

The Execution Engine

Not one model. A system of specialized on-device models, skills, and tools that finish real work while minimizing resources.



MODELS

A system, not one model

- Cephable Gemma 4/Qwen fine-tunes: S / M / L / XL / XXL (XL/XXL are MoE)
- Medium (~4B) is the default; auto-selected to the device, user-adjustable
- 4-bit quantized; ~64 silicon-optimized builds
- Separate models for reasoning, tool filtering, summarizing, generation



TOOL SELECTION

Dynamic tool selection

- A local non-LLM model ranks every tool per request
- Sorts into: definitely need / might need / won't need
- Loads only what's needed, so context and memory stay lean
- Applies to MCP tools and built-in tools alike



TOOLS

What tools can do

- Read, find, create, edit files
- Create any text based files, code, scripts, markdown, etc
- Search the CRM, generate Canva designs, work in Outlook
- Web research via embedded browser; run terminal / CLI commands
- Computer use via the OS accessibility tree, not screenshots or recording



SKILLS

Skills: reusable knowledge

- Build from scratch, import Anthropic/OpenAI-compatible skills, or generate from a run
- Bundle markdown instructions, references, files, scripts, templates
- Favorite a workflow and convert it into a shareable Skill
- Roadmap: learn a skill by watching you perform it



WORKFLOWS

Automate workflows

- Multi-step chains: research, draft, generate docs, pull CRM, create designs
- Run in the background while you keep working
- Pause for approvals, resume, branch, and track progress
- Export results to JSON-L; promote any run into a Skill



ON-DEVICE

Built to finish the job

- Fine-tuned for task completion, not just generation
- Pre- and post-processing at each step keeps work on track
- Human-in-the-loop before destructive or non-reversible actions