



DATA INSIGHT · BA-DI-3

Two ways to count an llms.txt file, two adoption rates

A site is said to have adopted llms.txt when it publishes a /llms.txt file.

DOCUMENT	BA-DI-3
KIND	Data insight
DATE	2026
SUBJECTS	Measurement & statistics; Crawlers & access; Conventions & policy
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

A site is said to have adopted llms.txt when it publishes a /llms.txt file. But 'has one' can be operationalized two ways, and the 2026 crawler-access census measured both: present, an HTTP 200 at /llms.txt, and valid, a content check that the 200 is markdown-like rather than an HTML page returned for any path. The two disagree widely. Counting presence, e-commerce adoption is 18.2% and government 15.2%; counting validity, the same sectors read 8.8% and 1.8% — a gap that reaches better than eight-to-one for government, where 76 domains answer /llms.txt with a 200 but only 9 return something that resembles the file. The census already reported these counts; this note isolates what the gap is about. Neither number is wrong, and the choice between them is a measurement decision, not a fact about the world. An 'adoption rate' is interpretable only when it says which rule it counted — the same disclosure discipline the minimum-disclosure convention requires of any visibility claim.

CONTENTS

1	Two operationalizations	3
2	The same sector, two rates	3
3	An adoption rate has to name its rule	3
4	Limitations	4

An /llms.txt file is a proposed convention: a markdown file at a site’s root that offers language models a curated, readable guide to the site’s content [1]. Whether a site has adopted it looks like a yes-or-no question, and reporting an adoption rate looks like a matter of counting the sites that say yes. It is not quite, because “has an /llms.txt file” can be operationalized in more than one way, and the operationalizations do not agree. The 2026 crawler-access census measured two of them on the same 2,000 domains, and the distance between the results is the subject of this note.

1 Two operationalizations

The census probe recorded two fields for each domain, over a denominator of the full 500 domains per sector — the probe runs regardless of a site’s robots.txt outcome.

The first is **present**: the request GET /llms.txt returned HTTP 200. This is the most permissive reading of “has one” — the path answers with a success status.

The second is **valid** (looks_valid in the tool), a content-sanity check that the 200 is a file rather than a web page. Its operationalization, taken from the census code, is exact: the response body, after a leading byte-order mark and whitespace are stripped, must **not** begin with an HTML marker (<!doctype, <html, <head, <?xml, or <body); and it must then satisfy at least one of three conditions — it begins with #; or its content-type is text/markdown or text/plain and its body is non-empty; or the body contains a markdown heading (# through ##### followed by a space), a markdown link ([...] (...)), or a bullet (- or * followed by a space). The check exists because a bare 200 is weak evidence: many sites answer every path, /llms.txt included, with the same HTML page, and that page is not an llms.txt file.

2 The same sector, two rates

Applied to the corpus, the two counts diverge sharply, and consistently in the same direction:

SECTOR	PRESENT (OF 500)	VALID (OF 500)	PRESENT RATE	VALID RATE
E-commerce	91	44	18.2%	8.8%
Government	76	9	15.2%	1.8%
News	75	24	15.0%	4.8%
Universities	57	16	11.4%	3.2%

A recount of the census records reproduces every cell. Read one way, e-commerce leads at 18.2% adoption; read the other way, its rate is 8.8%. Government is the extreme case: 15.2% of its domains return a 200 at /llms.txt, but only 1.8% return something that looks like the file — 76 present against 9 valid, a ratio better than eight to one. Across all four sectors the totals are 299 present and 93 valid, so the permissive count is roughly 3.2 times the stricter one. The census already published these figures and the present-versus-valid framing; what this note adds is the reading of the gap itself.

3 An adoption rate has to name its rule

The gap is not a discrepancy to resolve, and neither column is the “true” one. Presence and validity measure different things — that a path answers, and that the answer is a file of the intended kind — and both are legitimate quantities. What the gap shows is that “llms.txt adoption,” stated as a single percentage, is underspecified: for government it names either 15.2% or 1.8%, and the two differ by an order of magnitude entirely because of an unstated measurement choice. A reader given only the number cannot

tell which was counted, and so cannot compare it to any other adoption figure, which may have counted the other.

This is the same discipline the minimum-disclosure convention requires of visibility claims: a rate is evaluable only when the rule that produced it is stated alongside it. An adoption figure for `/llms.txt` — or for any of the newer `robots.txt` signals whose take-up is now being tracked — should say whether it counts a successful response or a valid file, because the choice moves the headline by threefold overall and eightfold in the sector where it matters most. The count is cheap to disclose and expensive to omit. The per-domain `present` and `looks_valid` values are in the crawler-access dataset for anyone who wants to recount either way.

4 Limitations

`valid` is a heuristic, not a validator of the `llms.txt` proposal: it screens out HTML pages returned for any path, which is what inflates the presence count, but it can admit a malformed file that happens to open with a heading and can reject an unusual but legitimate one. The `present-versus-valid` gap is robust to the heuristic's exact threshold — the pages it removes are plainly not `llms.txt` files — but the valid counts should be read as “resembles a valid `llms.txt` file,” not “conforms to the convention.” Both fields are point-in-time probes on the census access date and speak only to whether a file is published and well-formed; whether any model requests or uses such a file is a separate question this note does not address, and one the census takes up directly.

REFERENCES

1. Jeremy Howard (Answer.AI). *The /llms.txt file* (2024). <https://llmstxt.org/> Accessed 2026-07-10. [\[archived\]](#)

HOW TO CITE

Barkhausen AI (2026). Two ways to count an `llms.txt` file, two adoption rates. <https://barkhausen.ai/notes/llms-txt-adoption-gap/>

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



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