

ProductTeam v2.6.0

Architecture & User Guide

Structured AI Software Delivery Pipeline

Free local AI or fast cloud APIs
You choose. The wizard handles the rest.

A LOCAL AI

Free | Ollama | ~20 min/step

B CLOUD AI

Deeper & faster | API Key

Scott Converse

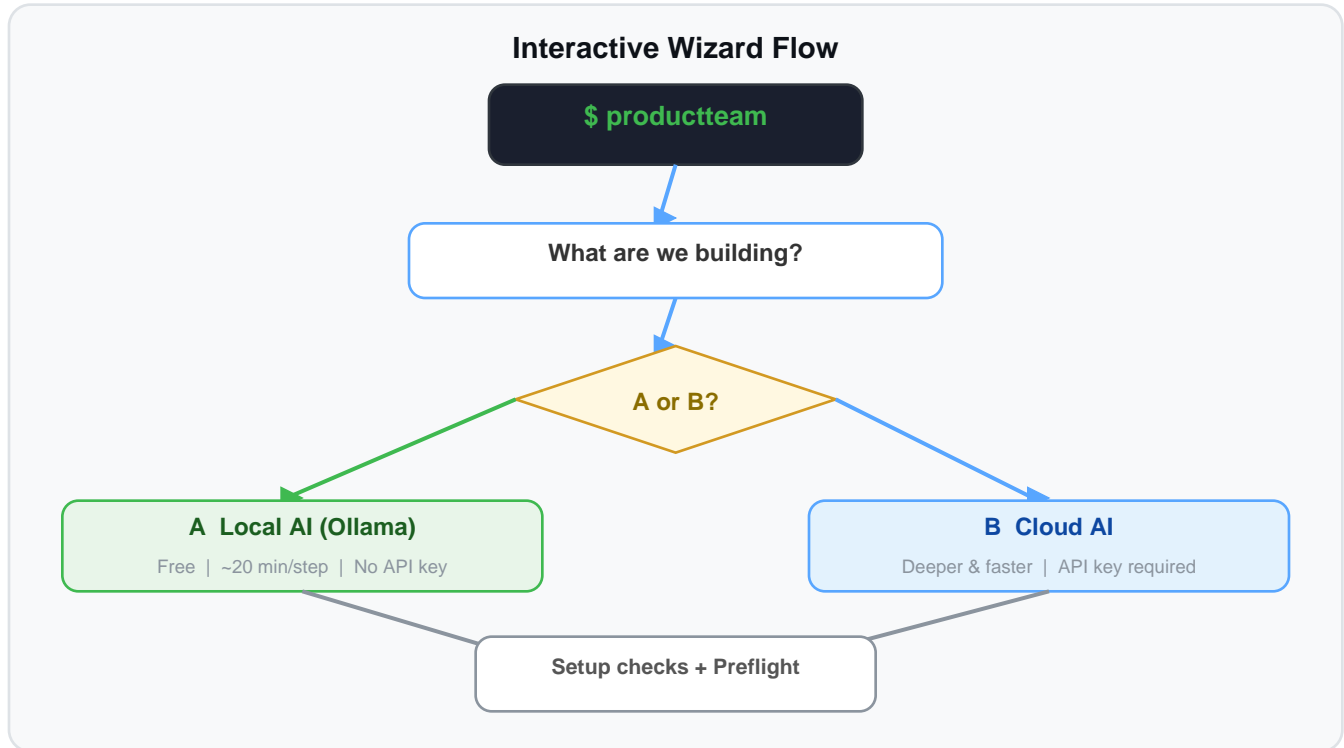
March 2026

Table of Contents

1. Two Ways to Run
2. Pipeline Architecture
3. Agent Roles
4. Local AI Setup
5. Cloud AI Setup
6. Cost Comparison
7. CLI Reference
8. Safety & Recovery

1. Two Ways to Run

ProductTeam offers two paths to run the pipeline. Just type **productteam** (no arguments) and the interactive wizard walks you through setup.



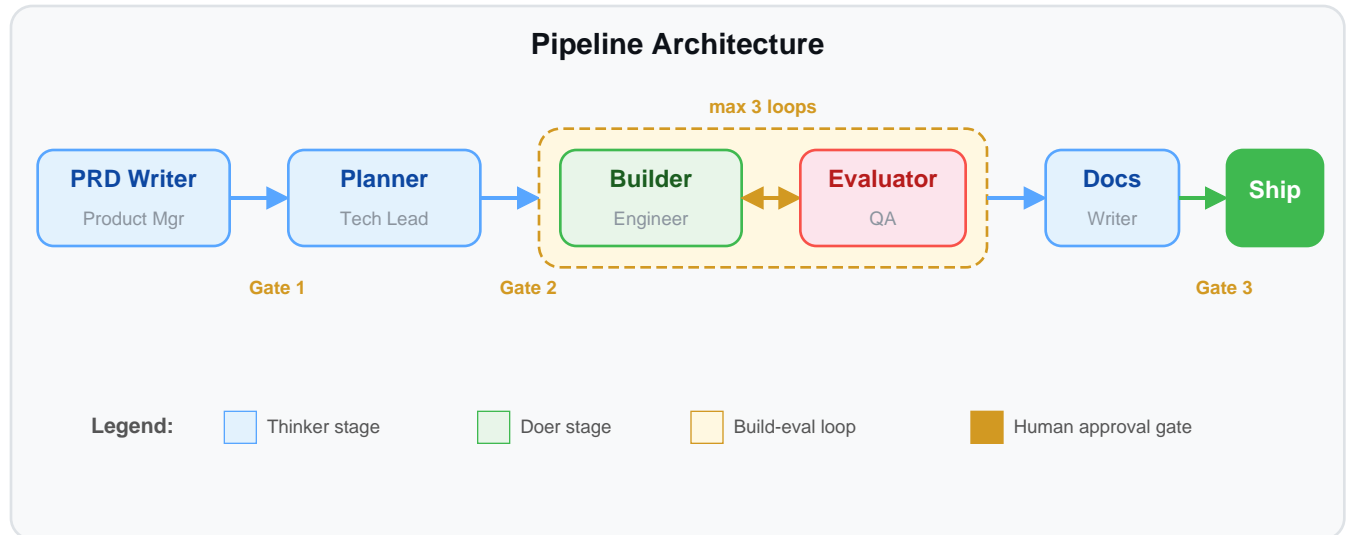
Comparison

	Local AI (Ollama)	Cloud AI
Cost	Free	Standard API costs
Speed	~20 min/step	Faster with cloud models
Setup	Install Ollama + pull model	Paste API key
Recommended model	gpt-oss:20b (13 GB)	Claude Sonnet / GPT-4o
Internet required	No (after model download)	Yes
Privacy	Everything stays on your machine	Prompts sent to provider API

Note: Local models are free but slower. Each pipeline step takes ~20 minutes on a 20B parameter model running on consumer hardware with 32K context, so a full project takes hours. Cloud APIs are significantly faster.

2. Pipeline Architecture

The pipeline transforms a product concept into working code through seven specialized agents. Three human approval gates let you confirm intent, scope, and readiness. The builder never grades its own work -- a separate, skeptical evaluator does.



Thinker stages (PRD Writer, Design Evaluator) take context in and produce a text artifact out. One LLM call. No filesystem access.

Doer stages (Planner, Builder, Evaluator, Doc Writer) use an agentic tool-use loop with exactly four tools: `read_file`, `write_file`, `run_bash`, `list_dir`.

The **Build-Evaluate loop** runs up to 3 iterations. If the Evaluator grades `NEEDS_WORK`, findings route back to the Builder automatically. After loop 3, the plan is wrong -- not the implementation.

3. Agent Roles

Each agent is a standalone markdown skill file. Readable, editable, replaceable. Use the full pipeline or drop in only the skills you need.

Agent	Role	Description
prd-writer	Product Manager	Converts concept to structured PRD with requirements, constraints, and success criteria.
planner	Tech Lead	Decomposes PRD into sprint contracts with testable acceptance criteria. Writes sprint YAML.
builder	Engineer	Implements sprint contracts with production-quality code and tests. Declares 'ready for review.'
ui-builder	Frontend Engineer	Specialized builder for visual work. Landing pages, dashboards, web UIs.
evaluator	QA Engineer	Skeptical by default. Reads source, runs tests, verifies acceptance criteria. PASS / NEEDS_WORK
evaluator-design	Design Reviewer	Grades visual artifacts on Coherence, Originality, Craft, Functionality. 1-5 scale, 4.0+ to pass.
doc-writer	Technical Writer	Reads every source file. Produces README, changelog with real data only. Never fabricates.
orchestrator	Project Manager	Routes work between agents, manages loops (max 3), handles approval gates.

4. Local AI Setup

ProductTeam runs entirely free using Ollama, a local AI runtime. No API key, no cloud dependency, no per-token costs.

Installation

1. Download Ollama from <https://ollama.com/download>
2. Run the installer
3. Pull the recommended model:

```
ollama pull gpt-oss:20b
```

Recommended Models

Model	Size	Role	Notes
gpt-oss:20b	13 GB	Primary	Best tool-calling reliability and speed. OpenAI open-weight.
devstral:24b	14 GB	Backup	Mistral coding agent. Strong code generation.

Preflight Check

Before committing to a long pipeline run, verify your model works:

```
productteam preflight
```

Preflight runs three quick tests: basic response, tool calling, and multi-turn tool use. Takes about 30-60 seconds. If all three pass, the model is pipeline-ready.

Auto-Tuning

When you choose Local AI, ProductTeam automatically:

- Sets stage timeouts to 60 minutes (vs 5-10 min for cloud)
- Disables design review (saves time)
- Sets all approval gates to auto-approve
- Recommends 32K context window in Ollama settings

5. Cloud AI Setup

For deeper and faster pipeline runs, use a cloud API provider.

Supported Providers

Provider	Default Model	Cost	Context Window
Anthropic	Claude Sonnet 4	Standard API costs	200K tokens
OpenAI	GPT-4o	Standard API costs	128K tokens
Google	Gemini 2.0 Flash	Standard API costs	1M tokens

API Key Handling

When you select Cloud AI, the wizard prompts for your API key. The key is stored locally in `~/.productteam/prefs.json`. It never leaves your machine and is never sent to ProductTeam or any third party. Only the LLM provider receives API calls.

If your API key is already set as an environment variable (e.g. `ANTHROPIC_API_KEY`), the wizard detects it automatically and asks if you want to use it.

6. Cost Comparison

Estimated costs for a typical small project (2-3 sprints, CLI tool complexity):

Path	Model	Cost	Notes
Local AI	gpt-oss:20b (Ollama)	Free	Runs on your hardware. ~20 min/step.
Cloud (cheap)	Claude Haiku	Standard API costs	Best value for simple projects.
Cloud (balanced)	GPT-4o	Standard API costs	Good balance of cost and quality.
Cloud (powerful)	Claude Sonnet	Standard API costs	Best output quality.

Costs scale with concept complexity (more features = more sprints = more tokens), quality level (strict costs 3-5x more than standard), and model choice.

The **cost circuit breaker** (default \$2.00) kills the pipeline if cumulative cost exceeds the limit, saving all work to disk. Set with `--budget` or in `productteam.toml`.

7. CLI Reference

Command	Description
productteam	Interactive wizard. Concept input, provider selection, auto-setup.
productteam preflight	Test Ollama model capability (basic, tools, multi-turn).
productteam init	Initialize a project directory.
productteam run "concept"	Run the full pipeline with a concept.
productteam run	Resume from current state.
productteam run --auto-approve	Headless / CI mode.
productteam run --budget 1.50	Set cost limit (default \$2.00).
productteam run --step prd	Run only a specific stage.
productteam recover	Reset stuck stages and re-run.
productteam status	Show pipeline status.
productteam doctor	Check environment and config.
productteam config set KEY VALUE	Set configuration value.
productteam test	Run the test suite.
productteam test --live	Run live integration tests.
productteam forge "idea"	Submit an idea to the Forge queue.
productteam forge --listen --dashboard	Start the Forge daemon + dashboard.
productteam forge status [JOB-ID]	Check job status.

8. Safety & Recovery

ProductTeam runs LLM-generated shell commands on your machine. That is inherently risky. Here is how it is mitigated:

Path validation

All file operations are locked to the project directory. No ../ traversal, no absolute paths.

Environment isolation

Builder subprocesses receive a minimal allowlisted environment (PATH, HOME, TMP, locale). API keys, tokens, and credentials from the parent process are not forwarded.

Command filtering

Known credential-adjacent paths (.ssh/, .aws/, /proc/envron) are blocked in run_bash.

Loop detection

If the LLM calls the same tool with identical arguments three consecutive times, the loop breaks automatically.

Tool call limits

Maximum 75 tool calls per doer run (configurable). After that, the stage stops and escalates.

State persistence

state.json is written on every state change. Crash at any point, resume with productteam run.

Timeouts

Every stage has a configurable timeout. Default: 300s for thinkers, 600s for doers. Auto-tuned to 3600s for Ollama.

Budget circuit breaker

The --budget flag sets a hard dollar limit (default \$2.00). Kills the pipeline mid-loop and saves all work to disk when exceeded.