



Introducing Tanuki: The Local only AI Productivity Helper

Shouldn't AI do the boring, repetitive things? That's what we think, so we made Tanuki; a private, local only inference pipeline tool designed to process voice, text and documents.

Tanuki isn't going to wow your dinner guests, but it also won't empty your bank account while you sleep. In fact tanuki can't ever access anything other than what you pass in. It's up to you where it forwards your output, be that local, to Dropbox or one of five other common storage providers.

Designed for those who don't want to think about how the pipeline works AND for those who do. Out of the box Tanuki can summarise documents, transcribe and clean up voice recordings and work with you in realtime while you write emails or other documents. Follow the [getting started walk-thru](#) and you'll be processing your first summarisation seconds after Tanuki pulls down a model.

For those who do want to touch the models we expose configuration parameters, and allow you to craft your own workflows and context templates, to truly customise your text processing capabilities.

Tanuki's design is unique, a local only AI productivity tool, leveraging your choice of models, utilising the cutting edge MLX toolchain for fast inference. Tanuki offers a private and user-friendly experience for anyone looking to enjoy inference without giving away their information.

For small businesses, a second in-app purchase enables an HTTP server, so your local office can share a single Tanuki instance, using up all that Apple Silicon headroom on the office's least used Mac.

Key Features

- **No Token Cost:** You use your own hardware so there are no token or monthly limits of any kind.
- **Local Inference:** Tanuki uses local models powered by MLX, ensuring quick and efficient processing without relying on external servers, enhancing user privacy and security.
- **Workflow Customisation:** Users can create and customise reusable workflows to suit their specific needs, allowing for flexible and tailored AI-driven tasks.
- **Real-Time Editing:** The app features a real-time document editor for brainstorming with access to both workflows and a local chat instance.
- **Desktop Widget:** power users can assign a keyboard shortcut to record short voice recordings and drop them into the currently active application.
- **BYO-Model:** Tanuki can use most models on HuggingFace, public or private, don't like the Qwen model we load during onboard? No problem pull another and swap it out.
- **Network Subscription:** Tanuki offers a secure local network server model with HTTPS and OIDC options.

Also Available

A headless network appliance version of Tanuki. While its presently pre v1 the headless Tanuki has been run through some serious load testing, we are happy to demo, share the results, and the tests, on request.

Coming Soon

In August we will release Tanuki TinCan, a basic Go executable HTTP proxy for your office, allowing users to push requests to your HTTP enabled Tanuki without crafting your own page or using the provided Swagger interface.

Availability

Tanuki is available on subscription for \$7.99US per month

Local network subscription an extra \$9.99US per month

Annual subscriptions and forever purchases available at a discount.

Download on the App Store: <https://apps.apple.com/app/id6762837679>

About the Developer

Hypermedia Tech is a platform engineering company that thinks Apple Silicon is future of desktop computing. So now we build quality apps for modern Apple machines, taking advantage of local machine learning capabilities while designing for scale.

We believe your machine is yours and you shouldn't have to give your valuable data to us, or some inference provider.

Press Contact

Contact: Bruno Watt

Email: contact@hypermedia.au