

Predictable AI Spend.

No Cloud. No Compromise.

The on-device AI productivity assistant that runs routine work at zero marginal cost, never sends data to the cloud for inference and capitalizes on the underutilized compute across end user devices.

60-80%

of daily AI tasks are routine work, today billed at frontier-model rates

\$0

marginal cost when that routine work runs on-device with Cephable

Flat price, no per-token meter

Jun 2026

metered, per-token AI billing arrives as Cloud AI goes base + usage billed

[github.blog, Apr 2026](https://github.blog/2024-04-26-gemini-on-device/)

The problem

- Metered AI billing has no natural ceiling. Spend scales with usage, not value.
- Most of the workday's AI is routine work, priced at frontier-model rates.
- In a regulated firm, every cloud call is client data leaving your perimeter.

Cephable's answer

- Routine work runs on-device at zero marginal cost.
- The cloud bill reflects only work that truly needed frontier capability.
- Nothing leaves the device for inference, so there is no risk surface to govern.

One agent. Three capabilities. All on-device.



Control

Operate any app and the OS by voice, dictation, and Quick Actions. App-aware, on-device.



Create

Intelligent Tasks, Generate, Rewrite, Summarize, Translate, in any running app, bringing AI across platforms.



Automate

Completes multi-step workflows in the background while your team keeps working.

Built for financial services



Security by inheritance

Inherits each user's existing roles and governance. It can't do anything the user couldn't.



Jailbreak-resistant rules

IT can hard-block apps and sites (banking, payroll) before the agent acts. Not prompt-based.



Runs alongside Copilot & Gemini

Adds cross-ecosystem coverage without displacing your existing licensing while limiting spend.



Deployed in hours

SSO, IT Admin Portal, and license management. Runs on the fleet you already own.