20 Percent of All 'Ok' Web Responses AI Gets Are Not Okay

Originally published Weds 4 Nov 2026 at https://www.unite.ai/20-percent-of-all-ok-web-responses-ai-gets-are-not-okay/



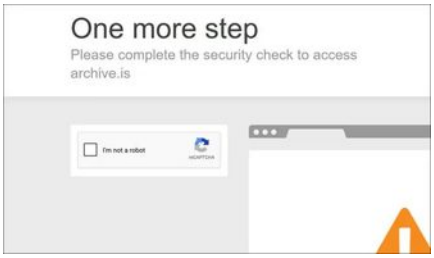
My recent experiments in AI-assisted web-scraping have brought home to me the growing extent to which the web is becoming 'locked-down'. The motivating event for this retrenchment, of course, is the shocking growth in AI web-scraping, whose uncompromising demand for new data can end up taking domains offline.

This is a good excuse for domains to do what many of them have wanted to do for a long time: make their content accessible only when the user is logged in (though naturally, Googlebot typically gets a free pass). Facebook completed its gradual adoption of this policy in 2023, and Twitter/X did likewise the same year. Reddit will adopt it later this month.

In this way, compliance with any country-specific age-limitation enforcement becomes a little easier for the domain, since the user is already registered. More importantly, the user's data can be saved, exploited for local analytics, and – local legislation permitting – sold to data brokers and AI companies.

Elsewhere, measures squarely aimed at defending against bots (rather than enrolling users) have risen more than 58% year-on-year, according to Cloudflare; and all that escalating bot activity naturally encourages websites to scrutinize VPN and proxy addresses, potentially denying legitimate users access, together with any 'malicious' traffic.

The anecdatal aspect of this is that I seem to come up against 'blocking' responses or challenges more and more, to the point where I may even abandon a site that begins loading them; and since CAPTCHA -walls have been found to deter users and potential customers, one can imagine the pressure that causes domain owners to adopt them in spite of that downside.

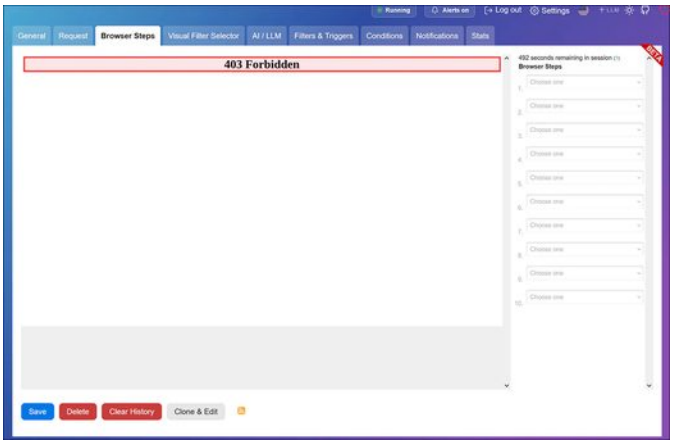


A captcha from archive.is, one of the many sites that has adopted challenges, often in response to excessive automated (rather than human) requests.

© Martin Anderson and/or commissioning entities

All this new and increased friction is partly logistical (because not all domains can withstand the onslaught) but also often a kind of negotiating tactic to see if AI companies will offer a single or collective form of remuneration for all that data they so desperately crave, and which has been free for so long, until now.

Since AI crawlers have to work at scale, they tend to use 'headless' instances of Chrome under the Playwright framework. These days, for an increasing number of domains, that will get you a 403 ('forbidden') error (i.e., host knows you're automated traffic, now scram):



A local installation of changedetection.io attempts to poll a site, but the site does not like the detected browser that the framework uses, and returns a 403 ('Forbidden') response.

What's wanted is the '200' (*'OK'*) client response, even if one has to hand over a pile of money to a residential proxy farm, in concert with other diverse non-human sneaky tricks, in order to obtain it.

There's nothing better than the 200 response; which is to say, there is no client response code with a higher level of positive signal. Neither is it an ambiguous response; it's never *'nearly okay'*, *'going to be okay later'*, or *'getting more okay'*; it's *'okay!'*. If it's not okay, it's not a 200 response.

Hello! Go Away

Except that this lack of granularity, according to a new paper, could be significantly skewing AI datasets, statistical estimations, and reporting 'good' status when that is not the case.

The study* estimates that around a fifth of 200-category responses *contain no usable page content at all*, instead delivering empty website shells, CAPTCHA challenges, access restrictions, or substitute pages.

The paper states:

> *'20.9% of the pages that a static client receives with status 200 do not contain the page's main content: app shells, bot challenges, block pages, error and placeholder pages, and partial content.'*

In this regard, impatient automated systems are tripped up by their own sparseness, since the '200' response is allowed to be equated with successful fetches, even when the fetch just produces a CAPTCHA-wall, a paywall, or some other obstacle.

The paper outlines the seven types of content found to frequently complicate a 200 response:

© Martin Anderson and/or commissioning entities

	Type	Example	Auto	Rev.
A	Content	article, product, documentation or listing text	3,575	199
B	Block	geo-block ("location not supported"); WAF "request unsuccessful"	83	3
C	Challenge	Cloudflare Turnstile; publisher "Client Challenge"; image captcha	159	12
D	Shell	product grid that only says "Loading..."; live scores missing; empty body	354	48
E	Error or substitute	random path redirected to the homepage; unrelated page served at a random path	341	31
F	Parked	hosting placeholder; parked domain redirecting to a lander	27	6
G	Consent wall	publisher privacy gate in front of the article	21	1

The seven categories of content that are anything but 'okay' in a 200 response. Source

For the study, over three days, 10,307 web addresses were tested using both a basic web-fetching program and a browser capable of loading JavaScript, with responses classified according to the *content retrieved*, rather than their HTTP status codes.

The largest categories of 'unusable' 202 responses were empty website shells and error or substitute pages, together accounting for more than two-thirds of the failures.

Regarding the empty site shells, a website might display its menus and navigation while leaving the requested article or product listing blank; in the latter case, an unrelated page or homepage might appear instead of the requested content – but both could receive the same HTTP 200 status as a successful retrieval offering usable and useful content.

By way of a remedy, the paper recommends that web-fetching tools report *what was actually retrieved*, rather than relying on HTTP status codes, since simple checks for unusually short pages, suspicious redirects and content resembling nonexistent pages could identify most failures.

More resource-intensive browser tools could then be used only when a page appears to have failed to load properly. Additionally, AI evaluations could check whether cited sources *actually contain the expected information*, rather than assuming that an HTTP 200 response means the content was successfully retrieved.

The author of the new work also notes that 'cheap heuristics' can trivially identify around 86% of unusable responses, using basic checks on page length, redirects and similarities to nonexistent pages:

Policy	Usable %	Missed %	Rendered %	Time (s)
Never render	74.3	25.7	0.0	2.5
Always render	80.3	0.0	100.0	20.6
Render if H-best flags	81.4	7.8	19.6	5.8
Render if M2 flags	81.6	6.4	22.9	6.6
Render if M3 flags	81.8	6.2	23.6	6.7

Comparing web-fetching strategies: rendering only pages flagged as suspicious recovered slightly more usable content (81.4%) than rendering every page (80.3%), while reducing average retrieval time from 20.6 to 5.8 seconds per page.

Why So Many Errant 200 Responses?

There has been some argument over the years, in developer communities, that the 202 ('Accepted') error (*'Requested accepted but not yet completed'*) could disambiguate these 'failed' 200-status requests; but really, 202 cannot accommodate all the cases outlined in the new paper, and there is no existing code that can.

Though the paper does not deal with the issue in depth, the reason *why* domain owners permit 200 errors might be worth further investigation in future related works.

One reason could be that anything that is not a 200 is essentially some kind of error page or obstruction, including the most common, 404 (*'not found'*) errors – where the target content is simply missing. Some users believe that an excess of such errors will lead to penalties in SEO (and, presumably, GEO), though Google denies this.

© Martin Anderson and/or commissioning entities

In fact, erroneous 200 errors are considered as 'soft' 404s by Google, and clearly likely to be treated as more of a problem than plain 404 errors, because these are essentially 'deceptive' outcomes, intentional or not.

Citation Drift

If all this seems like pedantry, one has to consider the petabyte-level scale of the data that is routinely gathered (or attempted to be gathered) for the purposes of AI training. If a full fifth of that data is unreliable or outright deceptive, the downstream consequences are difficult to shrug off as some kind of 'collateral inaccuracy'; rather, such false responses could significantly and adversely affect the distribution in datasets and the outcomes of training such data into AI models.

For instance, one notable issue around erroneous 200 responses concerns 'citation drift', where an AI-generated answer is no longer supported by the web page that was originally cited as evidence, i.e., the page may have changed, the relevant text may have been removed, or the link may now lead somewhere else.

The new study found that apparent citation drift rose from 5.8% to 21.9% when all HTTP 200 responses were accepted as successful retrievals, *regardless of what they contained*.

This means that a source could be wrongly judged to have changed, disappeared or ceased to support an AI-generated claim, simply because a CAPTCHA, empty page or other obstruction had been retrieved in place of the original material.

Just Say It's So!

The problem was also demonstrated by tests involving deliberately *nonexistent* web addresses. When random, invented paths were requested from websites, 13.1% of those that responded returned a 200 ('OK') status, even though the requested pages did not exist†.

> *'13.1% of the sites that answered the random-path probe returned 200 (S1 10.5%, S2 13.3%, S3 17.3%, S4 13.7%, S5 13.1%). Every such response is a soft 404 by construction, but not every one says so.*
>
> *'In an exploratory breakdown of the 172 soft-404 responses, the labeller called 98 an error or substitute page, 32 a shell, and 28 content: a homepage or listing served at a random path reads like a normal page.'*

Conclusion

If the problem was recognized at a higher level, the IANA could conceivably create a new and more apposite http status code, such as [X] (*'Request succeeded, but the requested content was not delivered'*). While this would actually describe what is happening, it would naturally require an unlikely amount of insight into the transaction on the host's side.

Similar codes already exist, though none really fit: 451 (*'Unavailable For Legal Reasons'*), is ambiguous and rather alarming (!); 206 (*'Partial Content'*), is technically apt, but intended for incomplete data (fetches) ; and the surprising-it-exists 203 (*'Non-Authoritative Information'*), which was created in 1997, and is only peripheral to the use case at hand.

Perhaps the nearest candidate is actually the unintuitively-named 103 (*'Early Hints'*), which indicates that not all the URL's resources are loaded yet (though this would clearly be a gross misinterpretation of the code's intent).

* *Titled 'When 200 OK Is Not an Answer: Silent Fetch Failures in Agentic Web Content Acquisition.' Please be aware that the paper is written, apparently, to be primarily machine-readable, perhaps to be queried by AI, and is not an easy read in itself.*

† *I apologize for the mechanistic quote, but the entire paper is written this way, as I note above, to be 'machine readable' rather than human-interpretable, though it stops short of being presented in JSON format.*

First published Wednesday, November 4, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Website: martinanderson.ai
Contact: martin@martinanderson.ai