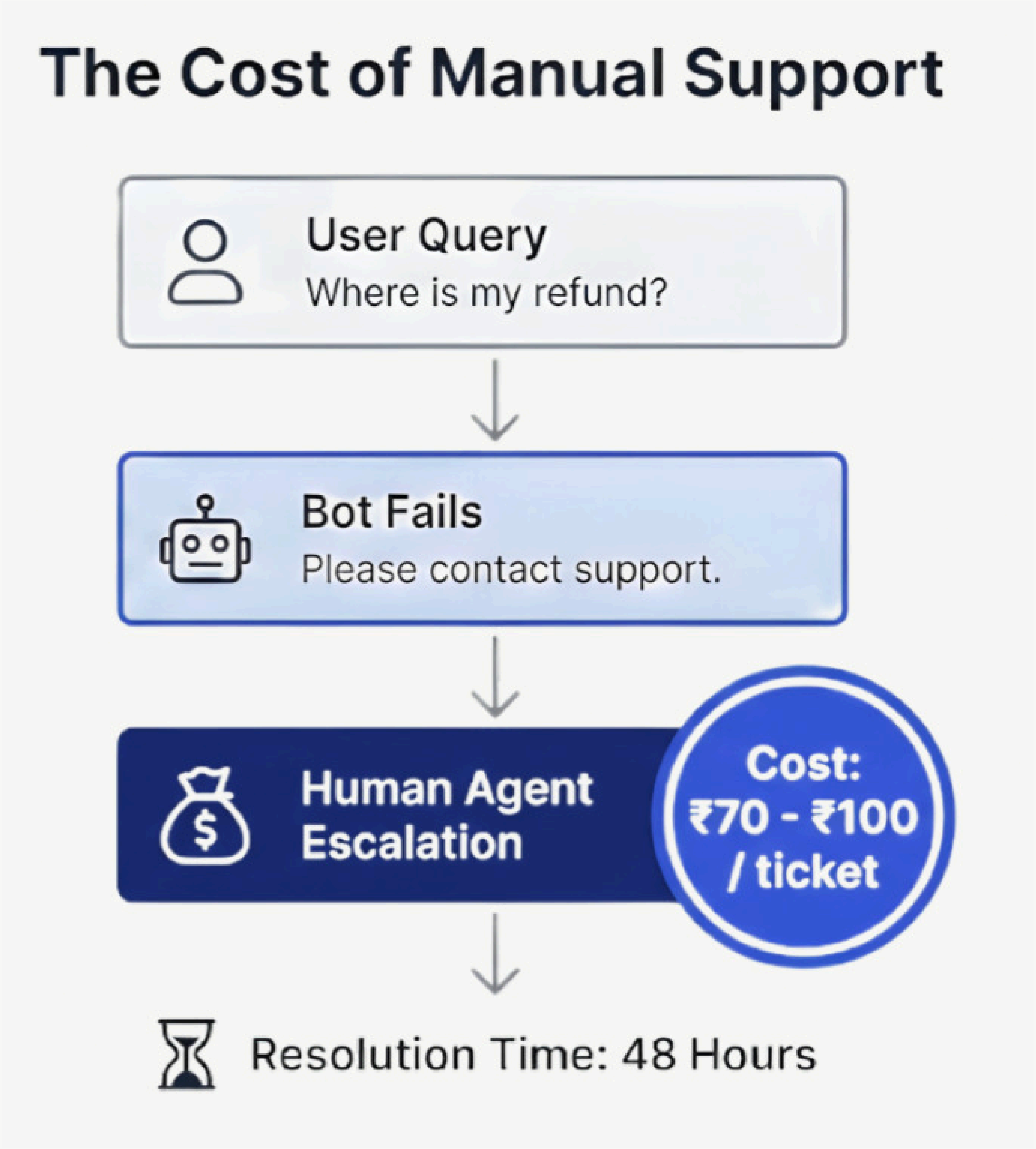
- Traditional chatbots rely on rigid scripts. They cannot recall past issues or check real-time order status.

The Action Gap

- Bots cannot perform physical tasks (refunds, replacements). They force customers to repeat themselves, leading to frustration.

The Escalation Trap

- 50% of tickets escalate to humans because the bot fails at simple logic.
- Result: A massive operational bottleneck with high churn risk.



The Cost of Manual Support

User Query
Where is my refund?

Bot Fails
Please contact support.

Human Agent Escalation

Cost: ₹70 - ₹100 / ticket

Resolution Time: 48 Hours

SOLUTION APPROACH

1. The "ReAct" Logic (Observe-Reason-Act)

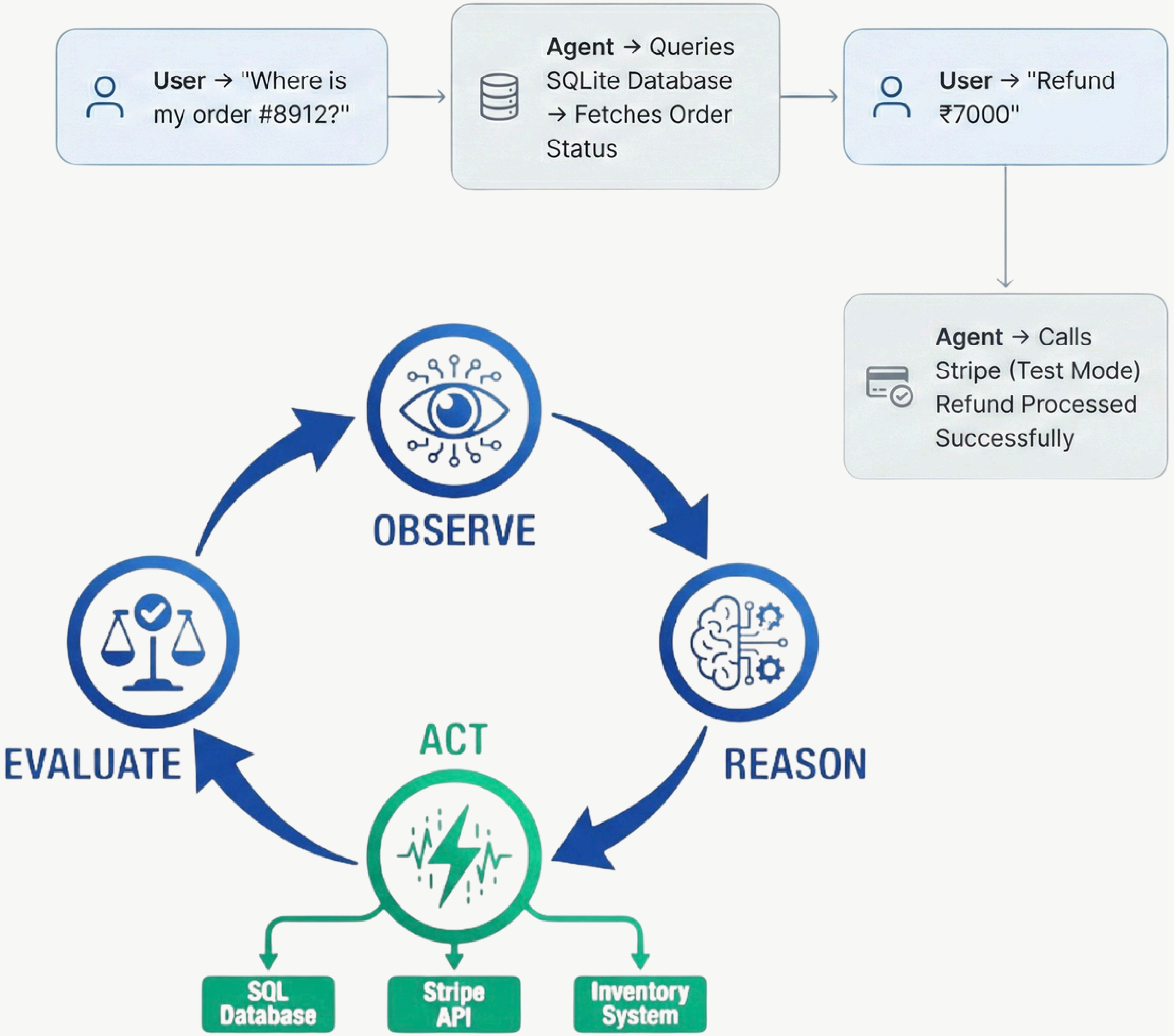
- Instead of blind replies, **Our agent** follows a strict logical path. LangChain ReAct + local LLM inference.
- Example: User asks for refund ➔ Agent checks SQLite (Order delivered?) ➔ Checks Policy (<₹10k?) ➔ Triggers Stripe API.

2. Intelligent Tooling

- Router Agent: Instantly classifies intent to pick the right tool.
- SQLite: Verifies real-time order status.
- Stripe (Test): Executes secure refunds.
- Safety: Auto-escalates to human if confidence < 60%.

3. Performance vs. Human Agents

- Resolution time: 48 hrs (today) → <60 seconds with our agent.
- Cost: ₹100 / ticket ➔ ₹0.50 / ticket
- Scale: Handles 10,000+ concurrent tickets instantly.



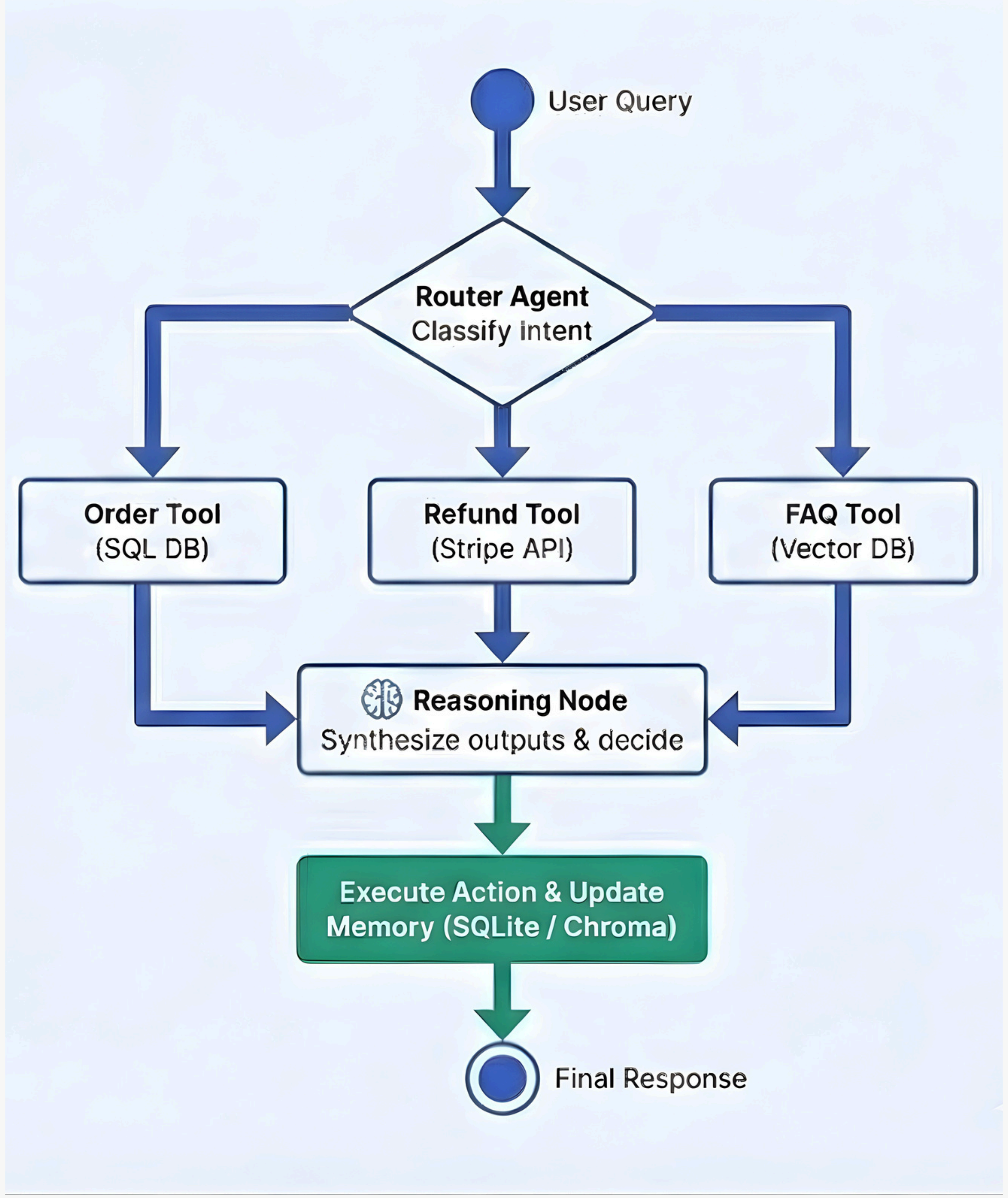
User → "Where is my order #8912?"

Agent → Queries SQLite Database → Fetches Order Status

User → "Refund ₹7000"

Agent → Calls Stripe (Test Mode) Refund Processed Successfully

TECHNICAL METHODOLOGY



1) **Data & Knowledge Layer**

- SQLite stores orders, customers, returns, and inventory.
- Chroma vector store powers semantic FAQ and policy retrieval.
- Synthetic edge cases for reliability and failure-mode testing.

2) **Agent Execution Pipeline**

- LangChain ReAct Agent coordinates reasoning + tool calling.
- Local LLM (via Ollama) handles intent, policy interpretation, and decision-making.
- Tools:
  - Stripe Tool → automated refund execution
- Structured responses optimized for customer clarity and accuracy.

3) **Backend & Channels**

- FastAPI backend integrates all tools & agent logic.
- Refunds triggered in Stripe test mode with webhook confirmation.
- Slack or Web UI channel for the live demo conversation.

4) **Safety & Reliability**

- ReAct ensures step-by-step verified reasoning.
- Policy checks → DB validation → action execution → confirmation.
- Fast, lightweight, low-cost due to fully local LLM + local DB + local Chroma.

TOOLS, MODELS & ARCHITECTURE

1. Core Intelligence (The Brain)

- LangChain ReAct Agent: Drives reasoning using step-by-step ReAct loops (think → act → observe).
- LLM Engine: LLaMA 3 fast, deterministic performance; integrates smoothly with FastAPI and LangChain tools.

2. Data & Memory Layer

SQLite (Operational Database)

- Stores structured data: orders, users, payments, inventory.
- Lightweight, portable, perfect for MVP and real-time tool queries.

Chroma Vector Store (Knowledge Memory)

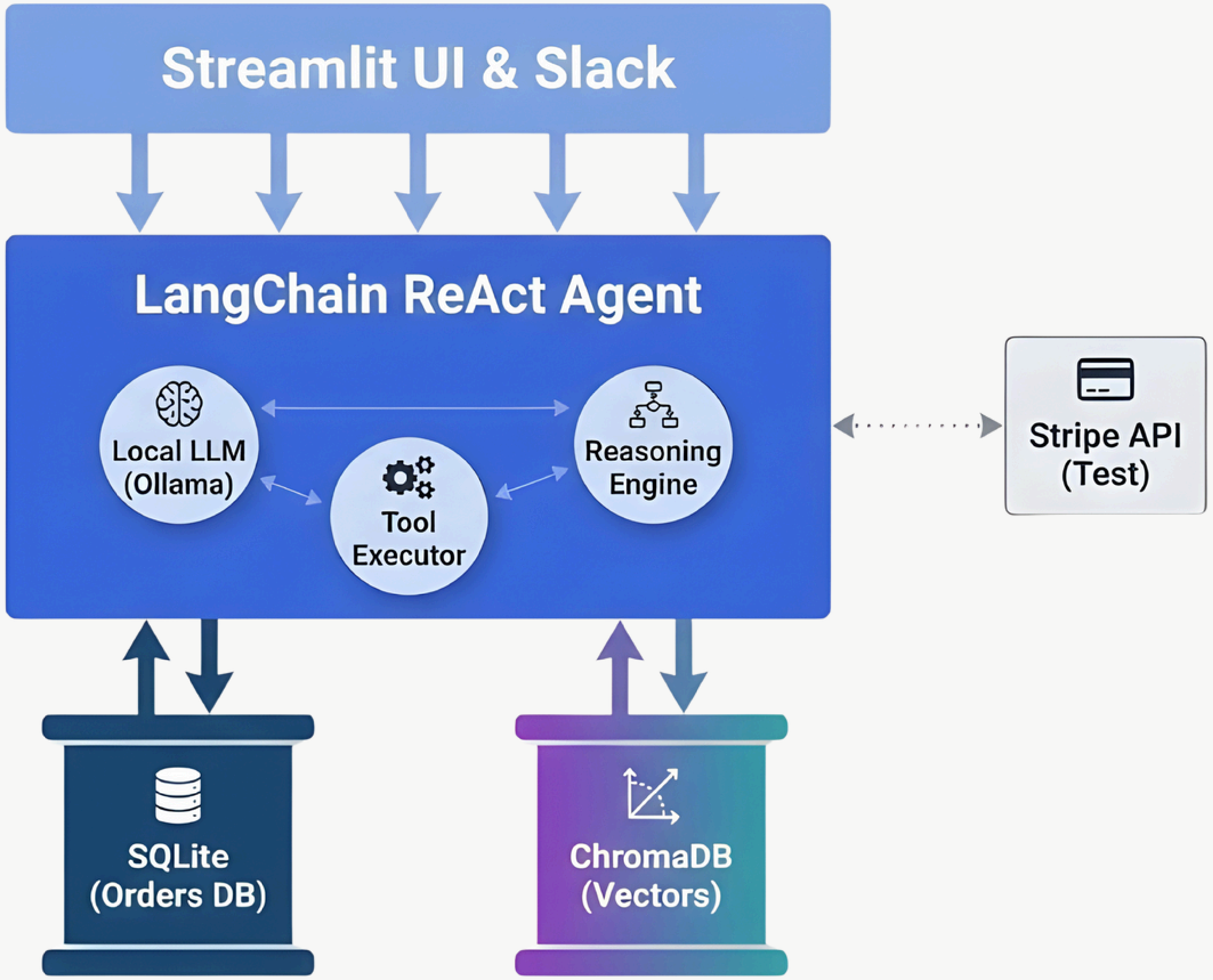
- Embedding-powered search over FAQs, policies, SLAs.
- Enables contextual, knowledge-grounded responses using local vector retrieval.

3. Action & Integration Ecosystem

- Stripe API: Handles secure, idempotent refund processing.
- Slack API: Provides a "Human-in-the-Loop" fallback channel for low-confidence queries.

4. Backend Infrastructure

- FastAPI: High-performance async Python framework serving the agent endpoints.
- Streamlit: Lightweight frontend for real-time chat simulation and demonstration.



EXPECTED IMPACTS

OPERATIONAL IMPACT

RESOLUTION TIME: 48 HRS → <60 SECONDS.
10× WORKLOAD REDUCTION FOR SUPPORT TEAM.
>95% ACTION ACCURACY WITH POLICY/TOOL VERIFICATION.
10–30% HIGHER FIRST-CONTACT RESOLUTION.

EXPERIENCE IMPACT

INSTANT REFUNDS + VERIFIABLE RECEIPTS.
INSTANT ORDER LOOKUPS & REPLACEMENTS.
PERSONALIZED RESPONSES VIA AGENT MEMORY.
24/7 SUPPORT WITH CONSISTENT QUALITY.

FINANCIAL IMPACT

GLOBAL CUSTOMER SUPPORT SPEND: $400B+ ANNUALLY
AI AUTOMATION GROWING AT 20–30% CAGR
SUPPORT COST DROPS FROM ₹5 CR/MONTH → ~₹50L/MONTH (90% REDUCTION)

TECHNICAL & SCALABILITY IMPACT

HORIZONTALLY SCALABLE: 1 → 10K TICKETS/DAY.
SAFE ACTIONS VIA TRUST-BUT-VERIFY ARCHITECTURE.
ROBUST: RETRIES, FALLBACKS, AUTO-ESCALATION.
MULTI-CHANNEL READY: SLACK, WHATSAPP, WEB WIDGET.

WHY OUR SOLUTION STAND OUT!!

- **Not a chatbot — a true autonomous agent**
- **Actually executes actions (refunds, DB lookups)**
- **Uses ReAct reasoning instead of generic LLM replies**
- **Has real guardrails (₹10k limit, confirmation steps)**
- **End-to-end ticket resolution under 60 seconds**
- **Local LLM → no cloud cost, privacy-friendly**
- **Business-ready architecture (FastAPI + tools + memory)**

GITHUB LINK

GITHUB