

Cloudflare's Clairic: A Non-Generative Extractor, No Output Tokens, And No Published Benchmark

TechRounder PDF Edition

Live article: <https://www.techrounder.com/ai/cloudflare-clairic-system-one-extraction-model/>

By Vipin PG | Published October 20, 2026 | Updated October 20, 2026 | Format: Deep Dive | 8 min read

In brief

Cloudflare Clairic extracts typed values with no output tokens at \$0.10 per million input, but the vendor's own pipeline tables do not reconcile.

Cloudflare's new extraction model finished the company's own three-step support test in 19.3 seconds. The same test, run entirely on a large language model, took 61.7 seconds. Add up the three step times on the all-LLM side of Cloudflare's table, though, and they come to 54.4 seconds. The last 7.3 seconds are never explained.

That gap sets the tone for Cloudflare Clairic, which the company published as a preview on Workers AI on 19 October 2026. The mechanism is a real change to a real bill line: extraction stops generating tokens, so the output side of the meter disappears. The evidence is three worked examples, one 21-day forecast, and one pipeline run whose figures do not reconcile with the rates Cloudflare publishes.

The fair verdict on 20 October 2026 is narrower than either the announcement or the scepticism: promising mechanism, unmeasured model. No independent test exists that this article could find, and every number in circulation traces back to the launch post.

What Cloudflare actually shipped

Clairic runs on Workers AI as '@cf/cloudflare/clairic'. The post is timestamped 2026-10-19T17:48:26Z and carries three bylines: Alex Reneau, Michelle Chen and Kevin Flansburg. Cloudflare calls it "the first model in a new class: System One extraction models."

The operative word is preview. The model reference describes the endpoint as "in preview from the Cloudflare Workers AI team," and its terms are blunt: the model is offered "AS IS" and "AS AVAILABLE," Cloudflare "may cease providing it without entitling you to any refund, credit, or other compensation," and it "has no obligation to indemnify you."

Weights are promised twice and dated never: "soon," then "after some more work." No licence appears on the post or the model reference, unlike Clef, which shipped with Apache 2.0 weights on Hugging Face the day it launched.

The class Cloudflare is claiming, and who named it

Cloudflare splits model work three ways: System One extract (Clairic), System One decide (Jev, Clef), and System Two, which covers frontier LLMs. The first two columns get one parallel pass; the third works "one token at a time." A decision model needs "answers you can list out in advance," and an invoice total or tomorrow's airport traffic has no such list, hence a new class. What you get instead is a typed answer: "every answer is a value of its declared type (or null), inside your bounds," against LLM strings that "must be parsed and checked."

The vocabulary is TypeSafe's. On 15 September 2026 the startup published its System One launch and a model called Jev at \$0.042 per million input tokens, with output tokens free and 70ms-500ms latency. Cloudflare credits it by name. Inside five weeks a startup coined the class and Cloudflare defined a subclass above it.

Cloudflare's exact claim is "the first model in a new class," not the first extraction model. NuMind's NuExtract, published 24 June 2024, was "a lightweight text-to-JSON LLM" trained from 0.5B to 7B parameters and released under MIT. The distinction that survives is architectural: NuExtract emits JSON as text, which under Cloudflare's own taxonomy is generation, not extraction.

How a request works

A request is a 'state' - text plus up to four images - and a map of named fields, each with a type and optional instructions. Types run from text returned as a span with character offsets to estimates returned as a distribution inside your bounds and forecasts for up to 64 steps. Scalar answers carry 'p_null' and a confidence score; estimates carry a median and an 80% interval from p10 to p90.

Field count is uncapped. Fields that will not fit beside the state in one 8,192-token pass are split across parallel passes that each see the whole input, against a 65,536-token context. The docs' example returns 'passes: 3', 'state_tokens' 61, 'schema_tokens' 472, 'input_tokens' 655 and 'output_tokens' 0.

Nulls, spans, and the failure mode that replaces bad JSON

In the post's support email the tracking number does not exist. Clairic returns 'null' at 'p_null' 1 and, separately, a 'guess' of "\$52.10" at confidence 0.416 - the tax figure a few characters away. A JSON-emitting model would likely have returned that string as the answer, and the mistake would have surfaced downstream. Text and list answers carry offsets too, such as 'spans: [[120, 129], [131, 141]]' for two names in the docs example.

The same request shows the other half of the pitch. 'amount_paid' returns 701.1, the sum of \$649.00 and \$52.10, and 'days_since_delivery' returns 11, counted from 28 September to 9 October. Neither number appears literally in the email.

What "no output tokens" changes, and what it does not

Mechanically the claim is simple: Clairic "only processes input tokens with a single forward pass, and thus has no charge for output tokens." There is no reasoning step and no generated text, so extraction here is non-generative by construction. The consequence Cloudflare draws is structural rather than statistical - "The architecture makes every answer well formed, not necessarily correct" - and the company concedes the hole in the same sentence: it cannot produce a mistyped answer or hallucinate in the usual way, "though it can, of course, make mistakes." The documentation repeats the caveat twice: "Typed results provide a predictable structure but do not guarantee correctness," and constraints "do not guarantee that the answer is correct."

Cloudflare Clairic pricing: \$0.10 per million input tokens

Clairic is listed at "\$0.10 per M input tokens" with no output line beside it. The Workers AI price list gives the same figure as 9,091 neurons per million input tokens.

On the same rate card '@cf/cloudflare/clef' is \$0.240 per million input tokens and the comparison model, 'glm-5.3', is \$1.400 in, \$0.260 cached and \$4.400 out. Per input token Clairic is 14 times cheaper than GLM-5.3 and 44 times cheaper than its output rate - a rate 3.14 times GLM's own input price, which is the structural reason a step with no output is cheap at all.

The catch is that the schema is input too. In the docs example the document is 61 tokens and the field definitions are 472, so you pay mostly for describing what you want. "Asking for thirty fields costs about the same as asking for three" holds while the schema fits beside the state in one pass; 'passes: 3' shows that splitting is real.

The claims pipeline, as Cloudflare published it

For the pipeline test, Cloudflare ran a damaged-product support email through Clairic (extract), Clef (decide refund or repair) and GLM-5.3 (write a reply under 60 words), and through GLM-5.3 doing all three jobs:

Image: As published in Cloudflare's Clairic post, 19 October 2026. Cloudflare's summary: "69% less time, or about 3.2 times faster," and "costs 49% less."

Cloudflare also reports that both pipelines "got the facts and the decision right in every run," without saying how many, and that Clairic's repair estimate was identical each time while GLM-5.3 returned \$250, \$250 and \$220.

Reading that table against its own numbers

Start with latency, because the arithmetic needs no access to the model. The cascade's steps sum to 19.28 seconds against a stated 19.3, which rounds cleanly. The all-LLM steps - 16.1, 21.9 and 16.4 - sum to 54.4 seconds against a stated 61.7, leaving 7.3 seconds unexplained.

A benign explanation deserves stating first: sequential pipelines have overhead between calls, and Cloudflare's demo page notes that "latency is measured inside the Worker around each model call." The post does not say so, though, and 12% of the headline is a lot to leave implicit.

The second reading is worse for the framing. The reply step is the one step both pipelines share, and in the cascade it was slower: 18.6 seconds against 16.4. The entire 3.2×advantage comes from extraction and decision, where 0.68 seconds replaces 38. The headline percentages are correct roundings, but they describe a pipeline whose generative tail regressed.

The cost column splits the same way. The all-LLM figures reproduce exactly at list prices: 390 input tokens at \$1.40 is \$546, and 2,544 output tokens at \$4.40 is \$11,193.60, rounding to the published \$11,194, for \$11,740. The cascade's input line does not. Take 701 tokens at Clairic's \$0.10 and you get \$70.10; at GLM-5.3's \$1.40 you get \$981.40; neither is \$184. That works out to an effective \$0.26 per million, which matches no published rate, and since \$184 is exactly \$6,032 minus \$5,848 it reads as a balancing figure. Output is still 96.9% of that \$6,032, so the saving comes from the absence of generated text at the first two steps.

Twenty-one days of TSA counts

The forecasting test is the only part of the launch with a public ground truth. Over 21 days, Cloudflare fed Clairic the TSA's daily checkpoint counts, from 16 July to 13 October 2026, and asked for the next day's number.

Cloudflare reports the forecasts were "off by 3.3% on average, and 17 of the 21 actual counts landed inside its 80% interval." The worked day is 14 October: median 2,419,283 predicted, actual 2,419,474, interval 2,270,774 to 2,556,769. That one is checkable, and it holds. The series I opened lists 2,353,356 for 13 October, the last input in Cloudflare's example, and 2,419,474 for 14 October.

Three of the four misses fell on the Columbus Day weekend. "Clairic forecasts from the patterns in the series but doesn't account for events the series doesn't show, like holidays," the post says.

What cannot be checked is the aggregate. Per-day intervals are not published, so 3.3% and 17-of-21 rest on 21 observations of one series. At that size a single hit or miss moves the coverage figure by 4.8 percentage points, which cannot separate a calibrated interval from a fortunate one.

The caveats Cloudflare publishes, and the benchmark it does not

The limitations section is the most credible page of the launch. One- and two-step calculations "seem reliable," Cloudflare writes, but "longer chains of dependent steps are hard for any one-pass model without a scratchpad. The same goes for aggregating across many rows of a table." Intervals "are trained to be calibrated and are on most held-out sets, but they can run narrow on unfamiliar data," and forecasting "is basic" - "not designed to accurately replace or predict time-series models that are trained on specific data."

What the post does not contain is an accuracy benchmark: no eval table, no held-out set, no comparison against an extraction baseline. Against that, the opening sentence claims "a significantly higher accuracy" than traditional LLMs, and nothing published with the launch measures accuracy at all.

The contrast with Cloudflare's previous model is sharp. Clef shipped on 1 October with a ten-row benchmark table spanning BFCL to PhishNChips, a public evaluation site, Apache 2.0 weights, and a median latency of 209.3 ms against Jev's 524.1 ms. Two weeks later the same team shipped an extraction model with none of that apparatus, explaining that it "wanted to take the time first to introduce the concept." No independent evaluation, replication or hands-on review of Clairic was findable on 20 October 2026.

Where this leaves a buyer

The mechanism is worth a pilot, and that is a stronger statement than the evidence licenses - which is why the pilot should be built to fail visibly. Pick one high-volume extraction job, run it against your current LLM pipeline on a few hundred of your own documents, and score exact match, null precision and recall, span correctness and interval coverage. A public yardstick exists for the first: the ICDAR-SROIE receipt set has 1,000 scans, 712 annotated training images and 347 test images, and scores four fields by exact match.

Two risks sit outside any benchmark. The contract is one: a preview that may be withdrawn without compensation, with no indemnity and no licence for weights that do not yet exist, does not belong in a payment path. Portability is the other. Clef can leave Workers AI because its weights are downloadable under Apache 2.0; Clairic cannot yet.

The next real checkpoint is not another Cloudflare post. It is the first independent evaluation on a public extraction corpus, or the open-weights release, whichever comes first. Until one lands, treat Cloudflare Clairic's 19.3 seconds and 49% as the vendor's numbers, and the unexplained 7.3 seconds as the reason to check them.

References

1. [blog.cloudflare.com - clairic-extraction-model](https://blog.cloudflare.com/clairic-extraction-model/) - <https://blog.cloudflare.com/clairic-extraction-model/>

2. developers.cloudflare.com - workers-ai / models -
<https://developers.cloudflare.com/workers-ai/models/clairic/>
3. typesafe.ai - blog / introducing-system-one-models-and-jev -
<https://typesafe.ai/blog/introducing-system-one-models-and-jev>
4. numind.ai - blog / nuextract-a-foundation-model-for-structured-extraction -
<https://numind.ai/blog/nuextract-a-foundation-model-for-structured-extraction>
5. developers.cloudflare.com - workers-ai / platform -
<https://developers.cloudflare.com/workers-ai/platform/pricing/>
6. tsa.gov - travel / passenger-volumes - <https://www.tsa.gov/travel/passenger-volumes>
7. blog.cloudflare.com - clef-decision-models - <https://blog.cloudflare.com/clef-decision-models/>
8. huggingface.co - datasets / Voxel51 - https://huggingface.co/datasets/Voxel51/scanned_receipts