V1.1 · AUGUST 2026

Eight thousand pages taken, ten readers sent back

Two Greek sites, 58 observed days each, measured from their own access logs. How many pages an answer engine takes for every reader it returns.

Nikolaos Broikos · broikos.gr August 2026

2 sites · 58 observed days · 6,870 verified requests

2,323 : 1

PAGES TAKEN PER READER RETURNED, FIRST SITE

633 : 1

THE SAME RATIO ON A SECOND, UNRELATED SITE

1,069

PAGES ANTHROPIC TOOK FROM A SITE IT SENT NOBODY TO

12.9×

CRAWL GROWTH IN THIRTEEN MONTHS

What the AI crawlers took from two Greek sites, and what they returned, measured from the servers' own logs.

Nikolaos Broikos, August 2026

Summary

Every argument about AI and publishing turns on a ratio nobody publishes: how many pages an operator takes for each visitor it sends back. This paper computes it for two Greek sites, from their own access logs, over 58 observed days each between April 2025 and August 2026.

One site is the author's own portfolio and research site, named. Its robots.txt does not merely permit AI crawlers, it names five of them and allows them on purpose, on the stated bet that being read by assistants is worth being trained on. The second is a client's, a single-practitioner professional practice with a blog and online booking, read with its owner's agreement on the condition that it is not identified. Neither has ever blocked a crawler.

	PAGES TAKEN	READERS RETURNED	PAGES PER READER
broikos.gr, the author's own portfolio and research site	2323	1	**2323 : 1**
a second Greek site, a Greek single-practitioner professional practice with a blog and online booking	5700	9	**633 : 1**

- **On both sites the answer engines take between six hundred and two thousand

pages for every reader they return.** Two sites, different categories, different traffic, and the answer lands within one order of magnitude.

- **Only one operator sent anybody at all.** OpenAI returned nine readers to the

practice site and one Gemini reader reached the portfolio. Anthropic took 1,930 pages from the practice site and 1,069 from the portfolio, and sent nobody to either. Microsoft took 2,104 and 799, and sent nobody.

- **9,109 pages and 993.2 MB** left the practice site, **3,466 pages and 219.7

MB** the portfolio, counting only crawler addresses that survived an identity check against the operators' own published ranges.

- **95.8 per cent and 89.7 per cent of crawler requests verified.** Requests that

failed are excluded rather than credited to whoever their user agent named.

- **Crawl volume rose 12.9 times on the portfolio** between two comparable windows

thirteen months apart, and **1.6 times on the practice site**, which was already being crawled heavily in the first window.

- For scale, on the practice site over the same window, **ordinary Google search

sent 4,484 readers** and every answer engine on earth sent nine.

The asymmetry is not the surprising part. The surprising part is how close the two sites are, given that they share nothing except a country and a decision never to block anybody.

01 AI crawlers: deliberately ALLOWED. This site is a portfolio whose job is to be

The number nobody publishes

The public argument about AI and the web has two sides and one missing number.

One side says the crawlers take content and return nothing, so the trade is extraction. The other says assistants are the new distribution channel, so the crawl is the price of being found. Both are claims about a ratio: pages taken per visitor returned. Both are usually argued without it.

The operators do not publish it. They know it exactly, per site, and have no reason to. The large publishers who have measured it have mostly done so inside legal proceedings or press releases, on properties big enough that the answer says little about anybody else's site. What is missing is the ordinary case: a small site, no negotiation, no deal, no legal department.

This paper computes the ratio for one such site over sixteen months.

Why this site is the friendly case

The site is broikos.gr, a personal portfolio and research site. It is the author's own, which is why its logs could be read at all without asking anybody's permission.

Its robots.txt is not neutral. It names five AI crawlers and allows them, with the reason written in a comment beside them:

```
# AI crawlers: deliberately ALLOWED. This site is a portfolio whose job is to be
# found and quoted, including by assistants people ask for a developer. The
# trade (training use in exchange for reach) is one I take on purpose.
User-agent: GPTBot
Allow: /
```

That matters for how the result should be read. This is not a site that was crawled against its wishes. It is a site that made the bet the operators themselves describe, and the measurement is of what the bet paid.

What is being counted

Taken: a request from a verified crawler address that the server answered with a 200 or a 206, for a path that is not a stylesheet, script, font or image. A crawler that fetches one article and its twelve images has taken one page, not thirteen. The full request count is reported beside it, because that is what a bandwidth bill sees.

Returned: a page view, served, from an agent that names a browser, carrying a referrer from that operator's answer engine. An arrival from a library or a scanner carrying the same referrer is counted separately and is not a reader.

Everything else in this paper exists to make those two definitions honest.

What was read, and what was never written down

The blocker, checked instead of assumed

This paper was parked in early August 2026 on the belief that the logs had already rotated away. That belief was never tested. Testing it took four minutes: log_inventory.py connects over FTPS to each site with a stored credential, lists two directories, and disconnects without downloading anything.

The site held twelve monthly archives going back to September 2024. The paper was not blocked and had not been for a year.

It is worth naming the shape of that error, because it is common and it is expensive: the assumption that made the work impossible was cheaper to check than to hold.

The window

MONTH	DAYS WITH TRAFFIC	CRAWLER REQUESTS
Apr 2025	1	5
Jul 2025	25	1017
May 2026	1	71
Jun 2026	9	718
Jul 2026	1	26
Aug 2026	21	5332
whole window	**58**	**7169**

Two of these months are substantially complete and comparable: **July 2025, 25 days**, and **August 2026, 21 days**. The rest are fragments, the residue of a hosting panel that keeps a rolling archive rather than a full history. The fragments are counted in every total and are not used for any rate.

Every one of the 109,107 lines matched the combined log format, and none was discarded as unparseable. That is stated because a silent parse failure rate is the easiest way for a study like this one to under-report a crawler whose requests happen to be shaped differently.

PROPERTY	VALUE
lines parsed	109107
lines not matching the combined format	0
lines carrying a user agent	109107
lines carrying a referrer	69858
requests attributed to a known crawler	7169
of those, served with a 200 or 206	6291
of those, a page rather than an asset	3667
requests not attributed to a crawler	101938
arrivals from an answer engine	2
distinct page paths the site served to anybody	474

What is never written down

The logs carry visitor IP addresses, which are personal data. The rule this paper runs on is the same one paper 02 used for merchant records:

- pull_logs.py fetches archives into raw/, which is in the never-publish set.
- extract.py is the only script that reads a raw line, and it emits aggregates

only. No address, no user agent string and no full referrer URL leaves it. Only the referring host is kept, because a full referrer can carry a search term and a search term can identify a person.

- verify_agents.py holds addresses in memory to check them and writes counts.
- ratio.py and tables.py cannot reach a raw log line at all.

The consequence is that the entire analysis path can be published while the logs stay off the site, and the headline can be recomputed by anyone holding two CSV files and no personal data at all.

This is also the reason the paper covers one site rather than five. The other sites with reachable logs belong to clients. Reading a client's access log is their decision, not the author's, and it has not been asked for.

03

A user agent is a claim, not evidence

Anybody can send GPTBot in a header. Nothing stops them, nothing checks it, and the fake ones concentrate in exactly the agent strings a site is least likely to block. A paper that counts agent strings and calls the result "what OpenAI took" has counted a claim.

So every request here was checked against what the operator itself publishes, using three methods in descending order of what they prove.

Published address ranges. OpenAI, Anthropic, Perplexity, Google, Microsoft and Apple each publish a JSON list of the prefixes their crawlers use. An address inside one is the operator; an address outside it is not. Anthropic's file is not where it might be guessed: it is at claude.com/crawling/bots.json, linked from a single sentence in a support article about blocking by address. Google splits its crawlers across three separate files, and checking only the first would report Google impersonating itself.

Reverse DNS, confirmed forward. Google, Microsoft, Apple, Amazon and Meta document a hostname suffix instead of, or as well as, a prefix list. The address is resolved to a name, the name is resolved back to an address, and both must agree. A one-way lookup is set by whoever holds the address and is not accepted here.

Registry ownership, as a last resort. Meta, ByteDance and You.com publish neither for their crawlers. For those, and only for those, the address is looked up in the regional registry and matched against the operator's own organisation name. This is the weakest of the three and it is labelled as such wherever it is used: it says who holds the address, not who sent the request.

OPERATOR	REQUESTS	ADDRESSES	VERIFIED ADDRESSES	REQUESTS VERIFIED	HOW
Googlebot	1380	53	34	1337 (97%)	published range and reverse DNS
ClaudeBot	1350	46	45	1311 (97%)	published range
meta-externalagent	1344	130	130	1344 (100%)	reverse DNS and registry ownership
Bingbot	1170	226	225	1169 (100%)	published range and reverse DNS
Applebot	788	215	215	788 (100%)	published range
GPTBot	669	24	23	668 (100%)	published range
OAI-SearchBot	179	52	51	178 (99%)	published range
CCBot	132	9	0	0 (0%)	none published
Bytespider	68	32	0	0 (0%)	reverse DNS and registry ownership
Amazonbot	66	55	55	66 (100%)	reverse DNS
Claude-User	9	3	0	0 (0%)	published range
ChatGPT-User	5	5	4	4 (80%)	published range
YouBot	4	2	2	4 (100%)	registry ownership
PerplexityBot	2	2	1	1 (50%)	published range
Claude-SearchBot	1	1	0	0 (0%)	published range
Google-Extended	1	1	0	0 (0%)	published range and reverse DNS
Applebot-Extended	1	1	0	0 (0%)	published range and reverse DNS
total	**7169**			**6870 (95.8%)**	

Six thousand eight hundred and seventy requests of 7,169 verified, 95.8 per cent. The 299 that failed are excluded from every number in this paper.

The check asks live sources, so it is not perfectly repeatable and should not pretend to be. A prefix file is republished, a reverse record resolves on one run and times out on the next, a registry rate-limits. Two runs a day apart moved the excluded count by four requests out of 7,169. The figures here are from the run of 22 August 2026, the date is recorded with every table, and a re-run will differ in the third digit.

Three things in that table are worth stating plainly.

Nineteen addresses claimed to be Googlebot and were in none of Google's three published lists, and eleven of them had no reverse record at all. That is the classic shape of a scraper wearing a search engine's name, and it is why the check exists. It is also only 43 requests out of 1,380.

Claude-User, Claude-SearchBot and Google-Extended did not verify at all, on 9, 1 and 1 requests respectively. On these volumes that is a curiosity rather than a finding: three addresses that are not in a published file, over sixteen months. The requests are excluded and the exclusion changes no conclusion.

CCBot and Bytespider cannot be checked. Common Crawl runs from a cloud provider whose registry entry says only that it is a cloud provider, and ByteDance publishes nothing that resolves. Their 200 requests are counted in no table here. That is a gap in this study, not an accusation against either crawler.

An unverified request is not proof of impersonation. It can equally be a range published after this window closed, a proxy the operator uses and does not list, or a file that moved. It is reported because the alternative is to count a claim and call it a measurement.

04

What they took

OPERATOR	AGENT	PAGES TAKEN	ALL REQUESTS	DATA SERVED
Anthropic	ClaudeBot, Claude-User, Claude-SearchBot	**1069**	1311	31.7 MB
Microsoft	Bingbot	**799**	1169	61.2 MB
Meta	meta-externalagent	**720**	1344	29.6 MB
OpenAI	GPTBot, OAI-SearchBot, ChatGPT-User	**453**	850	45.2 MB
Google search	Googlebot	**259**	1337	23.7 MB
Apple	Applebot-Extended, Applebot	**144**	788	24.8 MB
Amazon	Amazonbot	**22**	66	3.3 MB
You.com	YouBot	**2**	4	0.0 MB
Perplexity	PerplexityBot	**0**	1	0.1 MB
total		**3468**	**6870**	**219.7 MB**

Counting only addresses that survived the identity check, the crawlers took **3,466 pages and 219.7 MB** from a site whose entire content is a few dozen pages.

The ordering is the first surprise. **Anthropic took the most pages of anyone, 1,069**, more than twice OpenAI's 453 and more than the Bing crawler that feeds an actual search index. Meta, which has no consumer answer engine that cites sources, took 720.

The second surprise is how differently the same number of requests turns into pages. Apple made 788 requests and took 144 pages: it is fetching assets, images and stylesheets, at roughly five requests per page. Anthropic made 1,311 requests and took 1,069 pages, which is almost pure text. Whatever ClaudeBot is doing, it is not

rendering the site. It is reading it.

Bandwidth tells a third story again. Microsoft moved the most data, 61.2 MB across 1,169 requests, because Bingbot fetches images at full size. The operator that took the most content and the operator that cost the most bandwidth are not the same one, and a site owner who measures only bandwidth will conclude the wrong thing about who is taking their work.

How much of the site that is

Over the window the site served **474 distinct page paths** to anybody at all, which is the honest size of it: a few dozen written pages, plus the research whitepapers and their assets.

Against that, Anthropic's 1,069 page fetches are a little over two full passes of the site, OpenAI's 453 slightly under one, and the whole set of crawlers together about seven. None of that is misbehaviour. Content changes, papers get added, and a crawler that never revisits is a crawler with a stale index. It is simply the scale, for a site with no news section and no store, and it is the numerator of everything that follows.

05

What came back

DAY	ENGINE	ARRIVALS	A READER IN A BROWSER
01 Jun 2026	copilot.microsoft.com	1	no, and it was not even a served page
19 Aug 2026	gemini.google.com	1	yes

That is the complete list. Two arrivals in sixteen months carried an answer engine's referrer, and one of them was not a person: the Copilot request was not even a served page view, and its agent named no browser.

One reader arrived from an answer engine over the whole window, from Gemini, on 19 August 2026.

A single event proves nothing on its own, so the question is what the site's normal traffic looks like beside it.

REFERRING SITE	PAGE VIEWS IT SENT
youtube.com	1291
(an address rather than a hostname)	59
ghost-rider	36
google.com	33
bio-gel.eu	11
l.instagram.com	9
m.baidu.com	9
accounts.google.com	9
facebook.com	4
bing.com	3
every other referring site	**15**
all of them	**1479**

Page views from an agent naming a browser: 14129. Page views from everything else that is not a listed crawler, which is scanners, monitors, libraries and feed readers: 15290.

Two of those rows are not really sites. Referrers that are bare addresses are masked, under the same rule that keeps every other address out of this paper, and ghost-rider is not a hostname at all. Both are left in the table rather than quietly filtered, because a filter that removes whatever looks wrong is how a number becomes whatever its author expected.

With those aside, over the same window and by the same definition of a reader: **YouTube sent 1,291, Google search sent 33, and every answer engine on earth sent one.**

The comparison the crawl invites

The operators' own argument for crawling is distribution: the assistant reads your site, cites you, and the reader follows the citation. On this site, over sixteen months, that mechanism produced one visit.

The counter-argument is that the visits are unmeasurable, because assistants strip referrers or answer without a link at all. That is true and it cuts both ways: an operator that will not pass a referrer has also made its side of the trade unauditable, and a trade whose return cannot be measured is not a trade a site owner can consent to on the evidence. This site allowed the crawlers on the stated bet that being read was worth being trained on. Sixteen months later, the only measurable return is one visitor.

06

The ratio

Only operators that run a consumer answer engine can be given a ratio at all. A dataset collector or a general search crawler has no front door of its own to send anybody through, and giving it a divisor it never had would be arithmetic dressed as an argument. Those operators are reported in the load table and excluded here.

OPERATOR	PAGES TAKEN	VISITORS SENT BACK	PAGES PER VISITOR
OpenAI	453	0	**not one visitor**
Anthropic	1069	0	**not one visitor**
Google AI	0	1	a visitor, and no crawl that survived the check
Microsoft	799	0	**not one visitor**
You.com	2	0	**not one visitor**
every operator with a front door	**2323**	**1**	**2323 : 1**

> 2,323 pages taken, one reader returned. 2,323 to 1.

Per operator it is starker still, because the one reader did not come from either of the two operators that took the most:

- **OpenAI: 453 pages, not one visitor.**
- **Anthropic: 1,069 pages, not one visitor.**
- **Microsoft: 799 pages, not one visitor**, since its single referral was not a

reader.

- **Google's Gemini: one reader, and no crawl that survived the identity check.**

The last line is the odd one and it should not be smoothed over. The visitor from Gemini is real. The single request from Google-Extended came from an address in none of Google's three published files and is excluded. Google's crawling of this site happens almost entirely under Googlebot, 1,337 verified requests, which feeds search and is not attributable to the answer engine. So Gemini appears here as a front door with no measurable crawler behind it, which is a limitation of attributing engines to agents rather than a finding about Google.

What the ratio is not

It is not a claim that the crawlers owe this site traffic. Nothing was promised and no contract exists.

It is not a valuation. A page taken is not a page monetised, and a visitor returned is not a customer.

It is one site's answer to the question everybody argues about with no number at all: on a small site that deliberately allowed them, over sixteen months, the operators took a little over two thousand pages for each reader they sent back. The next section asks the same question of a second site that shares nothing with this one.

07

The second site

A portfolio is a strange thing to generalise from. It has no products, no customers arriving to buy, and its owner writes the research that the crawlers are partly there to collect. The obvious objection to everything above is that it is one site, and an unusual one.

So the whole instrument was run again, unchanged, on a second Greek site: a single-practitioner professional practice with a blog, a services page and online booking. It is a client's site, read with its owner's agreement on the condition that it is not identified, so it appears here only by category. It has never blocked a crawler and has no AI-specific rules in its robots.txt at all, so the default applies and everybody is allowed.

It is a genuinely different site: five times the human traffic, twice the pages, and a commercial rather than an editorial reason to exist.

What it looks like

PROPERTY	VALUE
lines parsed	343962
lines not matching the combined format	0
lines carrying a user agent	343962
lines carrying a referrer	223026
requests attributed to a known crawler	26646
of those, served with a 200 or 206	21580
of those, a page rather than an asset	10342
requests not attributed to a crawler	317316
arrivals from an answer engine	117
distinct page paths the site served to anybody	981

OPERATOR	AGENT	PAGES TAKEN	ALL REQUESTS	DATA SERVED
Google search	Googlebot	**2137**	6712	212.3 MB
Microsoft	Bingbot	**2104**	4314	252.3 MB
Anthropic	ClaudeBot, Claude-User, Claude-SearchBot	**1930**	6947	249.2 MB
OpenAI	GPTBot, OAI-SearchBot, ChatGPT-User	**1611**	2400	116.3 MB
Apple	Applebot-Extended, Applebot	**666**	2136	48.9 MB
Amazon	Amazonbot	**342**	861	66.2 MB
Meta	meta-externalagent	**264**	471	45.8 MB
Perplexity	PerplexityBot, Perplexity-User	**55**	72	2.1 MB
total		**9109**	**23913**	**993.2 MB**

9,109 pages and 993.2 MB, four times the portfolio's haul, from a site with roughly twice as many pages. The ordering is different too. Here the two general search crawlers lead, Google at 2,137 pages and Bing at 2,104, with Anthropic third at 1,930 and OpenAI fourth at 1,611. On the portfolio Anthropic led everybody.

What came back

OPERATOR	PAGES TAKEN	VISITORS SENT BACK	PAGES PER VISITOR
OpenAI	1611	9	179 : 1
Anthropic	1930	0	**not one visitor**
Perplexity	55	0	**not one visitor**
Microsoft	2104	0	**not one visitor**
every operator with a front door	**5700**	**9**	**633 : 1**

OpenAI is the only operator in this study that sent anybody a meaningful number of readers: nine, all from ChatGPT, across two days in July 2025. That is the whole of the answer-engine referral traffic to this site in sixteen months.

DAY	ENGINE	ARRIVALS	A READER IN A BROWSER
16 Jul 2025	chatgpt.com	53	yes
19 Jul 2025	chatgpt.com	64	yes

Two days, thirteen months ago, and nothing since. Whatever put this practice into ChatGPT's answers in July 2025 either stopped or stopped sending anybody.

Against that, ordinary search is not subtle about the difference:

REFERRING SITE	PAGE VIEWS IT SENT
google.com	4484
google.gr	82
bing.com	29
m.facebook.com	13
facebook.com	6
duckduckgo.com	6
search.brave.com	4
l.instagram.com	3
search.yahoo.com	3
baciakte.online	3
every other referring site	**31**
all of them	**4664**

Page views from an agent naming a browser: 44623. Page views from everything else that is not a listed crawler, which is scanners, monitors, libraries and feed readers: 71674.

Google search sent 4,484 readers over the same window. Every answer engine combined sent nine.

The two sites together

	PAGES TAKEN	READERS RETURNED	PAGES PER READER
broikos.gr, the author's own portfolio and research site	2323	1	**2323 : 1**
a second Greek site, a Greek single-practitioner professional practice with a blog and online booking	5700	9	**633 : 1**

633 to 1 and 2,323 to 1. Two sites that share nothing except a country and a decision never to block anybody, and the ratio lands within one order of magnitude on both.

The practice site does better, and the reason is worth stating precisely: not because less was taken from it, but because one operator sent it nine readers. Remove those nine and it is the portfolio's number again, which is to say no number at all.

08

Thirteen months apart

The archive holds two windows that can be compared honestly: July 2025 across 25 days, and August 2026 across 21. Same site, same log format, same robots policy, same extraction code, and the counts below use verified addresses only.

	JUL 2025	AUG 2026	CHANGE
days observed	25	21	
crawler requests	974	5170	5.3 times
pages taken	255	2761	10.8 times
pages taken per day	**10**	**131**	**12.9 times**

Pages taken per day rose from 10 to 131, a factor of 12.9. Requests rose by a smaller factor, 5.3, which says the crawlers did not merely arrive more often: they shifted towards text and away from assets. More of every visit is now content.

Two caveats on that comparison, both of which cut against the headline rather than for it.

The site grew. Between the two windows it gained the research whitepapers, which are long text pages of exactly the kind a crawler wants, so part of the rise is supply rather than appetite. Against that, 474 distinct page paths is not a large site in either window, and the pages taken in August 2026 are roughly six times the whole site.

August 2026 is the month this paper was written in, which means the site was being edited, published to and linked from during the window. A live month is not a representative month.

What the comparison does support is a direction, and the direction is not subtle: **over thirteen months the crawl on this site grew by an order of magnitude while the referrals stayed at zero and then one.**

09

What I am not claiming

Not that this generalises to your site. One site, one country, one subject area, one owner. A storefront, a news site or a forum would have different content, different crawl appetite and a completely different referral profile. The method generalises; the number does not.

Not that the referrals are fully measurable. Assistants strip referrers, answer without linking, and are increasingly used inside applications that pass no referrer at all. A reader who read this site inside a chat window and never clicked is invisible here and always will be. The honest form of the headline is: *the measurable return was one visitor*, and the unmeasurable return is unknown in both directions.

Not that the window is complete. The hosting panel keeps a rolling archive, so six of sixteen months survive and two of them substantially. Nothing here is a sixteen-month total; it is a 58-day observation spread across sixteen months, and every rate in the paper is per observed day.

Not that unverified means fake. Three hundred and three requests failed the identity check and are excluded. Some of them are certainly scrapers wearing a crawler's name. Others are ranges published after this window, proxies an operator does not list, or files that moved. The paper excludes them rather than accusing them.

Not a claim about Common Crawl or ByteDance. Neither publishes anything that can be checked, so their 200 requests appear in no table here. Their absence from the totals is a gap in the instrument, not evidence about their behaviour.

Not a claim that the trade is bad. This site allowed the crawlers on purpose and would do it again: being read by assistants may pay off in ways a referrer header cannot record, including the one that matters most to a portfolio, which is being named when somebody asks an assistant for a developer. What this paper establishes is only that the leg of the trade that *can* be counted has, so far, a value of one.

Limitations

One site, one owner, and it is the author's own. That is what made the logs readable without asking anybody. It is also the study's narrowest dimension, and the obvious extension, a Greek storefront with real commercial traffic, needs its owner's consent and has not been asked for.

A reader is inferred, not observed. A browser-shaped user agent with an external referrer is the best available proxy for a person, and it is a proxy. A headless browser is counted as a reader; a person using a text browser is not.

The page and asset split is by file extension. A page served from a path that ends in .js would be miscounted as an asset. No such path exists on this site, but the rule is crude and it is stated.

Bot detection cuts both ways. The baseline of 14,129 browser-shaped page views certainly contains automation that names a browser, and it certainly contains the author's own visits. It is an over-count of outside human traffic, which makes the comparison in the referral section conservative in the wrong direction: the real gap between YouTube and the answer engines is what matters, and both were counted the same way.

10

Reproducing this

The whole path is published except the logs themselves, which carry visitor addresses and never leave the machine they were pulled to.

SCRIPT	WHAT IT DOES	DOES IT TOUCH A RAW LOG LINE
log_inventory.py	lists what archives each host holds, downloads nothing	no
pull_logs.py	fetches archives into raw/, which is never published	it writes them
extract.py	the only reader of raw lines; emits aggregates only	yes
verify_agents.py	checks each crawler address against the operator's own published ranges, hostnames and registry entries; holds addresses in memory and writes counts	yes
ratio.py	pages taken per visitor returned, from the CSVs alone	no
tables.py	every table in this paper, from the CSVs alone	no
integrity.py	re-derives each headline figure from the CSVs and fails if the prose disagrees	no

The published aggregates are crawlers.csv as the agent strings claimed, crawlers_verified.csv after the identity check, referrals.csv, traffic.csv, verification.csv and format.md. None contains an address, a user agent string or a full referrer URL.

To recompute the headline from scratch you need the raw archives, which means you need your own site. To recompute it from this paper's data you need the two CSVs and ratio.py, and that path holds no personal data at all. That separation is deliberate: the claim should be checkable by someone who is not allowed to see the logs.

The operators' published sources, fetched at analysis time rather than transcribed:

- openai.com/gptbot.json, openai.com/searchbot.json, openai.com/chatgpt-user.json
- claude.com/crawling/bots.json
- perplexity.com/perplexitybot.json
- developers.google.com/static/search/apis/ipranges/ for googlebot.json, special-crawlers.json and user-triggered-fetchers-google.json
- bing.com/toolbox/bingbot.json
- search.developer.apple.com/applebot.json
- reverse DNS suffixes as documented by Google, Microsoft, Apple, Amazon and Meta

11

Claim ledger

Every factual claim in this paper, with where it came from and whether it was checked. The ledger is the contract: a claim not in it is not made, and a claim in it can be recomputed by anybody holding the published aggregates.

Two of the entries below are defects in my own instrument, found while adding the second site and recorded rather than quietly fixed. One of them wrote a hostname this paper had promised not to publish into a published table.

integrity.py re-derives each figure from the CSVs and fails if the assembled paper no longer contains it, so a number cannot survive the data changing under it. It also scans every published file for an address and for the protected hostname, which is the check that did not exist when the leak happened.

ID	CLAIM	SOURCE
C001	The site's robots.txt names five AI crawlers and allows them deliberately, with the reason stated in a comment, so this is the consenting case rather than the adversarial one	sources/robots.txt
C002	Twelve monthly log archives were available going back to September 2024, disproving the assumption that the logs had rotated away, which had parked this paper for a year	data/log_inventory.py
C003	109,107 log lines were parsed and none failed to match the combined log format	data/sites/broikos.gr/format.md
C004	The window is 58 days carrying crawler traffic between April 2025 and August 2026, of which two months are substantially complete: July 2025 at 25 days and August 2026 at 21	data/sites/broikos.gr/crawlers.csv
C005	6,870 of 7,169 crawler requests, 95.8 per cent, came from an address the operator publishes, a hostname it documents, or a registry entry in its own name	data/sites/broikos.gr/verification.csv
C006	19 of 53 addresses claiming to be Googlebot appear in none of Google's three published range files and 11 of those have no reverse record at all	data/sites/broikos.gr/verification.csv
C007	CCBot and Bytespider publish nothing that can be checked, so their 200 requests appear in no table in this paper	data/sites/broikos.gr/verification.csv
C008	Across verified addresses the crawlers took 3,466 pages and 219.7 MB during the window	data/sites/broikos.gr/crawlers_verified.csv
C009	Anthropic took the most pages of any operator, 1,069, more than twice OpenAI's 453	data/sites/broikos.gr/verification.csv
C010	Apple took 144 pages across 788 requests while Anthropic took 1,069 across 1,311: the same request count converts to pages at very different rates	data/sites/broikos.gr/verification.csv
C011	Microsoft moved the most data at 61.2 MB, so the operator that took the most content and the operator that cost the most bandwidth are not the same one	data/sites/broikos.gr/crawlers_verified.csv
C012	The site served 474 distinct page paths to anybody during the window, so the crawlers together took about seven full passes of it	data/sites/broikos.gr/format.md
C013	Exactly two requests in the whole window carried an answer engine's referrer, and only one was a served page view from an agent naming a browser	data/sites/broikos.gr/referrals.csv
C014	Over the same window and by the same definition of a reader, YouTube sent 1,291 page views and Google search sent 33	data/sites/broikos.gr/traffic.csv
C015	Operators running a consumer answer engine took 2,323 verified pages and returned one reader, a ratio of 2,323 to 1	data/ratio.py
C016	OpenAI took 453 pages and sent no visitor; Anthropic took 1,069 and sent no visitor; Microsoft took 799 and sent no reader	data/sites/broikos.gr/verification.csv
C017	The one Gemini reader has no verified Google-Extended crawl behind it, because Google's crawling of this site happens under Googlebot	data/sites/broikos.gr/verification.csv
C018	Between July 2025 and August 2026 pages taken per observed day rose from 10 to 131, a factor of 12.9, while requests per day rose by a factor of 5.3	data/sites/broikos.gr/crawlers_verified.csv

ID	CLAIM	SOURCE
C019	Part of the rise is supply: the site gained its research whitepapers between the two windows, and August 2026 is a month in which the site was being edited and published to	data/sites/broikos.gr/crawlers_verified.csv
C020	No address, user agent string or full referrer URL appears in any published file from this paper	data/extract.py
C021	A second Greek site, a single-practitioner professional practice read with its owner's agreement and not identified, was measured with the same code over the same 58-day window	data/sites/practice-a/
C022	On the second site 23,913 of 26,646 crawler requests verified, 89.7 per cent, and the crawlers took 9,109 pages and 993.2 MB	data/sites/practice-a/verification.csv
C023	On the second site the two general search crawlers led, Google at 2,137 pages and Bing at 2,104, with Anthropic third at 1,930 and OpenAI fourth at 1,611	data/sites/practice-a/verification.csv
C024	OpenAI sent the second site nine readers, all from ChatGPT, across two days in July 2025 and none since	data/sites/practice-a/referrals.csv
C025	Operators with a front door took 5,700 pages from the second site and returned nine readers, 633 to 1, against 2,323 to 1 on the portfolio	data/ratio.py
C026	Over the same window ordinary Google search sent the second site 4,484 readers while every answer engine combined sent nine	data/sites/practice-a/traffic.csv
C027	DEFECT: the first version of the anonymity mask wrote the client site's hostname into a published table, through a self-referral, a cPanel preview URL and a staging subdomain	data/extract.py
C028	DEFECT: duckduckgo.com was coded as an answer engine, which would have credited the answer engines with seven search referrals they did not send	data/extract.py
C029	The identity check queries live sources, so it is not perfectly repeatable: two runs a day apart moved the excluded count by four requests out of 7,169	data/verify_agents.py

ABOUT THE RECORD

Two Greek sites, 58 days of archived access logs each, every crawler address checked against the operator own published ranges, documented hostnames or registry entries.

WHAT IT SHOWS

Operators running a consumer answer engine took 2,323 pages from one site and 5,700 from the other, and returned one reader and nine. Ordinary search sent 4,484 to the same second site.

CORRECTIONS AND CONFLICTS

One site is the author own and is named. The second is a client read with its owner agreement and is not identified. Raw logs carry visitor addresses and are never published; two defects found while anonymising are in the ledger.