# The Video That Wasn't There

Seven Claude models reported no video on the only two pages that had one. Five other Claude models, given the identical task, got every field right. The difference was not the model. It was the tool the run reached for.

WHITE PAPER · MODEL EVALUATION SERIES

An independent field test of 44 AI models on a five-page web census — what they reported, what was actually on the pages, and why the difference came down to a single tooling choice.

Field report · Test run October 16, 2026 · Report prepared November 9, 2026 · Tygart Media · The websites tested are not named, at the operator's choice. Every model is named, because the models are the subject.

## Executive summary

Forty-four AI models were given the same plain task: visit five live web pages and report each page's title, top heading, embedded-video presence, and section-heading count, against an answer key fixed from raw page source before the run. Thirty-six models scored a perfect 20/20. Seven of the eight imperfect runs were Claude models, and all seven failed the same way: on the only two pages containing an embedded video, they reported that no video existed, and undercounted those pages' headings. Trajectory analysis — a review of each run's recorded tool calls — traced the failures to the page-fetch method: runs reading raw HTML were exact, while runs reading through a text-extraction layer inherited that layer's blind spots, because extraction strips embedded iframes. Two of the seven even ran an explicit verification pass that re-read the same lossy representation and certified the error. The finding generalizes beyond one model family: an agent's report describes its fetch, not the world, and verification only counts when it re-derives an answer from source.

## What was tested

On October 16, 2026, we ran an identical page-census brief to 44 AI models, each working as an agent in the same desktop coding environment. The brief was deliberately plain — a utility task, not a puzzle:

> Visit five live web pages. For each page report: (1) the exact HTML title, (2) the exact top-level heading (H1), (3) whether the page contains an embedded YouTube video, and if so its video ID, (4) the number of second-level headings (H2), and (5) a one-line summary of the page's topic.

The five pages were ordinary published pages on small business websites. Two of the five contained an embedded YouTube video. Three did not. One page had been chosen because its HTML title and its visible H1 deliberately differ — a page where a careless read produces a confident wrong answer. There were no trick pages, no paywalls, no logins.

## How it was graded

Before any model ran, we pulled the correct answers directly from the pages' raw HTML source and froze them as an answer key. Each run was scored out of 20 hard fields (four per page: title, H1, embed, H2 count) by an automated scorer comparing the model's report against that key. Because the agent environment records a full trajectory — every tool call, every fetch, every intermediate step — we also read how each model got its answers, not just the answers it turned in. That turned out to matter more than the scores.

## What happened

Thirty-six of the 44 models scored a perfect 20/20. Eight did not. Seven of the eight imperfect runs were Claude models, and their errors were not scattered — they were identical in shape:

| Model | Score | What it got wrong |
|---|---|---|
| Claude Fable 5.1 | 16/20 | Both video pages: reported no embed; H2 count short |
| Claude Opus 5 | 16/20 | Both video pages: reported no embed; H2 count short |
| Claude Sonnet 4.5 | 16/20 | Both video pages: reported no embed; H2 count short |
| Claude Sonnet 5.5 | 16/20 | Both video pages: reported no embed; H2 count short |
| Claude Opus 4.5 | 15/20 | Both video pages: reported no embed; H2 count short; one more H2 |
| Claude Sonnet 4.6 | 15/20 | Both video pages: reported no embed; H2 count short; one more H2 |
| Claude Haiku 4.5 | 14/20 | |

Both video pages: reported no embed; H2 count short; two more H2s

Every miss these seven made on the video pages was a miss of the same kind, on the same two pages, and on no others: **the video that was embedded in the page was reported as not there, and the heading count on those same pages came up a couple short.** Their titles and headings on all five pages were right. They were not hallucinating random facts. They were faithfully describing a version of the page that was missing things.

On the same day, on the same brief, five other Claude models scored a perfect 20/20 — Claude Fable 5, Claude Opus 4.6, Claude Opus 4.8, Claude Opus 5.5, and Claude Sonnet 5. Claude Opus 4.7 scored 19/20, missing a heading on a page with no video. Claude Haiku 5.5 found both videos and missed only heading counts. So this is not a story about a model family that "can't see." It is a story about what happened on particular runs.

## Why it happened

The trajectories show it plainly. Accuracy in this test tracked one variable, field by field: **how the run fetched the page.**

- Runs that fetched the **raw HTML source** and parsed it locally got every field right — titles, headings, embeds, counts.
- Runs that fetched the page through a **built-in web-fetch / extraction tool** received the page as cleaned, extracted text. In that representation, embedded iframes are stripped out entirely and some secondary headings drop. A video that exists only as an iframe embed simply does not exist in the extracted copy.
- The seven Claude runs above read the pages through the extraction path — or parsed them with a method that shared its blind spot — and then reported the extracted copy as though it were the page. Their answers were fluent, specific, and wrong in exactly the places where the extraction was lossy.

The sharpest case: **Claude Opus 5 and Claude Fable 5.1 both ran an explicit verification step** — a step whose stated job was to verify the YouTube embeds — over the HTML they had fetched, and still reported no embeds on either video page. A verification step only counts if it re-derives the answer from source. When the verifier looks through the same parser that produced the error, it certifies the error. We have come to call this verification theater: the run performs the shape of checking without the substance of it.

For contrast, the strongest runs in the fleet showed the opposite habits. One model weighed the extraction tool against a raw fetch in its recorded reasoning and chose the raw fetch explicitly, "for raw HTML." Another fetched every page twice, the second pass with a different job. Another singled out the one page whose title and heading differ and went back to its source markup to confirm it by hand. Persistence mattered too: one run failed four times on tooling errors, re-fetched everything three times, and still arrived at a perfect score without ever faking a step.

## The full board

All 44 runs, scored by the same answer key. Cost is the run's metered cost in US cents; time is wall-clock run time. One Claude run (Sonnet 4) errored out and produced no report.

| Model | Score | Time | Cost | Notes |
|---|---|---|---|---|
| Claude Fable 5 | 20/20 | 62s | 86.3¢ | |
| Claude Opus 4.6 | 20/20 | 104s | 50.6¢ | |
| Claude Opus 4.8 | 20/20 | 64s | 42.8¢ | |
| Claude Opus 5.5 | 20/20 | 49s | 27.0¢ | |
| Claude Sonnet 5 | 20/20 | 53s | 18.9¢ | |
| Claude Opus 4.7 | 19/20 | 64s | 48.7¢ | Missed one H1 |
| Claude Fable 5.1 | 16/20 | 50s | 70.3¢ | Both embeds missed |
| Claude Opus 5 | 16/20 | 83s | 48.0¢ | Both embeds missed |
| Claude Sonnet 4.5 | 16/20 | 49s | 16.2¢ | Both embeds missed |
| Claude Sonnet 5.5 | 16/20 | 36s | 13.6¢ | Both embeds missed |
| Claude Haiku 5.5 | 16/20 | 35s | 0.8¢ | H2 counts only |
| Claude Opus 4.5 | 15/20 | 47s | 26.5¢ | Both embeds missed |
| Claude Sonnet 4.6 | 15/20 | 70s | 19.8¢ | Both embeds missed |
| Claude Haiku 4.5 | 14/20 | 39s | 6.5¢ | Both embeds missed |
| Claude Sonnet 4 | — | — | — | Run errored |
| Composer 2.5 | 20/20 | 31s | 13.2¢ | Fastest run in the fleet |
| Auto (router default) | 20/20 | 58s | 5.9¢ | |
| Gemini 3 Flash | 20/20 | 119s | 8.4¢ | |

| | | | | |
|---|---|---|---|---|
| Gemini 3.1 Pro | 19/20 | 59s | 15.1¢ | Missed one H1 |
| Gemini 3.5 Flash | 20/20 | 150s | 37.8¢ | |
| Gemini 3.6 Flash | 20/20 | 99s | 26.1¢ | |
| Gemini 3.7 Flash | 20/20 | 241s | 12.7¢ | Four failed attempts, persisted |
| Gemini 3.8 Flash | 20/20 | 75s | 7.6¢ | |
| Gemini 4 Argon | 20/20 | 59s | 19.8¢ | |
| GLM 5.2 | 20/20 | 33s | 5.9¢ | |
| GLM 5.3 Flash | 20/20 | 47s | 0.8¢ | Cheapest perfect run; chose raw fetch deliberately |
| GLM 5.3 | 20/20 | 59s | 9.1¢ | |
| GPT-5.1 | 17/20 | 81s | 15.9¢ | H2 counts only; narrated steps before taking them |
| GPT-5.2 | 20/20 | 103s | 56.3¢ | |
| GPT-5.3 Codex | 20/20 | 78s | 33.1¢ | |
| GPT-5.4 Mini | 20/20 | 75s | 5.3¢ | |
| GPT-5.4 Nano | 20/20 | 65s | 1.3¢ | |
| GPT-5.4 | 20/20 | 86s | 19.9¢ | |
| GPT-5.5 | 20/20 | 47s | 22.1¢ | |
| GPT-5.6 Luna | 20/20 | 74s | 1.5¢ | |
| GPT-5.6 Sol | 20/20 | 86s | 30.4¢ | |
| GPT-5.6 Terra | 17/20 | 36s | 12.1¢ | H2 counts only |
| Grok 4.5 | 20/20 | 48s | 35.7¢ | |
| Grok 4.6 | 20/20 | 46s | 22.7¢ | |
| Grok 4.7 | 20/20 | 52s | 29.9¢ | |
| Kimi K2.7 Code | 19/20 | 63s | 4.5¢ | Missed one H1 |
| Kimi K3 | 20/20 | 33s | 23.3¢ | |
| Muse Spark 1.3 | 20/20 | 71s | 10.1¢ | |
| Muse Wonder 1.0 | 20/20 | 50s | 18.9¢ | |

## What this does and doesn't show

It does not show that Claude models cannot perceive web pages. Five of them, on the same day and the same task, perceived everything. Nor does it show intent: nothing in any trajectory suggests a model chose to deceive. The extraction-path runs described, accurately and in good faith, the only version of the page they ever received.

What it shows is more useful than either of those stories:

- **A model's report describes its fetch, not the world.** When the fetch is a lossy extraction, the report inherits the losses — silently. Nothing in the seven reports disclosed that the page had been read through a layer that strips embeds. A reader of the final answer alone would have no way to know.
- **Errors cluster by method, not by model.** The misses were identical in shape across seven different models because the failure lived one level below the models, in the tooling path several of them shared.
- **Verification has to change the evidence, not repeat it.** Two premium-priced runs "verified" the embeds and certified their absence. A check that re-reads the same representation is not a check.
- **Grade the trajectory, not the summary.** The final answers alone would have supported a lazy conclusion ("these models can't see videos"). The execution record supported the true one. Any serious evaluation of an agent should score what it did, step by step, against ground truth established before the run.

Method notes: one run per model; five pages; 20 scored fields. Heading counts are method-sensitive (raw source vs. rendered page can legitimately differ), which is why the answer key was fixed to raw source in advance and stated in the brief's scoring. Costs are the environment's metered per-run charges. Run artifacts — every model's full report, the scorer, and the trajectory analysis — are retained by the operator and available on request.

— W. Tygart, Tygart Media. Questions welcome.