Small models that check their work.

An independent lab building compact open models, and the stack that lets them route, search, verify and act.



North ML · October 2026 · north-ml.space

Grounded AI is priced for the few.

Frontier APIs

$2–$50

per million output tokens. Closed weights, cloud only, and your data leaves your machine.

Local small models

Cheap, private

but stale knowledge, made-up facts and no tools. Fluent, not reliable.

The gap isn't raw parameters. Small models can't check their work.

Make the system smart, not just the weights.

Route

A 1,538-parameter router decides whether to search and whether to think.

Search

Live web and Wikipedia, reranked with BM25 and embeddings, passages picked per question.

Verify

Fiber-1 Pro checks every claim and number against the sources before answering.

Act

Python, maths, time, weather and geocoding tools ship with the runtime.

The model is one component. Routing, search, state and tools are the product.

The Flare-1 lineup.

Model	Parameters	Size (8-bit)	$ / 1M in	$ / 1M out	What it adds
Ink-1	~725M	755 MB	$0.01	$0.03	Starlight Small backbone with the Kestrel base
Fiber-1	~1.63B	1.76 GB	$0.04	$0.08	Starlight Pro backbone with the Kestrel base
Fiber-1 Pro	~1.63B	1.76 GB	$0.08	$0.15	Online GRPO plus a neural fact-verification pass
Flare-Preview	1M-class	—	—	—	Architecture preview and a published negative result

All Apache-2.0 on Hugging Face. Prices are post-launch preview rates; usage is free during the beta. 20–100× cheaper output than frontier APIs.

One file, one Mac mini.

Starlight backbone
24 layers: three gated DeltaNet, then one GQA attention, repeated.

Fiber trunk
a 158M transformer that routes, drafts tokens and holds associative memory.

AgentRegisters
memory slots whose state survives across calls.

Era
lossless snapshots and rollback of every state, for speculative decoding.

Kestrel runtime
MLX, streaming each stage as it happens.



The whole lineup runs on one M4 Mac mini with 16 GB: about 105 tokens a second.

Measured, then published.

Benchmark	Ink-1	Fiber-1
ARC-Challenge (1,165 q)	63.6%	**73.5%**
MMLU (280 q subset)	39.3%	**48.9%**
HumanEval (164)	18.3%	**37.2%**
GSM8K (300 q)	40.3%	unstable, flagged

19/20

Fiber-1 Pro on a held-out reliability eval (Fiber-1 and Ink-1: 17/20), with fictional entities and offline sources.

We publish failures too: Flare-1M showed no significant advantage over a matched GPT baseline, and we said so. English only; no claim of frontier parity.

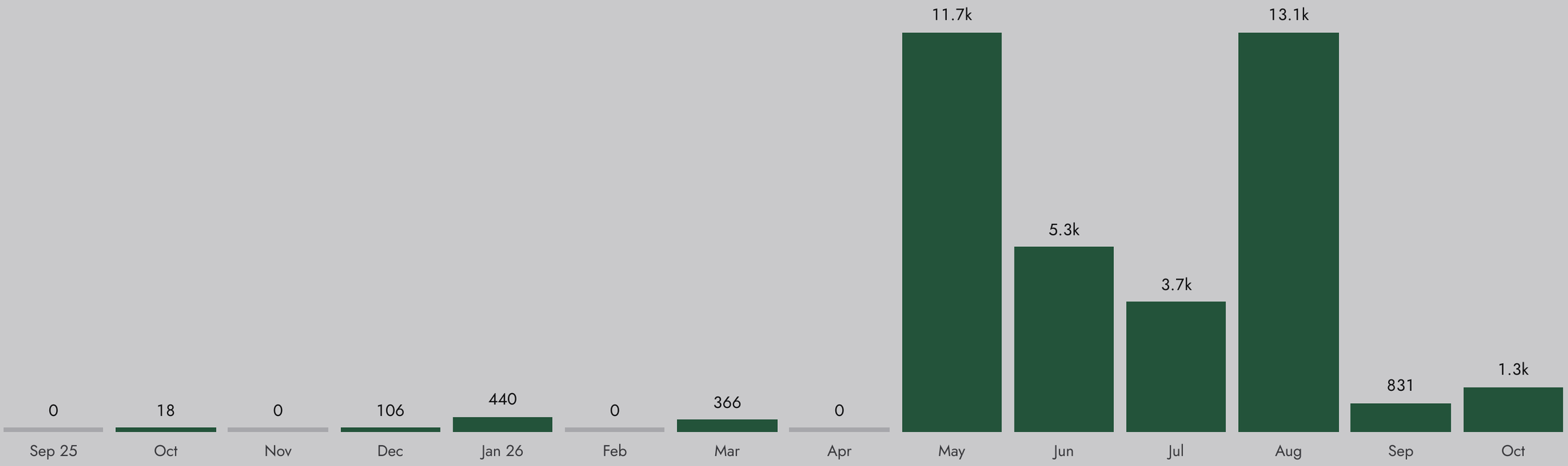
Thirteen months on Hugging Face.

55
public model repos

36.8k
model downloads, all time

4.2k
downloads in the last 30 days

Sep 2025
first upload



All-time downloads of the models released in each month, North-ML1 and arthu1 accounts. Source: Hugging Face API, 21 October 2026.

Shipping, every month.

Sep 2025

First uploads

snowy, Neura-1, then astrocoder-star

Jan 2026

starlight-mini

first North-ML1 release

Mar 2026

North Star family

trained from scratch on one Mac mini

May 2026

Wind-Edge 1.6 GGUF

3.8k downloads; starlight-mini GGUF 4.3k

Jun 2026

Willow, Forge-1, Curriculum

11 releases in a month

Jul 2026

Aurora One Mini

3.2k downloads, 6 likes

Aug 2026

Qwen3.8-Smol-uo, Aurora Proelia

11.4k downloads for Smol-uo

Oct 2026

Starlight, then Flare-1

Ink-1, Fiber-1, Fiber-1 Pro

What people can use today.

Parallax

A research workspace: plans, searches the live web, cites sources. Ink 1, Fiber 1 Lite, Fiber 1 Pro and Mix.

API

OpenAI-compatible and streaming, with a dashboard, keys, usage logs and credits. One model, north-auto.

Solace

An autonomous coding app that runs its own code and repairs it until tests pass.

Open weights

Every model Apache-2.0, many as GGUF. Download, run, fine-tune.

Accounts are approved by hand during the beta.

Free to start, cheap to scale.

Free tiers

API: 100,000 credits a month (1 credit = 1 token)

Parallax: 100,000 tokens a day, separate from API credits

Open weights: free forever, Apache-2.0

Credit packs

Starter	$5	25M credits
Builder	$20	110M credits
Scale	$50	300M credits

Purchased credits never expire. Billing switches on when Stripe is connected.

20–100× cheaper output than frontier APIs, and the weights are yours to run.

What's next.

Known limits, published

English only

Hallucination is still possible

The router is a linear probe

One Mac mini serves everything

Plans

More inference capacity beyond a single machine

Longer-context training to test the Flare mechanisms properly

A stronger router, then more languages

Paid credits live, with the free tiers kept

Everything we learn, including what fails, gets published.

Run it on your own desk.

Weights on Hugging Face. Papers on the site. An API and a research workspace for anyone we let in.

Site

north-ml.space

Research

north-ml.space/research

Weights

huggingface.co/North-ML1

Try the API

north-ml.space/try