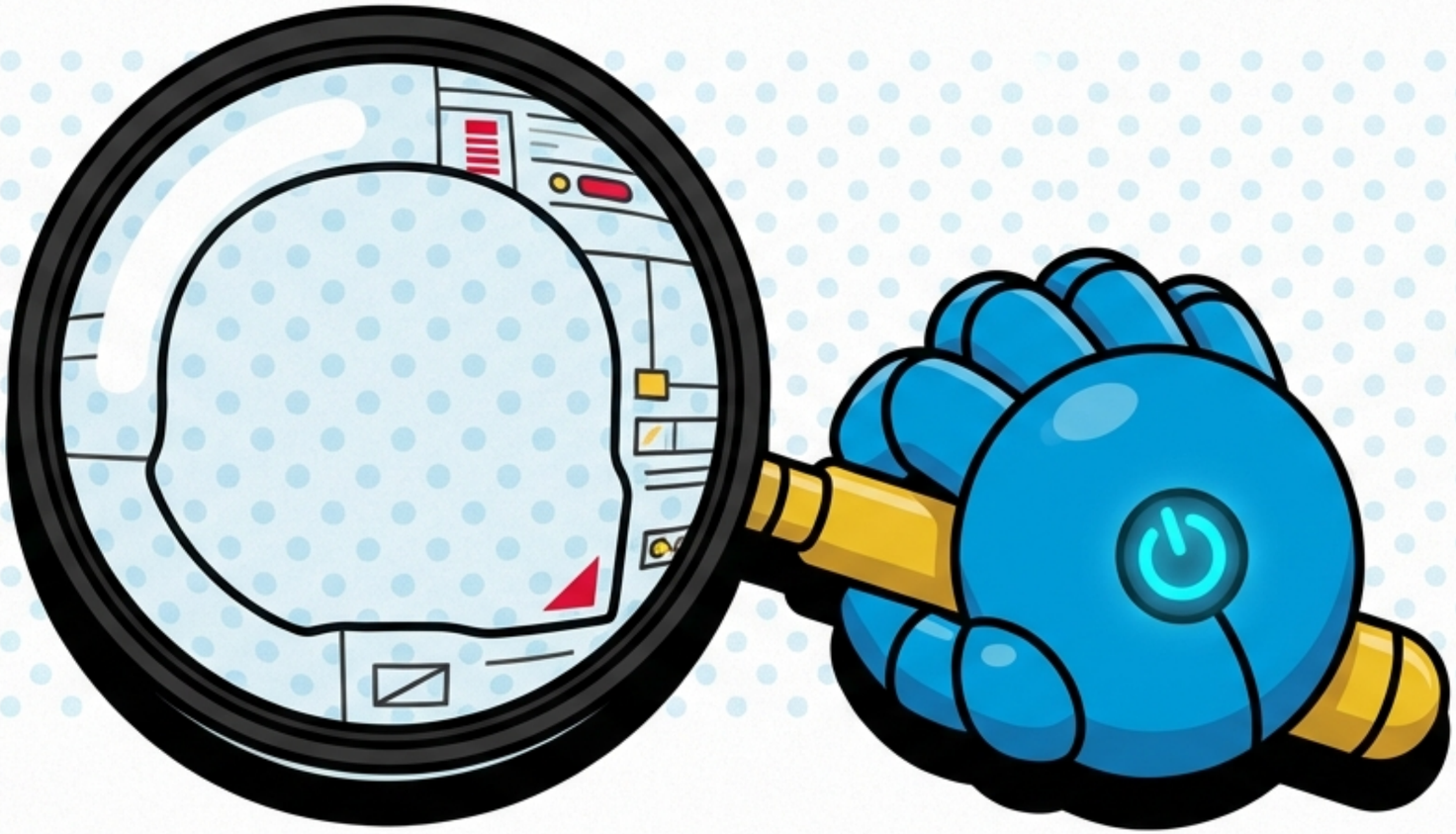
From Chatbot to Red Team

Introducing DevilAgent: The Dual-Mode Intelligent Auditor



v1.0.0 // OPEN SOURCE // PRODUCT ROADSHOW

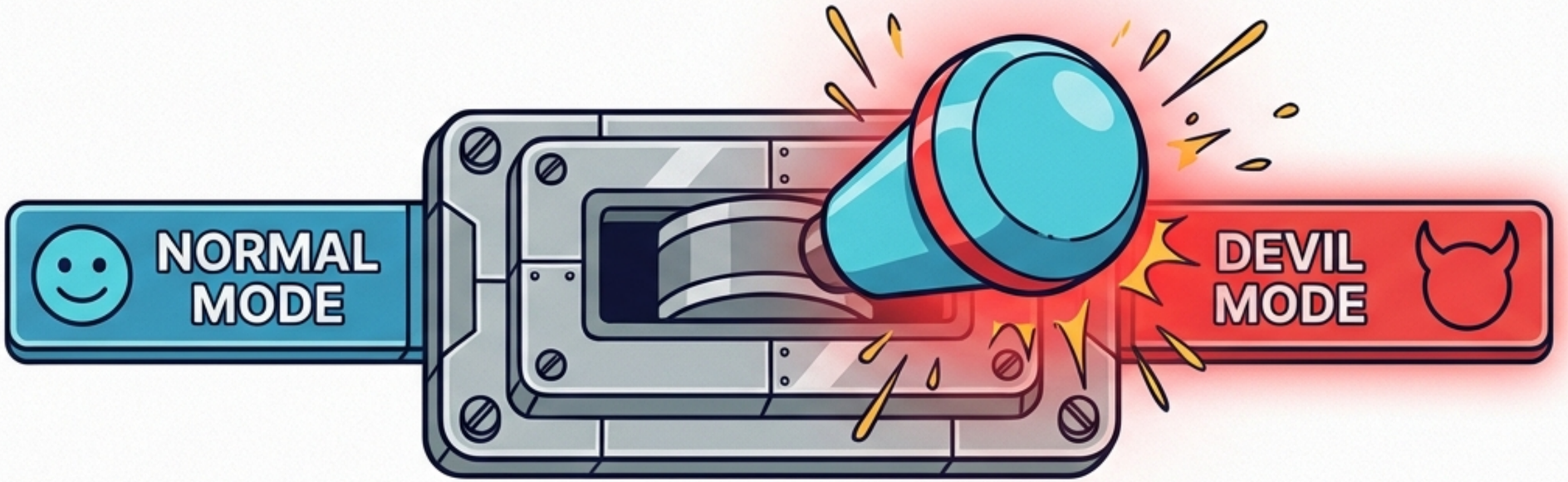
The Problem: Kill the “Yes-Man”

Standard LLMs suffer from sycophancy. They prioritize politeness over truth, leading to unverified hallucinations and shallow feedback loops.

Lack of Real-Time Verification

One Agent, Two Faces

Architecture-level toggle between Assistant and Auditor.



Normal Mode

- Polite & Conversational
- Standard Assistant behavior
- Helpful prompts

Devil Mode (Active)

- Ruthless ‘Red Team’ Persona
- Critical Logic Auditing
- Vulnerability Scanning

Mechanism 1: Truth Over Hallucination

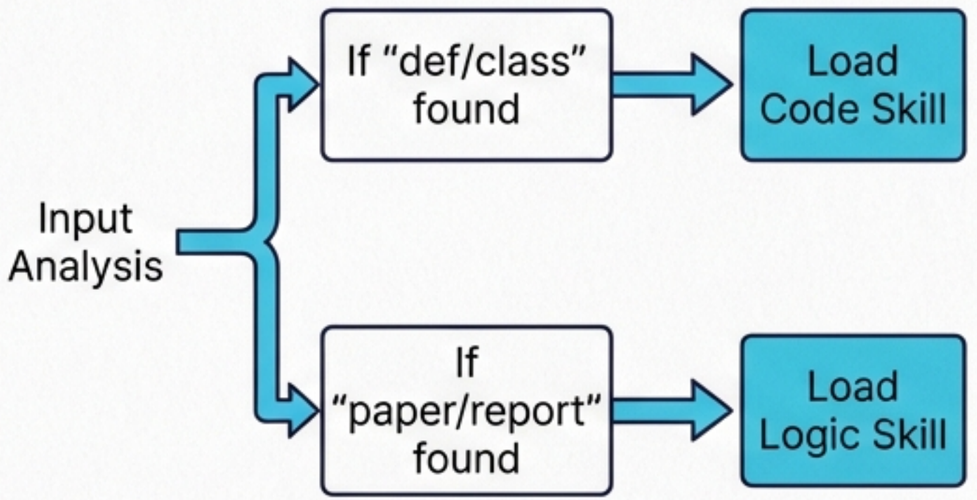
Integrated ReAct search workflow ensures answers are grounded in real-time data.

1. **Intent Extraction**
   Identifies search keywords (e.g., 'Tesla Stock Today').
2. **ReAct Execution**
   DuckDuckGo API fetches live results.
3. **Date Validation**
   SearchValidation model confirms data matches current date.
4. **Citation**
   Automatic footer links [1](url).

```python
class SearchValidation(BaseModel):
    is_satisfied: bool
    reason: str
```

Mechanism 2: The Right Tool for the Job

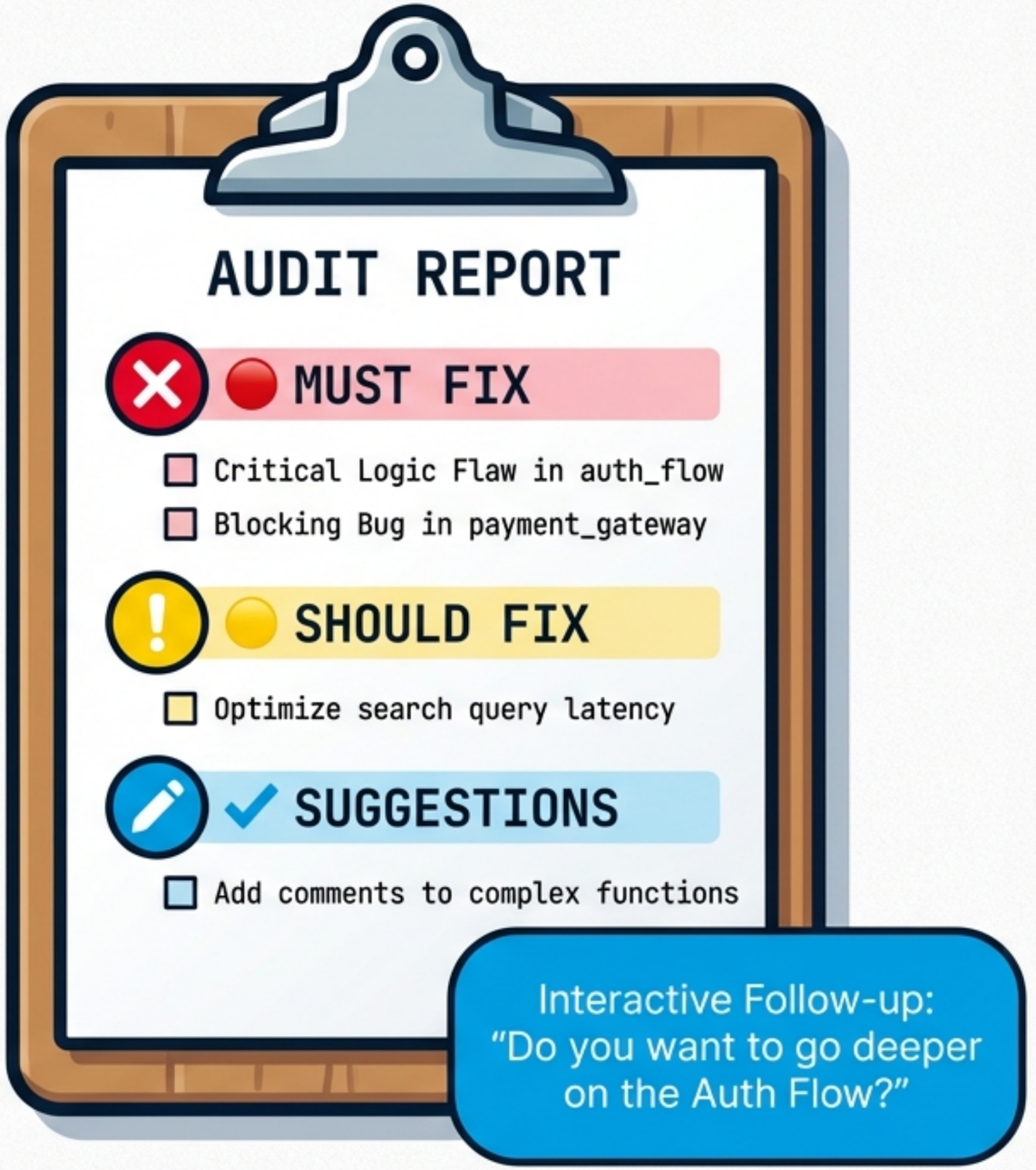
Context-aware skill routing dynamically loads instruction sets.



Extensible Architecture: Drop new .md files into /skills folder to expand capabilities instantly.

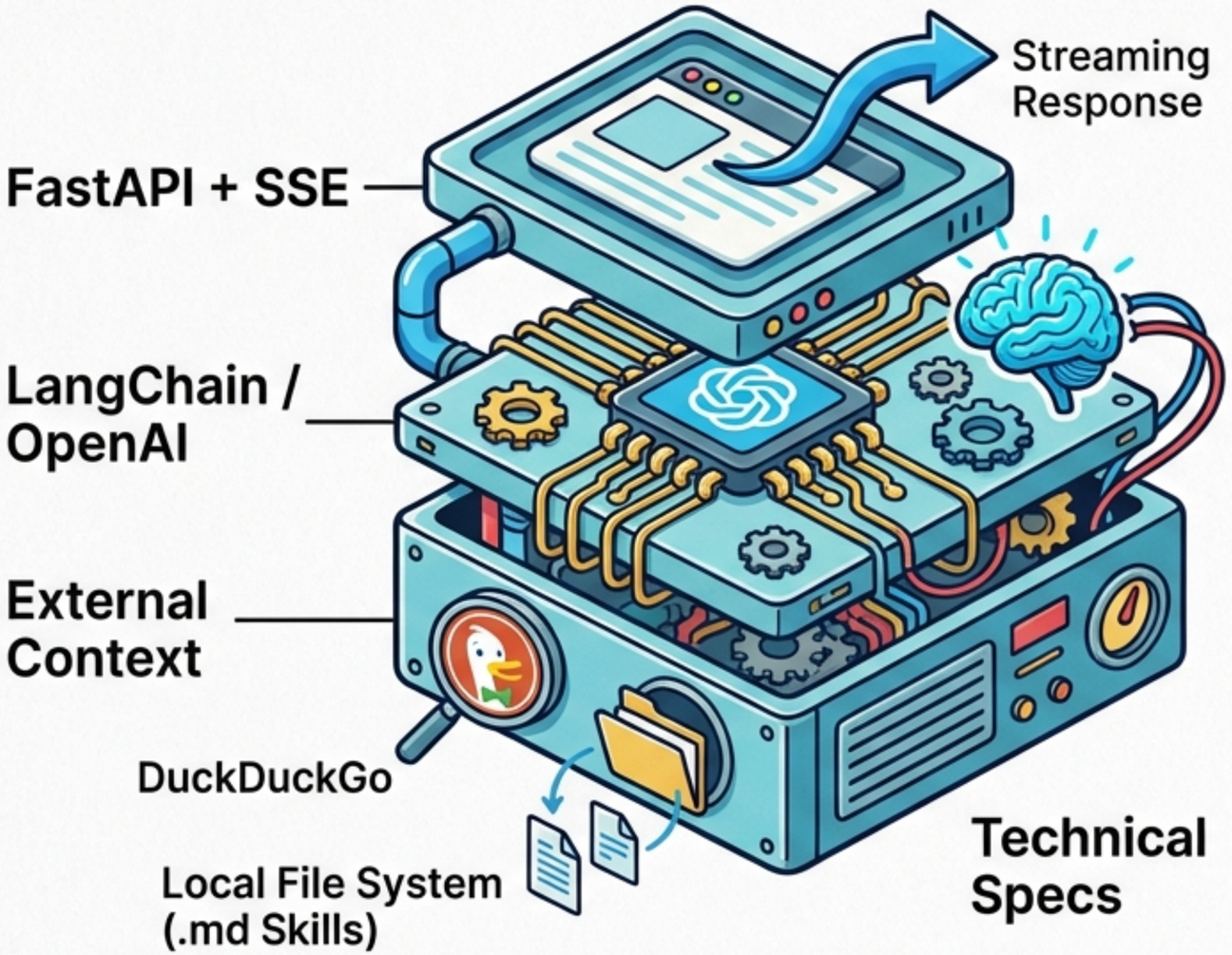
Ruthless Prioritization

DevilAgent doesn’t just chat. It grades. Output is structured by severity to force focus on critical issues.



Built to Scale

A modern, asynchronous Python stack.



- **Asynchronous:** FastAPI with Server-Sent Events
- **Structured:** Pydantic for validation
- **Configurable:** Simple .env setup, Docker-ready
- **Model Agnostic:** Compatible with GPT-4o-mini

Deploy the Devil

Stop asking for approval. Start asking for the truth.

```
$ git clone repo/devil-agent
$ pip install -r requirements.txt
$ python server.py
```

Ready for your pipeline. Open Source.