Cloudflare opens the weights of Clairic, and swaps the model behind an existing endpoint

20 November 2026 · 4 min read · blogs.iictp.org/cloudflare-clairic-open-weight-endpoint-swap

Cloudflare announced on 18 November 2026 that it has released updated versions of its Clairic data-extraction model, and a smaller Clairic-flash model, as open weights on Hugging Face. Both are also hosted on Workers AI. The announcement also changes which model sits behind the existing Clairic endpoint, which matters for anyone already using it.

TL;DR The short version

- On 18 November 2026 Cloudflare released its Clairic data-extraction model and a smaller 9B Clairic-flash as open weights, also hosted on Workers AI.
- It also changed which model sits behind the existing Clairic endpoint.
- If you call that endpoint, re-run your test set and decide which endpoint each workload uses, because an endpoint name is not a model version.

Read online, listen to it, or change the text size: https://blogs.iictp.org/cloudflare-clairic-open-weight-endpoint-swap

What was reported

- Cloudflare says the new Clairic was redesigned after its preview, with better extraction from complex documents, including images and, newly, PDFs, and more reliable handling of tables, lists and dates.
- Cloudflare describes Clairic-flash as a lighter, faster 9-billion-parameter model. It says the 9B model previously served at @cf/cloudflare/clairic moves to @cf/cloudflare/clairic-flash, and a 27B model now serves @cf/cloudflare/clairic.
- Cloudflare says existing users of the preview endpoint will automatically get the new model, which it is rolling out in place, and that the API schemas are unchanged.
- According to Cloudflare, both models cost $0.10 per million input tokens until 1 December 2026, after which Clairic returns to $0.24 per million input tokens. It says it does not charge for output tokens.
- Cloudflare reports internal evaluations against other extraction models and says image-reading errors on its internal benchmark fell from 207 to 0. It says there are no public extraction benchmarks yet and that its comparisons are self-reported and directional, so these are vendor claims.

What a practitioner should do

- If you call @cf/cloudflare/clairic, assume its outputs and latency have changed. Re-run your own test set of documents and compare field by field with your previous results.

- Decide before 1 December which endpoint each workload should use. Move cost-sensitive work to clairic-flash if its accuracy is good enough for you, and update budgets for anything that stays on Clairic.
- Keep using the confidence scores and character spans the API returns to route low-confidence extractions to a human, rather than trusting every field.
- If you need a fixed model version, consider self-hosting the open weights so that a provider cannot change the model behind your endpoint. Check the licence on the Hugging Face model card first.

The wider lesson

A model endpoint name is not a model version. When a provider can change the model behind a stable name, your evaluation set, not the release notes, is what tells you whether your system still works. Pinning versions where you can, and testing where you cannot, is the habit the Applied AI domain of the TPBOK asks for.

> Facts in this post were read from the sources listed below on 20 November 2026; this is a developing story or product and details may have changed since.

Sources and further reading

1. Alex Reneau, Michelle Chen and Kevin Flansburg, Cloudflare. An updated Clairic and Clairic-flash are now open-weight, from the new Workers AI Superintelligence team. 18 November 2026. blog.cloudflare.com/open-clairic/

Published by IICTP, an independent technical credential body, on 20 November 2026. The current version of this post is at https://blogs.iictp.org/cloudflare-clairic-open-weight-endpoint-swap. More guides: blogs.iictp.org.