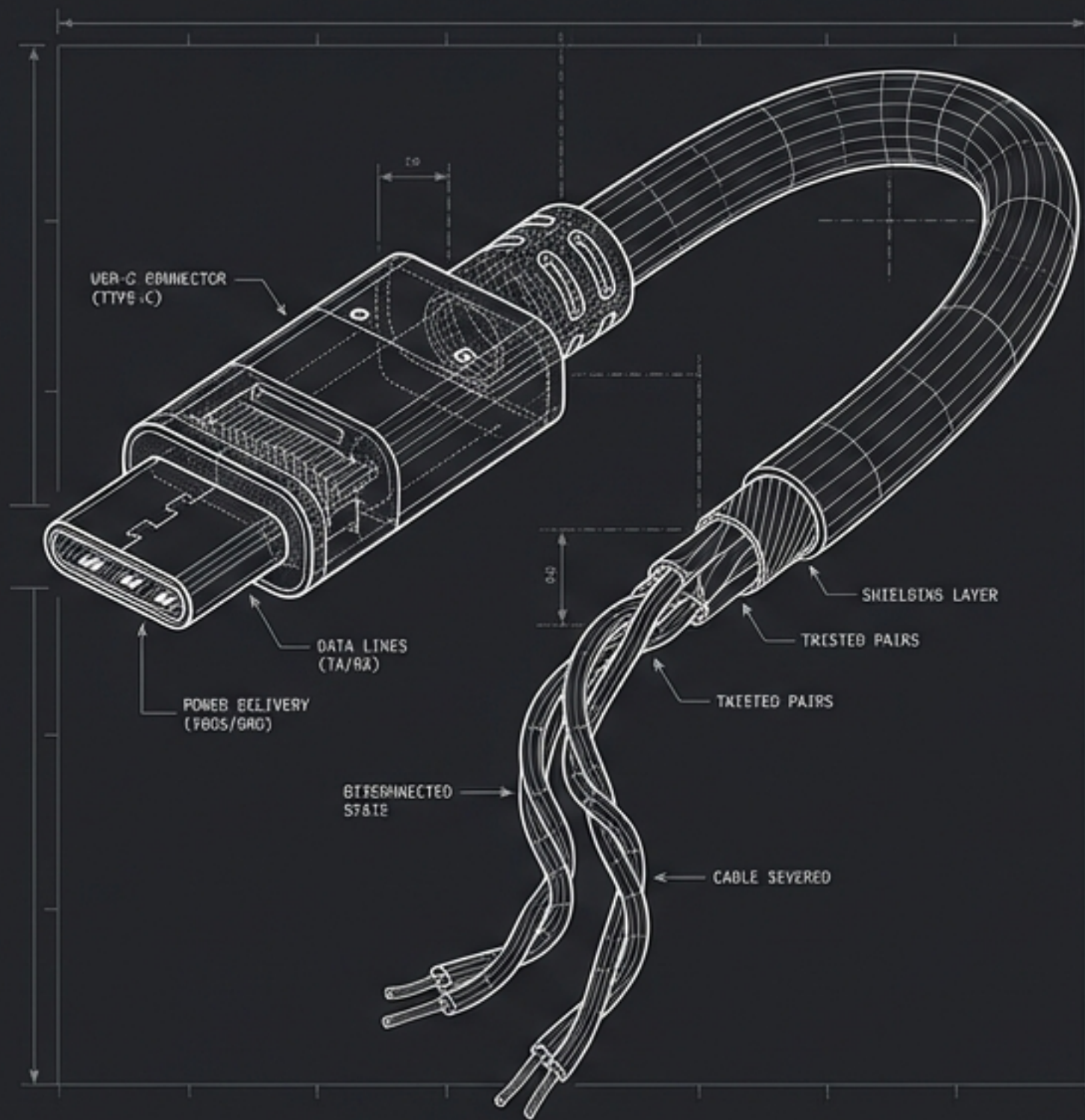




Offline AI-Assisted Development System

PRODUCT REQUIREMENTS DOCUMENT | VERSION 1.0 | JANUARY 2026

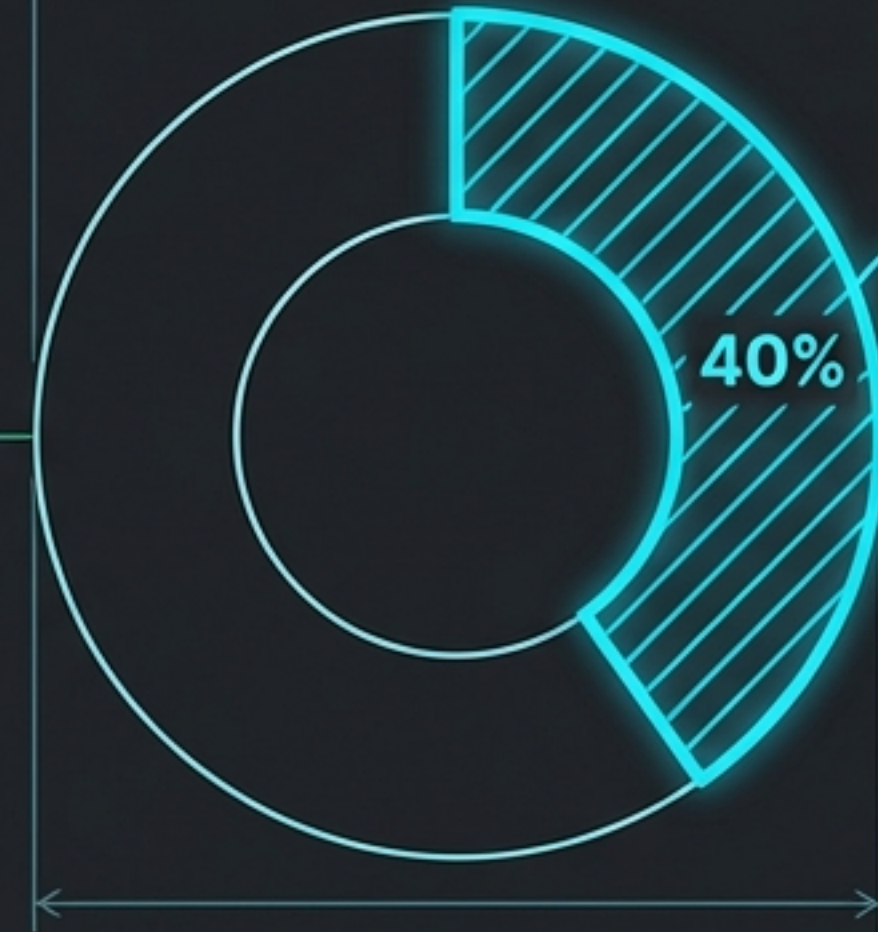
Seamless AI productivity, anywhere.

The Cloud Tether creates critical productivity gaps.

The Problem

Modern AI coding assistants like Claude Code and GitHub Copilot are tethered to the cloud. This creates dead zones in secure facilities, remote locations, and during travel.

Friction Point: Developers switching between online (Claude Code) and offline (Goose) tools currently suffer from context fragmentation and duplicated effort.



International flights account for ~40% of annual work time for traveling professionals.

**Goal: Zero internet dependency.
Feature parity with online tools.**

A comprehensive, sovereign development stack.

High-Level Solution Overview



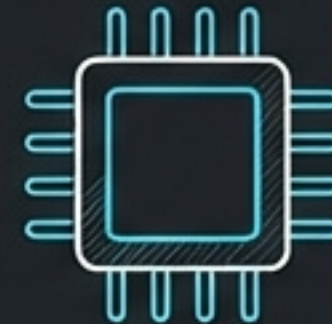
LocalCLI

Python wrapper replicating Claude Code UX. Handles approval workflows and output formatting.



Goose (goosed)

Rust-based agent execution engine using the Model Context Protocol (MCP).



Ollama

Local LLM inference server optimized for Apple Silicon (M-Series).

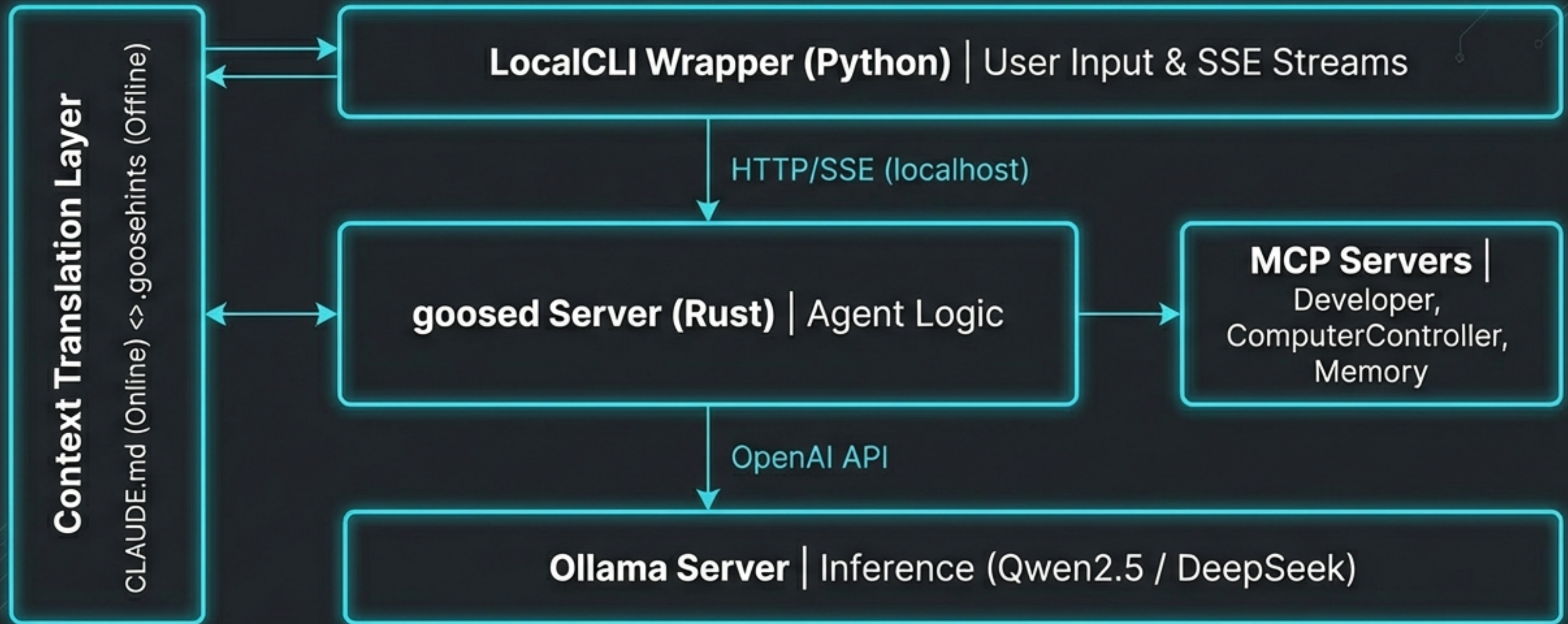


Context Translator

Bridging 'CLAUDE.md' and '.goosehints' for seamless portability.

Target Hardware: Validated for MacBook Pro M4 Max (128GB Unified Memory).

Zero-dependency system architecture.



All communication via localhost. No external data transmission.

Replicating the Claude Code experience.

LocalCLI is a Python 3.11+ wrapper designed to match the specific interaction patterns of Anthropic's online tool.

/compact	Summarize context to reduce token usage
/model	Switch between local models (Qwen vs DeepSeek)
/context	Display loaded files and token counts
/add <path>	Manage context files
/save	Session state management

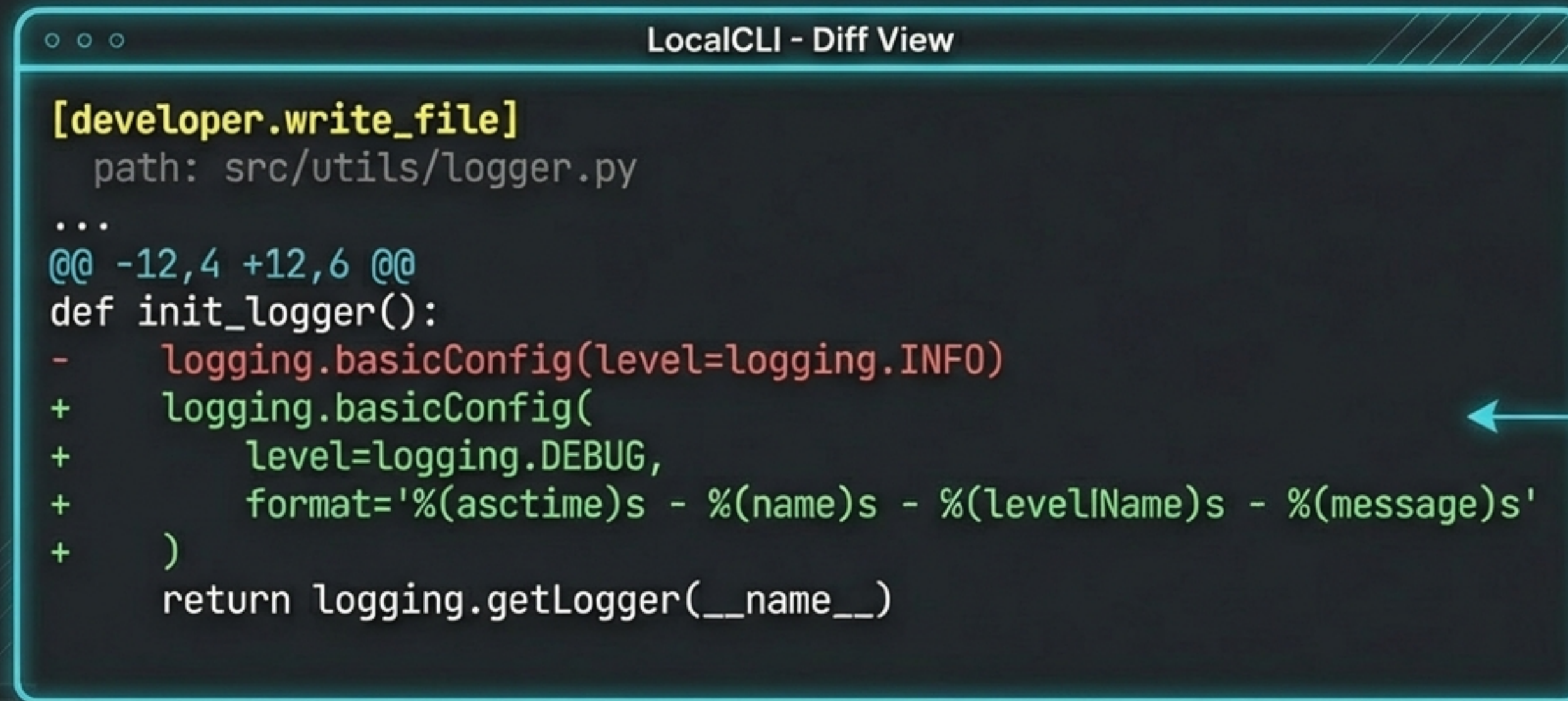
```
LocalCLI - v1.0

Prompt > /context

Loaded Context:
- src/main.rs (2.4k tokens)
- CLAUDE.md (1.1k tokens)
Total usage: 3.5k / 32k tokens
```

Clear, syntax-highlighted visual feedback.

Visualizing the Workflow: The Diff View



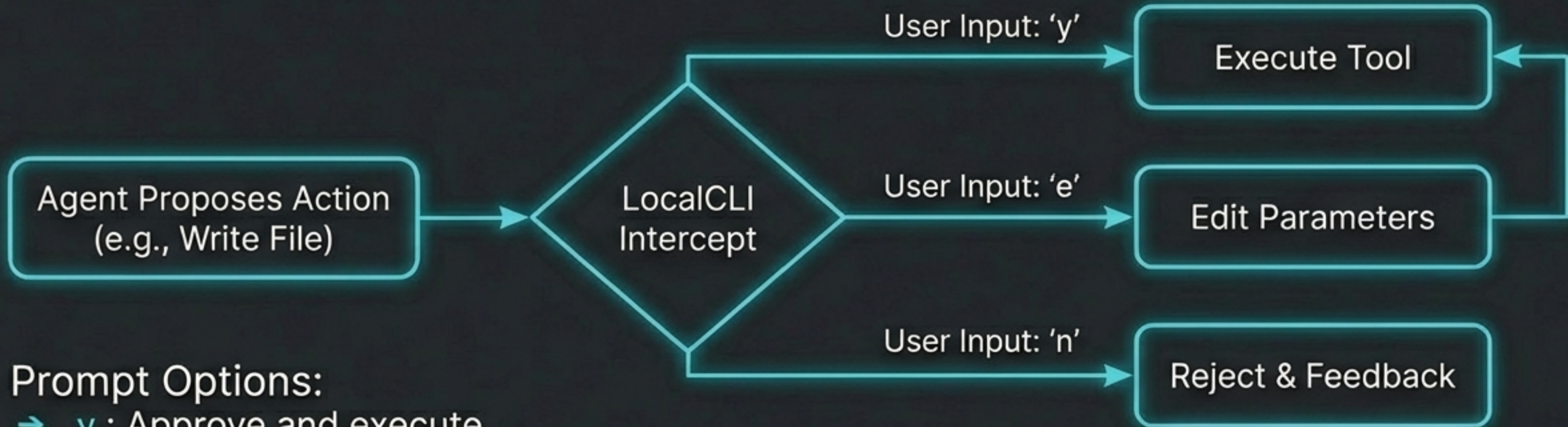
```
LocalCLI - Diff View

[developer.write_file]
path: src/utils/logger.py
...
@@ -12,4 +12,6 @@
def init_logger():
-     logging.basicConfig(level=logging.INFO)
+     logging.basicConfig(
+         level=logging.DEBUG,
+         format='%(asctime)s - %(name)s - %(levelname)s - %(message)s'
+     )
return logging.getLogger(__name__)
```

Unified Diff format
with clear tool
boundaries.

Tool Approval & Safety Protocols.

Human-in-the-loop Security.

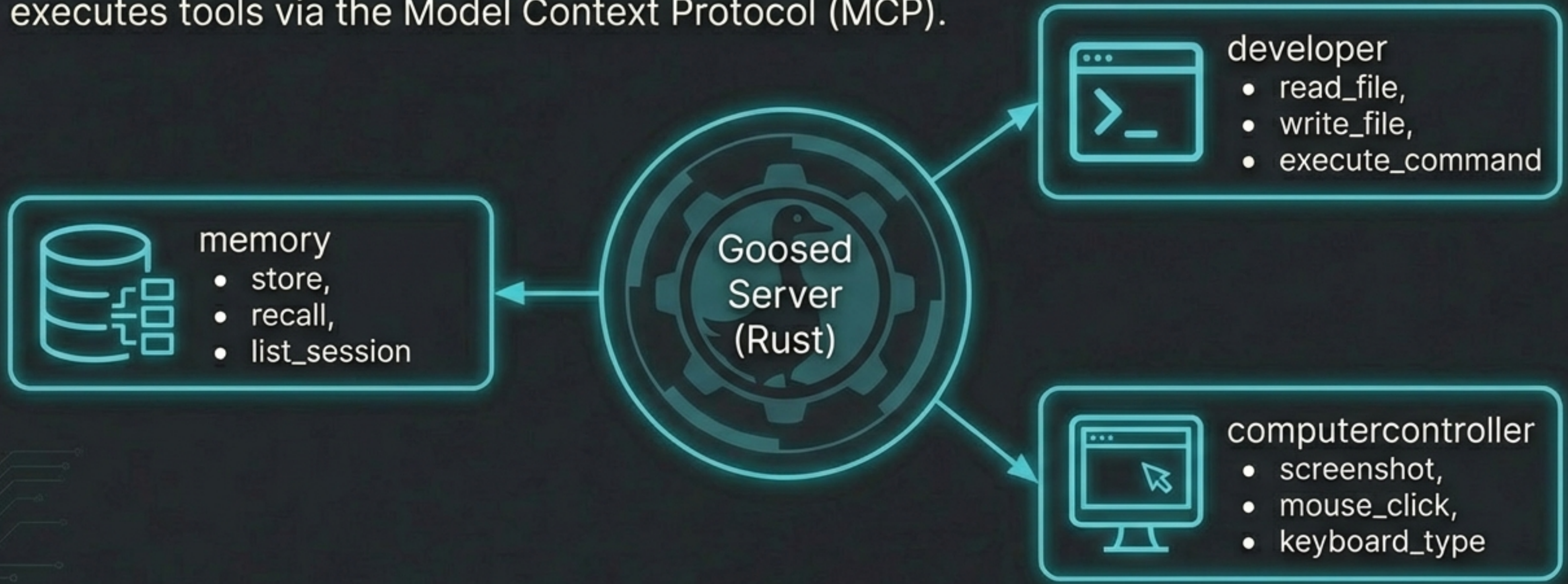


Prompt Options:

- y : Approve and execute
- n : Reject and provide feedback
- e : Edit parameters
- a : Always approve this action type
- ? : Display details

The Engine: Goosed & MCP Extensions.

The Rust-based 'goosed' server drives logic and executes tools via the Model Context Protocol (MCP).



Hardware Requirements & Model Tiers.

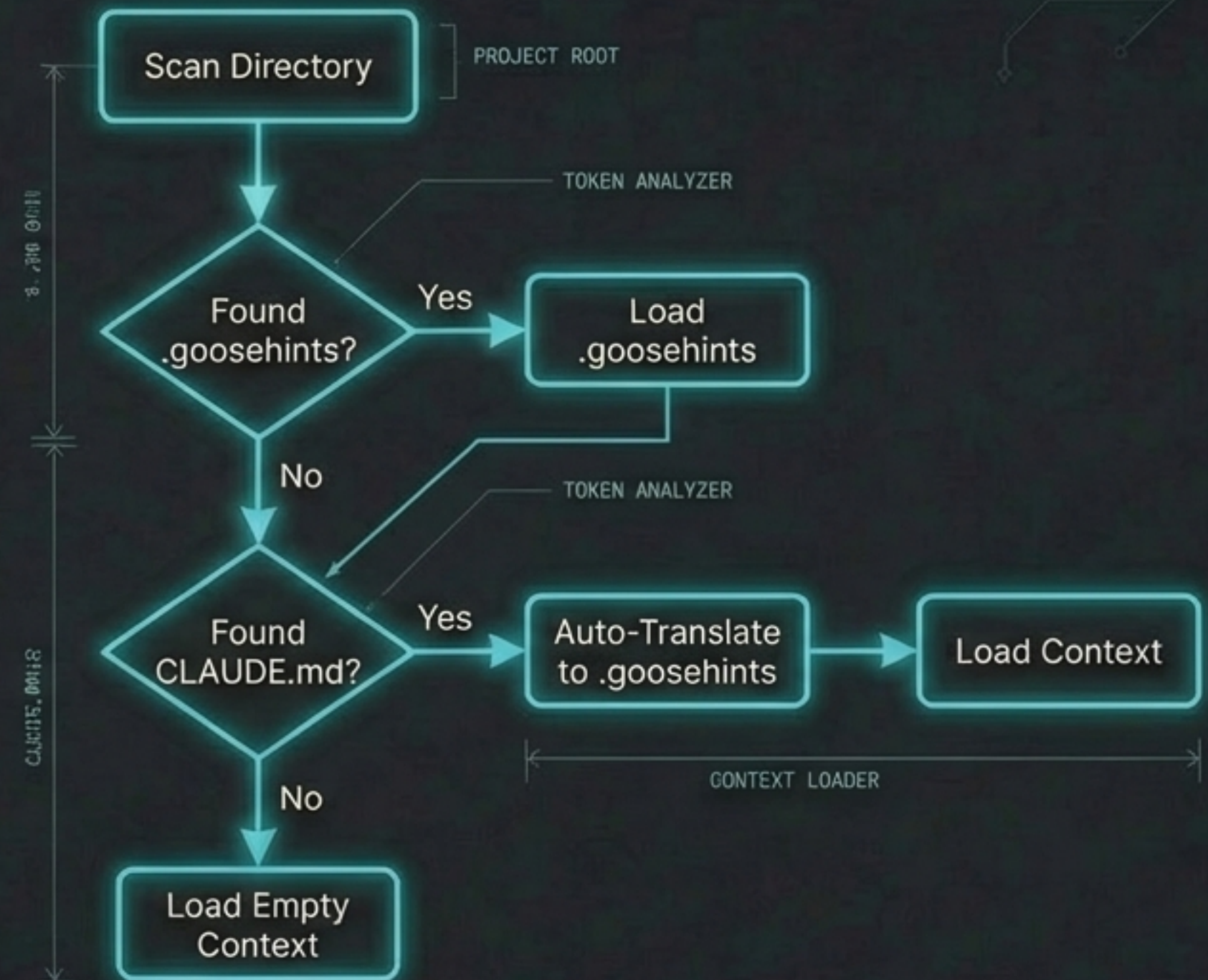
Primary Target: MacBook Pro M4 Max
Spec: 128GB Unified Memory | 100GB Storage

Tier	Model	VRAM Req	Speed
Best Quality	Qwen2.5-Coder 72B (Q4)	45GB	~15 tok/s
Balanced	DeepSeek-Coder-V2 33B	20GB	~30 tok/s
Constrained	CodeLlama 13B	8GB	~60 tok/s

Intelligent Context Management.

Auto-Detection: Identifies project roots via `.git`, `package.json`, or `Cargo.toml`.

Token Budgeting: Monitors the 32k context window limit.



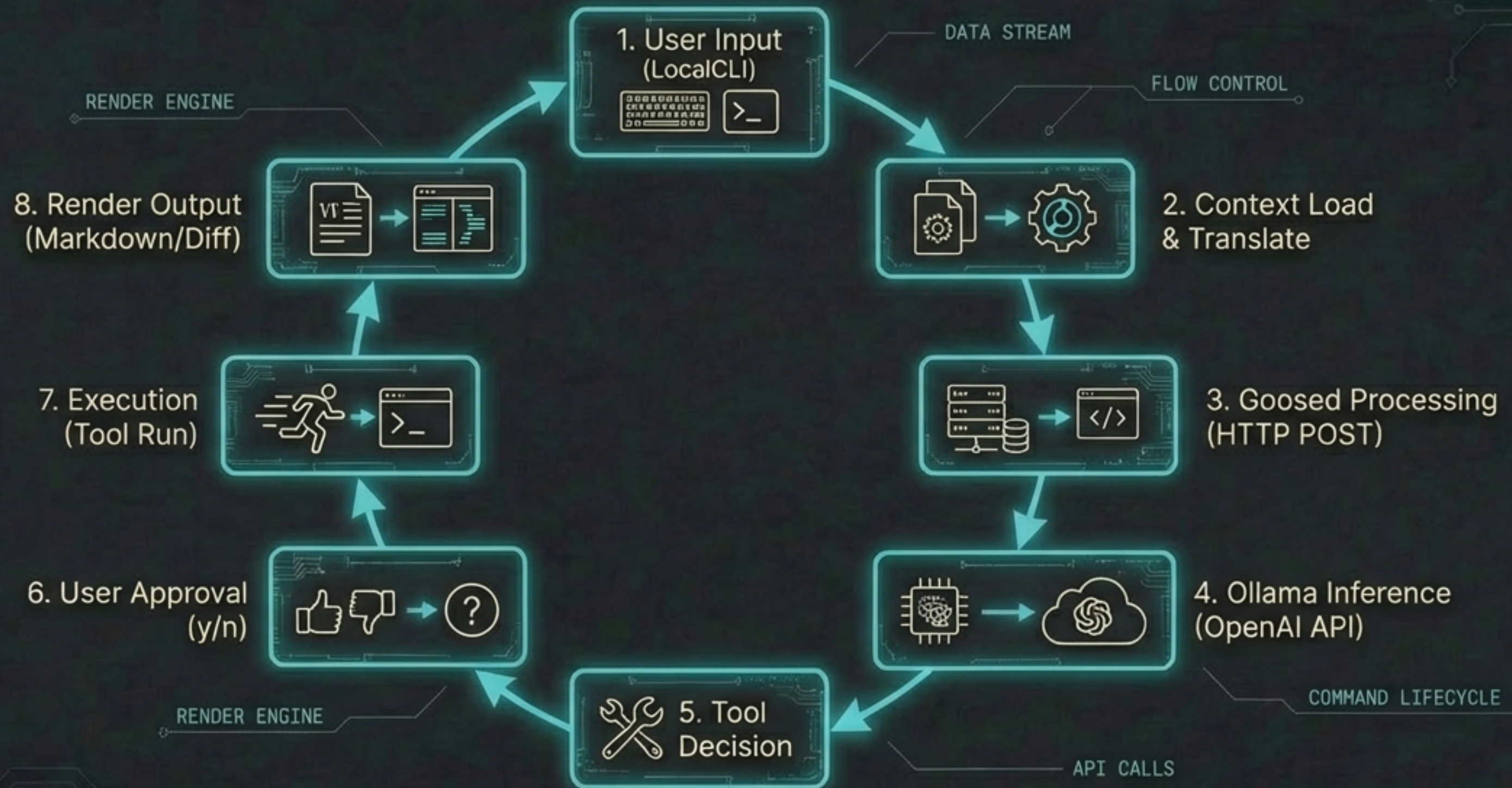
The Context Translator.

Bidirectional sync between Online and Offline environments.



Sync Trigger: Pre-commit Git Hook based on file timestamp.

Data Flow & Interaction Loop.



Configuration & Customization.

~/.config/localcli/config.toml

Model selection
per profile

```
[goosed]
url = "http://localhost:3000"
timeout = 120.0

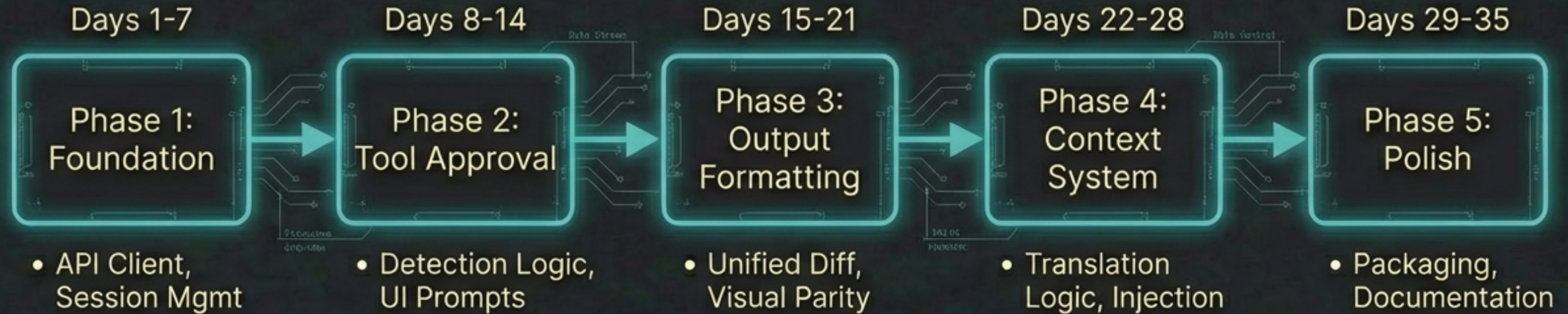
[ollama]
model = "qwen2.5-coder:72b-instruct-q4_K_M"

[ui]
prompt = ">"
show_cost = true

[context]
auto_translate = true
inject_defaults = true
```

Toggle auto-
translation logic

Implementation Roadmap (35 Days).



Pre-Flight Checklist.

Verifying readiness for offline operations.

- ✓ Verify Models (`ollama list`) - Qwen2.5 / DeepSeek available
- ✓ Verify Goose (`goose --version`) – v1.0 Installed
- ✓ Check Config - `profiles.yaml` validated
- ✓ Disk Space - 100GB+ available
- ✓ Test Stack - `curl localhost:3000` OK

STATUS: READY FOR OFFLINE DEVELOPMENT.