DefAcc Hackathon 2026 · Technical sheet

ECO

Measures how many independent sources stand behind a news story, a photo or a voice note, before you share it.

Team: Sadid Romero, Juan Felipe Jimenez, Ana Maria Vargas · ECO TRACE Team

Track: 4. Epistemic defense

Repository	Demo	Contact
https://github.com/SadidRomero77/eco	Live: https://eco.kibosecondbrain.com MCP: https://eco.kibosecondbrain.com/mcp	sadid_05@hotmail.com

Summary

AI makes it almost free to clone a story across dozens of sites, fake a photo's date or clone a voice, and both people and AI agents count that repetition as confirmation. We built ECO: a web app, a browser extension and an MCP server that measure how many independent sources back a claim, verify C2PA signatures and detect synthetic voices. On unseen test worlds it counted sources correctly **92 %** of the time (text similarity: 83 %).

1. The threat

Source laundering: one claim is copied and AI-paraphrased across many sites, illustrated with someone else's photo or a false date, or arrives as a WhatsApp voice note with a cloned voice. A search engine shows eight pages and they look like eight sources; a web-searching AI agent answers "several sources confirm it". It harms whoever receives and forwards rumors, especially in Latin America, where they spread on WhatsApp, which strips metadata. Human fact-checkers do not scale, and lab-trained voice detectors fail on compressed audio.

2. What we built

ECO is **a single investigating agent**, not a multi-agent system: code orchestrates its steps and the agent calls tools. It searches recent news with the **Exa API**, discards articles about a different event, and a model (Claude on Amazon Bedrock) extracts the checkable facts in each one. Then an **auditable, LLM-free rule** walks the articles in chronological order and counts one as a source only if it adds a new fact. Two reinforcements: **C2PA signatures** and a perceptual hash expose reused photos or fake dates, and an **AASIST-L detector** retrained with WhatsApp-like compression checks the voice. Input: a topic, URL, voice note (Whisper) or screenshot (Claude vision). Output: a one-sentence verdict, the origin tree and the recommended source to read first. It runs on the web, as a Chrome extension or as an **MCP server with 6 tools**, so other agents can check before saying "confirmed".

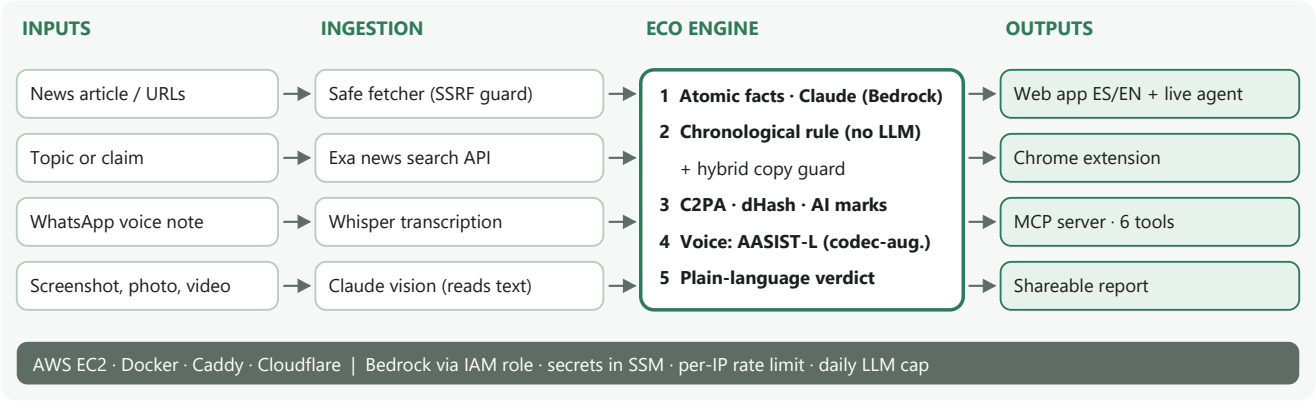


Figure 1. ECO architecture: a single agent orchestrated by code. Four input types go through safe ingestion (Exa search, Whisper transcription, Claude reading screenshots) and a five-step engine; results reach the web app, the extension, the MCP server and a shareable report. Runs on AWS behind Cloudflare.

3. How we evaluated it

We built synthetic "worlds" (fictitious outlets, people and places), each with its ground truth stored apart: 36 for tuning, 18 from new events as a final exam evaluated once, and 12 with signed photos. The baseline groups pages by text similarity. The metric is the **exact count of independent sources**, because that is precisely what source laundering fakes. Voice was measured on In-the-Wild: 1,200 clips from 17 unseen speakers, clean and compressed.

4. Results

Metric	Baseline	ECO	Difference
Exact source count (final exam)	83 %	92 %	+9 pts
Paraphrases recognized as copies	33 %	67 %	+33 pts
Agent says "confirmed" with a single origin ↓	51 %	22 %	−29 pts
Rewritten copies reusing the signed photo	0 %	100 %	+100 pts
Voice detector error, WhatsApp-like note ↓	49.7 %	28.7 %	−21 pts

Table 1. Final exam: 12 main worlds, 3 per type. Agent: Claude Haiku on 54 worlds, without and with ECO's report. Photos: 12 C2PA-signed worlds, baseline is text only. Voice: equal error rate (EER) under 16 kbps Opus compression; baseline is AASIST-L as published. ↓ = lower is better.

Key finding: measuring independence of facts, not textual similarity, defeats paraphrasing without merging independent outlets. **Where it fails:** if each copy rewrites and invents different details without reusing the photo, ECO also counts it as a source. We do not report the 100 % on tuning worlds: it was overfit.

5. Limits and dual-use risk

Independent does not mean true. C2PA proves who signed, not what is true; almost nothing is signed and we use a demo certificate authority. A wire agency published after a copy can look like a repetition. Voice was only measured in English and errs 29 times in 100; image forensics are unevaluated. **Dual use:** someone could polish copies until they pass; we mitigate with per-IP limits and a daily cap, unpublished thresholds, and we frame every verdict as an alert, never an accusation.

References

C2PA (2024). *C2PA Technical Specification 2.1*. Coalition for Content Provenance and Authenticity, c2pa.org.
Jung, J. et al. (2022). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. *ICASSP*.
Müller, N. et al. (2022). Does audio deepfake detection generalize? (In-the-Wild dataset). *Interspeech*.
Krawetz, N. (2007). A picture's worth: digital image analysis and forensics. *Black Hat*.