



CENSUS · BA-D-2026-05

The JavaScript wall at the university front door: a raw-versus-rendered census of homepage text

AI crawlers that supply answer engines largely do not execute JavaScript, so the text they read is what the raw HTML holds before any script runs.

DOCUMENT	BA-D-2026-05
SERIES	Census
DATE	2026
SUBJECTS	Crawlers & access; Measurement & statistics
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

AI crawlers that supply answer engines largely do not execute JavaScript, so the text they read is what the raw HTML holds before any script runs. This census measures how much of a university homepage's visible text exists only after client-side rendering. For the 429 analyzable homepages of companion census BA-D-2026-02, it compares the stored raw HTML's visible text (what a non-rendering crawler saw on 2026-07-09) against `document.body.innerText` from a headless-Chrome visit to the same URL (2026-07-10). Of the 400 that rendered cleanly, 354 (88.5%) were within 10% of their rendered text and only 16 (4.0%) at least doubled it; just 6 (1.5%) were hard walls, near-empty without JavaScript. Decomposing those six shows only two are genuine client-side rendering — the rest are bot-challenge shells or JavaScript redirects — so a 'universities are single-page apps' reading is unsupported. Findings span two adjacent days and should be assumed perishable.

IN BRIEF

When a prospective student asks an AI assistant about a university, the assistant's answer is assembled from what a crawler could read on the university's site — and most crawlers behind AI answers do not run JavaScript, so they see only the raw HTML the server sends, not the fully drawn page a browser shows. This census asks a narrow question of 429 university homepages: how much of the visible text appears only after JavaScript runs, and is therefore invisible to a non-rendering crawler? For the 400 homepages that a headless browser rendered cleanly, the answer is mostly 'very little.' Nearly nine in ten (354 of 400) showed essentially the same amount of text before and after rendering, and only six were near-empty without JavaScript. But those six are the interesting part, and they are not what they first look like: only two are genuinely built as client-rendered apps. Two others were empty for the crawler because a security service returned a challenge page to the crawler's non-browser identity, and two more because the homepage used JavaScript to redirect elsewhere. A mirror-image effect cuts the other way: on eleven homepages the headless browser itself was turned away with a block page that the previous day's simpler fetch had sailed through. The lesson for anyone measuring AI legibility is that 'what a browser sees minus what a crawler sees' is often a story about access control, not about rendering — and it runs in both directions. This is a two-day snapshot; sites change without notice, and every row is in the open dataset.

KEY FINDINGS

1. Across the 400 university homepages that a headless browser rendered cleanly, 354 (88.5%) had raw HTML text within 10% of their rendered text and only 16 (4.0%) at least doubled their visible text after JavaScript ran — in the server-delivered HTML a non-rendering crawler reads, most homepages are effectively already text-complete, in a frame built to over-represent internationally active universities.
2. Only 6 of the 400 (1.5%) were hard walls — near-empty raw HTML (under 200 characters) drawn out to a substantial rendered page (at least 1,000 characters) — and their decomposition is the census's central caution: just 2 of the 6 (uga.edu, msu.ru) are genuine client-side rendering, while 2 are bot-challenge shells that a security service served to the crawler's user agent (msu.edu, cityu.edu.hk) and 2 are client-side JavaScript redirects (cuhk.edu.hk, uitm.edu.my). A 'share of universities that are single-page apps' figure cannot be read off the hard-wall count.
3. At the median the gap runs the other way: across the 400 clean renders the median raw-over-rendered ratio is 1.35 and raw exceeds rendered by a median of 1,446 characters, and 321 of the 400 (80.2%) showed at least as much text in the raw parse as in the browser's rendered text — because the browser's innerText excludes CSS-hidden navigation and footers that the raw parser counts, a definitional difference between the two extractors rather than JavaScript removing content.
4. The bot/user-agent confound is bidirectional and is itself a first-order finding: 11 of the 429 renders (quarantined from the headline) returned a 403 or interstitial to the datacenter-IP headless browser — uci.edu, tcu.edu, dtu.dk, helsinki.fi, devry.edu and six others — even though the previous day's non-browser fetch had retrieved a full page, so the difference between what a crawler and a browser can read is frequently a question of who is being let in, not of client-side rendering.
5. Headline rates are computed only over the 400 homepages with a trustworthy render (rendered_status ok); the 17 navigation timeouts, 11 suspected challenge blocks, and 1 hard error carry an untrustworthy rendered-text count and are reported and excluded separately, and the raw-versus-rendered bytes were captured about one day apart (raw 2026-07-09, rendered 2026-07-10).

CONTENTS

1	Summary	4
2	Scope and method	4
3	Results — how much text is JavaScript-only	6
4	Results — the six hard walls, and what actually causes them	7
5	The mirror image: render-side blocking	8
6	What this means for reading AI legibility	9
7	Reproducibility appendix	9
8	Limitations	10

1 Summary

The companion census BA-D-2026-02 read the raw, unrendered HTML of a frame of 500 university homepages and found the international-targeting layer largely absent. It read that HTML for a reason stated in the maturity model of BA-C-1: the crawlers that feed AI answer engines largely do not execute JavaScript, so the content they can read is the content present before any script runs [4]. That claim is load-bearing for the whole raw-HTML method, and this census tests the quantity behind it directly. For the same 429 homepages that returned an analyzable page on 2026-07-09, it measures how much of each homepage’s visible text exists *only* after client-side rendering — the text a non-rendering crawler cannot see.

The method is a magnitude comparison. The RAW figure is the visible text of the stored 2026-07-09 HTML, recomputed with the exact extractor BA-D-2026-02 used, so it reproduces that census’s own character counts. The RENDERED figure is `document.body.innerText` from a headless-Chrome visit to the same final URL on 2026-07-10. Both strings are collapsed by one identical operation before counting, so the two are directly comparable, and the question asked of each homepage is narrow: is the rendered text many times larger than the raw text?

For most homepages it is not. Of the 400 homepages that rendered cleanly, 354 (88.5%) had raw text within 10% of their rendered text, and only 16 (4.0%) at least doubled their visible text under JavaScript. Just 6 (1.5%) were hard walls — near-empty in raw HTML, substantial once rendered. That small count is the headline, but the six hard walls are also the census’s central caution, because only two of them are genuinely client-rendered pages. The other four are near-empty to a crawler for reasons that have nothing to do with single-page-app architecture: two are bot-challenge shells a security service returned to the crawler’s user agent, and two are client-side JavaScript redirects. And the confound is bidirectional — on eleven other homepages it was the headless browser that was blocked, returning a near-empty block page where the previous day’s simpler fetch had retrieved a full one. The measurement is a two-day snapshot of homepages that change without notice; results should be assumed perishable.

2 Scope and method

The frame and the two texts

The population is the analyzed set of BA-D-2026-02: the 429 homepages, out of a documented frame of 500 universities, that returned an HTTP 200 with a non-empty body on 2026-07-09. That frame is 300 U.S. institutions ordered by enrollment from the College Scorecard [1] and 200 non-U.S. universities from the Hipolabs university-domains-list [2] restricted to and ordered by the Tranco research ranking [3]; its construction, and the reasons it over-represents larger and higher-traffic institutions, are documented in full in that census and are not re-derived here. This census adds a second reading of each of those 429 homepages and compares the two.

- **RAW** is the visible text a non-rendering crawler saw: the stored, unrendered HTML bytes BA-D-2026-02 fetched on 2026-07-09, re-parsed with that census’s exact extractor. Recomputing RAW for all 429 reproduced the stored character counts with zero mismatches, so RAW here is literally the earlier census’s `visible_text_chars`.
- **RENDERED** is `document.body.innerText` from a live headless-Chrome visit to the same final URL on 2026-07-10, captured with a real browser user agent. `innerText` is the browser’s rendered, CSS-

aware text of an element [6]; it is the reproducible proxy this census uses for “visible text a browser draws.”

Both strings pass through one identical whitespace-collapse before counting, and every count, ratio, and classification is computed in one place, so RAW and RENDERED share exactly the same counting semantics. For each homepage the census records `raw_chars`, `rendered_chars`, their ratio (raw over rendered), their absolute difference, the render’s outcome, and two raw-side signals discussed below.

Why raw-versus-rendered is the relevant lens

The reason this comparison matters is the same reason BA-D-2026-02 read raw HTML at all. In the Barkhausen Ladder (BA-C-1), rendering and crawler access is level BL-1, and its criterion is explicit: an entity’s primary content must be present in the initial HTML response without client-side JavaScript execution, because public analyses indicate that AI crawlers largely do not execute JavaScript, and content that exists only after client-side rendering is in practice invisible to them [4]. The crawler taxonomy BA-C-6 draws the same boundary across its functional classes: training, retrieval, and user-fetch crawlers are documented as fetching pages, not as driving a rendering engine. This census measures, per homepage, exactly the quantity BL-1 is about — how much visible text lives on the far side of a JavaScript execution that a non-rendering crawler will not perform. The whitepaper BA-W-2026-03 frames the same terrain as three different “readers” of a page — the raw HTML a non-rendering fetch receives, the rendered page a browser draws, and the accessibility tree assistive technology and some agents consume — and this census is a direct measurement of the gap between the first two.

Operationalizations, fixed before the batch

The classification thresholds were fixed in the analysis code before the 429 homepages were scored, and are applied uniformly. Let `raw` and `ren` be the RAW and RENDERED character counts and `ratio = raw / ren`.

- **js-light** — `ratio ≥ 0.9`. Raw and rendered are within 10%; the homepage is effectively text-complete for a non-rendering crawler. This band intentionally also contains the common case `ratio > 1` (raw larger than rendered), discussed below.
- **js-dependent** — `ren > 0` and `raw < 0.5 × ren`. The homepage’s visible text at least doubles once JavaScript runs, so a non-rendering crawler sees under half of it.
- **hard-wall** — `raw < 200` and `ren ≥ 1000`. The unrendered page is essentially empty — below the 200-character floor the parent census used — while the rendered page is substantial. A crawler that does not execute JavaScript sees nothing usable. Every hard wall is by construction also js-dependent.

Two further definitions govern which renders are trustworthy enough to count. A render is marked **challenge-suspected** when it navigated but returned under 200 characters *and* either its page title matches a known bot-challenge or WAF-block marker (for example “403 Forbidden”, “Access Blocked”, “Request Rejected”) or the stored raw page for that domain was itself substantial — the second arm catches a render-side block whose block page carries an innocuous title. A render is marked **timeout** when navigation hit the 45-second limit (a best-effort partial body is still captured) and **error** on a hard navigation or script failure. Headline rates are computed only over homepages with a clean render (`rendered_status ok`); the challenge-suspected, timeout, and error renders carry an untrustworthy `rendered_chars` and are reported and excluded separately. On challenge-suspected and error renders the three classification booleans are set to null rather than allowed to fire on a body that is not a valid measurement.

Finally, two signals are recorded from the raw bytes to explain *why* a raw page is empty when it is: `raw_incapsula_waf`, set when the stored HTML carries an Imperva/Incapsula challenge script, and `raw_client_redirect`, set when it carries an inline `location.*` assignment or a meta refresh. The second is a broad signal — analytics and consent snippets also call `location.*` — so it marks a redirect wall only when read together with a near-empty raw count, and the raw count of 50 flagged domains is not a count of 50 redirect walls.

3 Results — how much text is JavaScript-only

Of the 429 homepages, 400 rendered cleanly, 17 timed out, 11 are suspected challenge blocks, and 1 hard-errored. The 400 clean renders are the denominator for every rate here; the other 29 are taken up in the sections that follow.

Within those 400, the dominant result is that the raw HTML a non-rendering crawler reads is, for text purposes, already the page. 354 homepages (88.5%) are js-light — their raw and rendered text are within 10%. Only 46 (400 – 354) gained meaningful text on rendering, and the gain is concentrated: 16 (4.0%) are js-dependent, at least doubling their text under JavaScript, and within those, 6 (1.5%) are hard walls. The other 30 sit between, adding text without doubling it. For a frame deliberately built to over-represent internationally active, higher-traffic universities, JavaScript-dependence of the homepage’s *text* is the exception, not the rule.

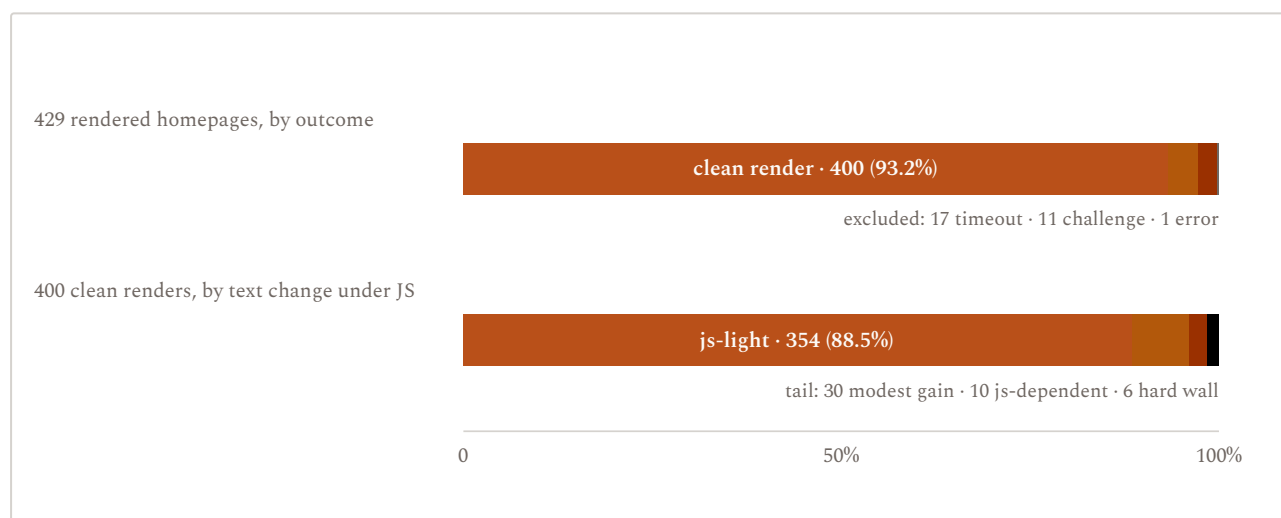


Figure 1. How the render pass resolves. Top: the 429 homepages by render outcome — 400 rendered cleanly and are the rate denominator; 17 timed out, 11 are suspected challenge blocks, and 1 hard-errored, and those 29 are excluded from the rates because their rendered-text count is not a trustworthy measurement. Bottom: the 400 clean renders composed by how much their visible text changed under JavaScript — js-light (raw within 10% of rendered), a modest-gain middle band, js-dependent (text at least doubles), and hard-wall (near-empty raw drawn out to a substantial page). The hard-wall segment is a subset of js-dependent, shown split out.

BA-D-2026-05, dataset js-rendering-2026 · n = 429 rendered (400 clean) · 2026-07-09 raw / 2026-07-10 rendered

The ratio usually exceeds one, and that is not JavaScript removing text

A reader expecting rendering to *add* text will be surprised that at the median it subtracts. Across the 400 clean renders the median raw-over-rendered ratio is 1.35, and at the median raw exceeds rendered by 1,446 characters; 321 of them (80.2%) have a ratio of at least 1 — the raw parse counted at least as many characters as the browser’s rendered text. This is a definitional difference between the two extractors, not evidence that JavaScript stripped content. The raw parser counts every text node that is not inside a script, style, template, or head element, regardless of CSS. The browser’s `innerText`, by contrast, reflects

rendered appearance and excludes text that is CSS-hidden at capture time — collapsed mega-menus, off-screen navigation, `display:none` accordions, cookie panels not yet shown [6]. A homepage with a large hidden navigation and footer therefore shows more raw text than rendered text. Because the classifications that matter here — js-dependent and hard-wall — fire only when raw is *much smaller* than rendered, they are unaffected by this asymmetry: a ratio > 1 homepage is unambiguously not JavaScript-walled, and the js-light band is defined to include it. The median above one is the expected shape of a set of mostly server-rendered pages, and it is reported so that no reader mistakes “rendered larger” for the common case when it is the rare one.

4 Results — the six hard walls, and what actually causes them

Six homepages are hard walls: near-empty in the HTML a non-rendering crawler receives, substantial once a browser runs their JavaScript. Taken bare, that count invites a headline it cannot support — “1.5% of universities are single-page apps.” The per-domain raw-side signals show why that reading is wrong. Only two of the six are genuinely client-rendered homepages. The other four are empty to the crawler for reasons unrelated to single-page-app rendering: two were served a bot-challenge shell in place of the page, and two use JavaScript to redirect the browser elsewhere. The census names the six here because the mechanism, not the institution, is the point, and the tone is descriptive: each row is a fact about what the crawler received, not a judgment of the site.

HOMEPAGE	RAW CHARS	RENDERED CHARS	RAW-SIDE SIGNAL	WHAT THE EMPTY RAW PAGE IS
uga.edu	0	2,225	none	Genuine client-side rendering — the raw HTML is a script shell with no visible text
msu.ru	0	1,677	none	Genuine client-side rendering — the raw HTML is a script shell with no visible text
msu.edu	0	5,228	Incapsula challenge	A 212-byte Imperva/Incapsula challenge shell was returned to the crawler’s user agent
cityu.edu.hk	0	6,917	Incapsula challenge	A 212-byte Imperva/Incapsula challenge shell was returned to the crawler’s user agent
cuhk.edu.hk	110	1,793	client redirect	A near-empty page whose script redirects the browser to the live homepage
uitm.edu.my	0	7,228	client redirect	A near-empty page whose script redirects the browser to the live homepage

The distinction is not pedantic. For `uga.edu` and `msu.ru` the finding is exactly BL-1: the homepage’s text is on the far side of a client-side render, so a non-rendering crawler sees none of it, and the fix is server-side rendering. For `msu.edu` and `cityu.edu.hk` the homepage is ordinary server-rendered HTML that a browser reaches fine; what the crawler hit was a security product returning a challenge to a non-browser user agent, so the “wall” is an access-control decision at the edge, not a rendering choice — and it is the same mechanism that put 33 homepages entirely out of the analyzed frame in BA-D-2026-02 as 403 responses. For `cuhk.edu.hk` and `uitm.edu.my` the raw page is a redirect stub that hands the browser to the real homepage in JavaScript; a crawler that does not run the redirect stays on the stub. Three different causes, one surface appearance. The honest statement the data supports is that of 400 cleanly rendered homepages, two are demonstrably client-rendered at the front door; the hard-wall rate is a ceiling on client-side rendering, not an estimate of it.

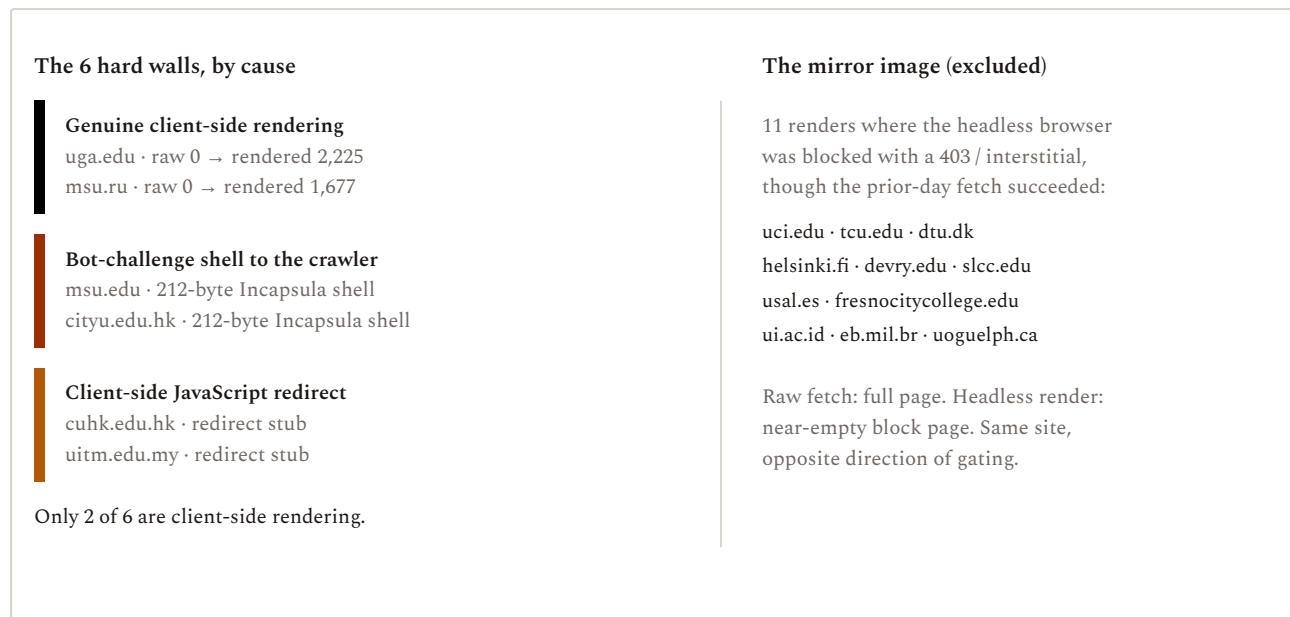


Figure 2. The six hard walls decomposed by cause, and the render-side mirror image. Left: the six near-empty-without-JavaScript homepages split by what their empty raw page actually is — two genuine client-side renders, two bot-challenge shells returned to the crawler's user agent, two client-side JavaScript redirects. Only the first pair is client-side rendering. Right: the mirror-image confound — 11 homepages (excluded from the headline) where the datacenter-IP headless browser was itself blocked with a 403 or interstitial, though the previous day's non-browser fetch retrieved a full page. In both directions, 'crawler sees less than browser' is frequently access control, not rendering.

BA-D-2026-05, dataset js-rendering-2026 · 6 hard walls · 11 render-side blocks · 2026-07-09 raw / 2026-07-10 rendered

5 The mirror image: render-side blocking

The four non-rendering causes above all sit on the raw side — the crawler's earlier fetch was gated or redirected. The render side carries the same confound in the opposite direction, and it is large enough to be a finding in its own right. Eleven homepages navigated on 2026-07-10 but returned under 200 characters — a 403 page, an "Access Blocked" or "Request Rejected" interstitial, or a near-empty body under an innocuous title — even though BA-D-2026-02 had retrieved a full, substantial page from the same domain the day before. The list includes uci.edu, tcu.edu, dtu.dk, helsinki.fi, devry.edu, slcc.edu, usal.es, fresnocitycollege.edu, ui.ac.id, eb.mil.br, and uoguelph.ca. Here the datacenter-IP headless browser was turned away while the previous day's ordinary httpx fetch was not — the reverse of the msu.edu and cityu.edu.hk pattern, where the browser was let through and the crawler was challenged. These eleven are quarantined as challenge-suspected and excluded from every rate, because their near-empty rendered body measures a block, not a homepage.

Read together, the raw-side and render-side blocks make the census's methodological point concrete: the difference between what a non-rendering crawler and a rendering browser see is, on a meaningful share of sites, produced by access-control systems that treat the two clients differently — by user agent, by IP reputation, by challenge — and not by client-side rendering at all. This matters for anyone trying to measure AI legibility by diffing a raw fetch against a browser render, because the naive diff attributes to "JavaScript" a gap that is often about who is being let in. A robots.txt policy, in the sense of RFC 9309, is a published request a crawler is asked to honor and is explicitly "not a form of access authorization" [7]; the gating seen here is the enforced kind — a WAF or IP filter returning a challenge — and it operates whether or not any robots rule is present. The two mechanisms are distinct, and only the enforced one produces the near-empty bodies quarantined here.

Two further renders belong to this same family without entering the eleven. `smu.edu.sg` had an empty raw page — one of the same 212-byte Incapsula shells as `msu.edu` and `cityu.edu.hk` — and a headless render that reached 7,140 characters but only after the navigation timed out, so it is recorded as a timeout and excluded from the headline on the conservative rule that a timed-out body may be partial. `usu.edu` is the single hard error, an HTTP/2 protocol error mid-navigation. Neither is counted as a js-wall, and both are in the open dataset for inspection.

6 What this means for reading AI legibility

For the two genuinely client-rendered homepages, the finding is the one BL-1 anticipates: text that exists only after a client-side render is, for the non-rendering crawlers that feed answer engines, not there. For the frame as a whole, the reassuring-sounding version — “88.5% of homepages are already text-complete for a crawler” — is true as stated but narrower than it sounds. It is a statement about the *quantity* of visible text on the homepage, not about the international-targeting and structured-data signals BA-D-2026-02 found missing in that same raw HTML; a page can deliver all its body text without JavaScript and still omit the hreflang, canonical, and schema.org markup that make that text legible as a specific entity’s. Text-completeness at BL-1 is necessary, not sufficient, and it sits below the structured-data rung (BL-2) in the same ladder.

The sharper contribution is the caution the six hard walls and the eleven render-side blocks deliver together. A single-number “share of sites that require JavaScript” is not safely readable from a raw-versus-rendered diff, because the diff conflates three things that look identical at the surface — client-side rendering, edge bot-challenges, and JavaScript redirects — and because the measurement apparatus itself (a datacenter-IP headless browser versus a prior non-browser fetch) is gated differently by the sites it measures, in both directions. The defensible claims are narrow and count-level: in this frame, on these two days, two homepages were demonstrably client-rendered at the front door, a handful more were gated or redirected in ways a non-rendering crawler cannot follow, and the large majority delivered their homepage text in server HTML. The open dataset carries every row, every signal, and the rendered-text hash, so any reader can re-derive these counts and inspect any domain directly.

7 Reproducibility appendix

The frame, the two texts, and the classification are specified to permit replication.

Frame. The 429 homepages with `analyzed = true` in the BA-D-2026-02 dataset — an HTTP 200 with a non-empty body on 2026-07-09 — drawn from the 500-university frame documented there (300 U.S. institutions by enrollment from the College Scorecard [1]; 200 non-U.S. universities from the Hipolabs list [2] restricted to the Tranco ranking [3]).

RAW. The visible text of each homepage’s stored 2026-07-09 HTML bytes, re-extracted with the parent census’s parser and identical whitespace collapse. Recomputing RAW for all 429 reproduced the stored `visible_text_chars` with zero mismatches, so RAW equals what the earlier non-rendering fetch saw.

RENDERED. A single headless-Chrome visit per homepage to the same final URL on 2026-07-10, driven by Puppeteer over the Chrome DevTools Protocol [5], in one shared browser with a fresh page per site and at most six concurrent pages. Requests for images, fonts, and media were blocked to speed the pass; stylesheets, scripts, and XHR were allowed, because `innerText` depends on CSS visibility [6] and blocking CSS would corrupt the quantity measured. Navigation waited for network-idle with a 45-second timeout followed by a fixed 3-second settle, then captured `document.body.innerText` and the page title. Each homepage was rendered once, with no retry; transient failures land in timeout or error rather than being

re-attempted. The render window was 2026-07-10T00:41:04Z to 00:51:19Z; median render time was 6.4 seconds.

Classification. RAW and RENDERED were collapsed identically and counted, and each homepage was labeled js-light ($\text{ratio} \geq 0.9$), js-dependent ($\text{ren} > 0$ and $\text{raw} < 0.5 \times \text{ren}$), and hard-wall ($\text{raw} < 200$ and $\text{ren} \geq 1000$) by the thresholds fixed before the batch. Rates were computed over the 400 homepages with a clean render; challenge-suspected, timeout, and error renders were reported and excluded separately, and their classification booleans nulled where the body is not a valid measurement. The full dataset — one row per homepage, every column documented, with the rendered-text SHA-256 pinning each render — is published alongside this census.

8 Limitations

One render, two days. RAW bytes are from 2026-07-09 and RENDERS from 2026-07-10, about one day apart. Homepage news tickers, event lists, and A/B variants change between days, so per-domain counts carry small drift. The metric is a magnitude comparison — is rendered many times larger than raw? — which is robust to that drift; it is not a diff of exact strings, and no per-domain ratio should be read as precise to the character.

The two clients are gated differently. RAW was fetched with a non-browser research user agent from one network; RENDERED used a real Chrome user agent from a datacenter IP. Sites that serve different content by user agent, or that challenge by IP reputation, will show a raw-versus-rendered gap that is access control rather than rendering — demonstrably so in both directions here (two Incapsula shells on the raw side, eleven render-side 403s). This confound cannot be separated from genuine client-side rendering by a single raw-versus-rendered comparison, which is precisely why the hard-wall count is decomposed rather than reported bare, and why it is a ceiling on client-side rendering, not an estimate.

innerText is a proxy for rendered visible text, not for indexable text. innerText approximates the text a browser draws for a human [6], and excludes CSS-hidden elements. A search or AI crawler’s own extractor may differ at the margins — it may or may not include alt text, aria-label, or visually-hidden but screen-reader-readable spans — and an agent that reads the accessibility tree rather than the rendered text is a third reader again, as the note *Reading the accessibility tree* and BA-W-2026-03 discuss. The figures here are a consistent, reproducible proxy for “visible rendered text,” matched to the human-visibility framing of the raw parser, not a claim about any specific engine’s extraction.

Headline denominator. Rates are computed over the 400 clean renders only. The 17 timeouts captured a best-effort, possibly partial body; the 11 challenge-suspected renders captured a block page; the 1 error captured little. Including any of them would let an untrustworthy rendered_chars distort the rates — a challenge body of a few characters against a substantial raw page would spuriously read as raw-heavy — so they are quarantined by construction and reported as counts. smu.edu.sg, whose raw side is an Incapsula shell and whose 7,140-character render arrived only after a timeout, is excluded on the same conservative rule.

Frame, not population. The 429 homepages are the analyzable subset of a documented 500-university frame that over-represents larger, higher-traffic, and internationally active institutions; the U.S. portion is ordered by enrollment and the international portion by traffic rank. Findings describe this frame on these two days and do not generalize to all universities, and the client-side-rendering rate in particular is a small integer count in this frame, not a population estimate.

Homepage only, single pass. Only the homepage of each domain is measured. An institution may serve server-rendered homepage text and client-render its inner pages, or the reverse; nothing here speaks to

pages below the front door. And each homepage is rendered exactly once, so a transient block or timeout is recorded as such rather than smoothed away by retries — which is why the excluded categories are reported openly rather than back-filled.

Within these bounds, the census’s conclusion is narrow and well supported: in the raw HTML that non-rendering crawlers read, the visible text of most homepages in this frame is already present without JavaScript; the small number of exceptions are mostly not client-side rendering; and the difference between what a crawler and a browser can read is, on a meaningful share of sites, a matter of access control that runs in both directions.

REFERENCES

1. U.S. Department of Education, College Scorecard. *College Scorecard data (institution-level open data and API)* (2026). <https://collegescorecard.ed.gov/data> Accessed 2026-07-09. [archived]
2. Hipolabs. *university-domains-list (world universities and domains)* (2026). <https://github.com/Hipo/university-domains-list> Accessed 2026-07-09. [archived]
3. Tranco (Le Pochat et al., NDSS 2019). *A research-oriented top sites ranking hardened against manipulation (list ID 46XZX)* (2026). <https://tranco-list.eu/> Accessed 2026-07-09. [archived]
4. Vercel engineering (with Merj). *The rise of the AI crawler* (2025). <https://vercel.com/blog/the-rise-of-the-ai-crawler> Accessed 2026-07-10. [archived]
5. Puppeteer (Chrome for Developers project). *Puppeteer — headless Chrome/Chromium automation over the Chrome DevTools Protocol* (2026). <https://pptr.dev/> Accessed 2026-07-10. [archived]
6. WHATWG. *HTML Living Standard — The innerText and outerText properties* (2026). <https://html.spec.whatwg.org/multipage/dom.html#the-innertext-idl-attribute> Accessed 2026-07-10. [archived]
7. M. Koster, G. Illyes, H. Zeller, and L. Sassman, IETF. *RFC 9309: Robots Exclusion Protocol* (2022). <https://www.rfc-editor.org/rfc/rfc9309.html> Accessed 2026-07-09. [archived]

HOW TO CITE

Barkhausen AI (2026). The JavaScript wall at the university front door: a raw-versus-rendered census of homepage text. <https://barkhausen.ai/research/university-js-rendering-census-2026/>

Published under the Creative Commons Attribution 4.0 International (CC-BY-4.0).



Barkhausen AI · Est. MMXXVI · *Every leap begins as a crackle.*