Inference Utility (publicai.co/utility)

Header: Inference Utility
Subheader: AI for everyone

The Public AI Inference Utility is the public access point for open and sovereign AI models. Imagine a water or electric utility—but for AI.

It's a lightweight, open-source frontend and deployment layer that runs on compute from public and private partners around the world, giving citizens, researchers, and policymakers a simple chat interface to interact with vetted open models.

Instead of depending on private APIs, the Utility offers direct access to models built **by public institutions**, backed by transparent governance and a sustainable funding model.

Why it matters

Today, nearly all AI access is mediated by private companies. Even when open models exist, **public access is fragmented, opaque, and fragile**.

The Inference Utility fills that gap:

- **Public access to public AI** — a shared service for citizens, researchers, and institutions
- **Clean, easy-to-use interface** — chat with open models, integrated with search and national knowledgebases
- **Sovereign infrastructure in progress** — anchored in publicly funded deployments

What you can do with it

- **Open source interface.** Inspectable code and transparent routing.
- **Multilingual & jurisdiction-aware.** Tuned for European and international deployments.
- **Multiple vetted models.** Access to a range of open-weight models from national labs and research centers.
- **Privacy-first.** By default, user prompts and outputs are not logged.
- **Public governance.** Funding, model selection, and operating principles are openly documented.

Who is it for?

- **Citizens & communities** seeking a trustworthy alternative to corporate APIs
- **Public sector initiatives** piloting or scaling sovereign AI models

- **Academics & nonprofits** who need reliable access to open models
- **Policy institutions & regulators** testing transparent AI deployments

How it’s sustained

The Utility is a **pilot in building a sustainable business model for public AI access**. It combines:

- **Donated compute** from academic, nonprofit, and industry partners
- **Advertising subsidies** to offset costs of free public access
- **State and institutional funding** to guarantee long-term availability

The goal is to make AI inference a **public service**, not a private privilege.

Governance & affiliation

The Public AI Inference Utility is built by Metagov, a nonprofit research lab, under its Airbus for AI initiative. It is part of the broader movement for public AI — an effort to make AI a form of public infrastructure, like highways, water, or electricity.

Roadmap

After the Apertus launch, we hope to expand the Utility with new launch partnerships in countries like Singapore, Spain, and Canada. Other improvements: integrated web search, support for image models and multimodal queries, and more jurisdiction-aware handling to reflect different legal and cultural contexts. On the sustainability side, we are refining both the advertising-supported and utility-style business models, while testing “Plus” and “Pro” tiers that remain accessible. Longer-term, we want the Utility to support national data flywheels and nation-scale inference infrastructure.

Developers

Looking to help with backend, frontend, ops, or evaluation?

Check out the Developer Guide for information on our architecture and platform.

Key areas we need help on:

- Evaluation & monitoring
- Specialized telemetry and data pipelines
- Setting up (and getting something out of) Mixpanel
- Optimizing our authentication flow
- Design and deploy a reputation system based on data contributions
- Load balancing and gateway configuration
- vLLM / HF CICD pipeline with national labs
- General chat UI improvements

Civic & Creative Contributors

You don’t need to write code to make a meaningful impact.

We’re especially looking for people who can:

- Run and manage distributed contributor teams (e.g. product, policy, or ops)
- ☝️ is especially important, because we need to coordinate other contributors
- Translate content and documentation (i18n / l10n)
- Support design and storytelling
- Draft and communicate policy guidance, especially in local governments

Code of Conduct

We’ll have a code of conduct soon. In the meantime, don’t be a jerk.