

Placebo Failures in Domain-Stratified Interrupted Time Series of Mean Lexical Diversity on the Post-2022 Web: A Methodological Caution on Common Crawl MATTR/MTLD Break- point Identification (with Secondary LLM-Associated Vocabulary Markers)

Muhammad Ahmad

ahmadshaukat328@gmail.com

Research Article

Keywords:

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11117458/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Placebo Failures in Domain-Stratified Interrupted Time Series of Mean Lexical Diversity on the Post-2022 Web: A Methodological Caution on Common Crawl MATTR/MTLD Breakpoint Identification (with Secondary LLM-Associated Vocabulary Markers)

1

Introduction

Every major communication technology leaves a linguistic fingerprint. Generative language models are unusual because they do not merely transmit language; they produce it at scale, sampling from smoothed distributions that favor high-probability continuations. If model text—and human writing polished by models—enters the open web in volume, several distinct ecological changes become plausible: fewer rare types on average (mean lexical decline), convergence of authors toward a shared complexity band (variance homogenization), and rising use of model-favored collocations (stylistic imprint).

These are empirically separable. Controlled comparisons find that ChatGPT often uses fewer distinct words than humans on matched tasks . Recursive training on synthetic text reduces lexical and syntactic diversity in model outputs . A large multi-dataset study reports a 21–50% reduction in *writing-complexity variance* after LLM adoption . At the same time, tests of English online news find little MATTR/MTLD homogenization even as LLM-associated vocabulary rises , and academic writing sometimes becomes *more* lexically diverse after ChatGPT—especially among non-native authors .

The open question for **naturally occurring web text** is therefore not “whether LLMs affect language at all”—that claim is already supported—but:

Does widespread LLM adoption produce a detectable decline in *mean* lexical diversity in Common Crawl English, does any pattern differ by domain, and can stylistic imprint appear without mean-diversity collapse?

Three design constraints shape the answer.

1. **Detection is not measurement.** Commercial AI detectors disagree and mislabel human text.
2. **Mean diversity $\neq$ variance homogenization $\neq$ style.** Sourati et al. primarily document shrinkage of complexity *variance*. Fitterer et al. document style without MATTR collapse in news.
3. **ChatGPT's release is not a clean experiment.** November 2022 coincides with other model releases, SEO shifts, platform policy changes, and crawl-composition dynamics. We estimate **post-ChatGPT-period change**, not a causal “ChatGPT effect.”

Hypotheses and analysis-plan revision

- **Initial H1 candidate.** Post-ChatGPT-period mean MATTR/MTLD decline relative to pre-period trend, with domain stratification.
- **Analysis-plan revision (identification).** Placebo ITS failures demoted **news** and, under a consistent Option A rule, also demoted **blogs MATTR**, **forums MTLD**, and kept **commerce** out of confirmatory inference (convenience sample). The **remnant placebo-eligible associational tests** are Post on {**blogs MTLD**, **forums MATTR**} only (2 tests; BH-FDR descriptive only). We do **not** claim a frozen pre-registered confirmatory family.
- **H2 (exploratory).** Any decline larger in lower-gatekeeping domains than in news.

Contributions

1. A longitudinal, domain-stratified Common Crawl ITS (2020–2024), extending two-period news comparisons .
2. Separation of **mean lexical diversity** from **LLM stylistic imprint**, against variance-homogenization evidence .
3. Identification discipline: **consistent placebo demotion** (Option A) across strata; commerce excluded from confirmatory; FDR; MDE/power; wild-bootstrap p and CIs; uncontrolled ITS; host-label audit.
4. Evidence that raw post-2022 mean declines can disappear after trend adjustment, and that powered breakpoint nulls are **not** placebo-robust for blogs MATTR in this panel.

Background and related work

Why homogenization is plausible

Autoregressive models under-sample rare words under typical decoding. Writing assistance can pull heterogeneous drafts toward a shared stylistic basin . Cultural feedback can imprint model-favored lexicon into human communication . An industry report documents steep rises in primarily AI-classified publishing after late 2022 — motivation only, not a dependent variable and not peer-reviewed evidence.

Why mean lexical collapse is contested

Laboratory lexical poverty need not imply open-web MATTR decline . Two-period snapshots cannot separate LLM-period change from long-run drift; an ITS with a dense pre-period can.

Positioning relative to Fitterer (2025) and Sourati (2026)

Table 1 should appear about here.

Positioning of this study relative to closely related work.

Study	Object	Design	Scope
Fitterer et al. (2025)	MATTR/MTLD/SWR	Cross-sectional 2018 vs 2024	English online news
Sourati et al. (2026)	Writing-complexity variance	Multi-dataset LLM-adoption comparison	> 880k texts, seven datasets
This study	Mean MATTR/MTLD + SWR	Bimonthly ITS 2020–2024	News (diagnostic), blogs, forums, convenience commerce

Data and methods

Causal language and identification

November 2022 is the **primary period marker**. Concurrent confounders include other LLM releases, AI writing tools, SEO incentives, platform moderation, COVID-era content shifts, and Common Crawl composition changes.

Primary reporting sample: crawl-month indexing (crawl_full). Secondary robustness: dated_only (news only has usable coverage).

For month t and domain s :

$$Y_{t,s} = \beta_0 + \beta_1 T_t + \beta_2 \text{Post}_t + \beta_3 (T_t \times \text{Post})_t + \gamma_1 \overline{\text{tokens}}_{t,s} + \gamma_2 \log N_{t,s} + \varepsilon_{t,s},$$

with HAC/Newey–West standard errors (maxlags = 3).

Endpoint families and multiplicity

Table 2 should appear about here.

Endpoint families after identification review.

Family	Tests	Decision rule
Placebo-eligible associational (Option A)	Post on { blogs MTLD, forums MATTR } only (2 tests)	BH-FDR (descriptive); underpowered
Secondary	Post and $T \times$ Post on strict SWR (all strata)	BH-FDR; imprint
Diagnostic	News; blogs MATTR (placebo fail); forums MTLD; commerce (convenience); full placebo grids; seasonal FE	Not confirmatory
Exploratory	Maas, Shannon, hapax, SDs; H2; dated-only; date grid	Raw p with caution

News is diagnostic-only. Placebo ITS on news fires at fake January 2021 and May 2022 dates; the breaks are driven by outlier bimonthly jumps (Nov 2020→Jan 2021 crash; Mar 2022→May 2022 spike), not a smooth trend misspecification. Odd-month fixed effects do not restore placebo validity (May 2022 still $p < .01$; Jan 2021 MTLD still $p \approx 4 \times 10^{-6}$). We therefore **drop news from confirmatory H1** rather than claim a repaired specification. The same Option A rule later demotes blogs MATTR (fake May 2022, $p \approx .022$), forums MTLD, and keeps commerce outside confirmatory inference.

Corpus, commerce caveat, and cell size

Table 3 should appear about here.

Domain strata. Commerce is a convenience slice, not a peer probability sample.

Domain	Source	Role in inference
News	CC-NEWS	Diagnostic only (failed placebos)
Blogs	CC-MAIN host match	MATTR diagnostic (placebo fail); MTLD placebo-eligible but underpowered
Forums	CC-MAIN host match	MATTR placebo-eligible but underpowered; MTLD diagnostic (placebo fail)
Commerce	CC-MAIN hosts + paths	Diagnostic only (convenience ; not confirmatory)

Domain assignment is disjoint (forums > blogs > commerce). **Commerce** mixes Shopify/Etsy/eBay/Amazon-style hosts and /product//review-style paths: $\approx 50.8\%$ of

commerce documents match a commerce host; $\approx 49.2\%$ are path-only; $\approx 82\%$ of unique commerce hosts are path-only. It is not interchangeable with blogs/forums as a sampling frame.

Panel. Bimonthly 2020-01 through 2024-11 ($T = 30$). Operational cell size ≈ 100 docs/domain/month—below the design ceiling of 300—because of wall-clock Common Crawl fetch constraints. That choice materially raises month-level noise: bootstrap SE of month means at $n = 100$ is ≈ 0.026 for forums MATTR vs ≈ 0.007 – 0.012 for news/blogs/commerce, and is the main reason forums/MTLD tests are underpowered. Final $N = 11,050$.

Host-label audit. We sampled 20 unique hosts per stratum (80 total; fixed seed) and assigned gold labels by inspecting URL/host against the written domain definitions. Overall agreement with the pipeline was 80/80 (100%) in this sample; the audit is a definition-consistency check, not an independent content taxonomy. Disagreement risk remains highest for path-only commerce.

Outcomes and robustness

Primary outcomes: MATTR (window 50) and MTLD (threshold 0.72), computed with `lexicalrichness` on whitespace/punctuation-tokenized lowercase tokens.

Strict SWR (secondary; operational definition). Strict style-word rate is the document share of tokens that match a fixed closed list of LLM-associated surface markers stored in `config_v2.json` (`metrics.swr_strict_words`; study freeze `run_v2`). The list is: `delve/delves/delved/delving`, `meticulous(ly)`, `boast(s|ed)`, `underscore(s|ed)`, `tapestry`, `realm(s)`, `embark(s|ed|ing)`, `pivotal`, `intricate/intricacies`, `nuanced`, `testament`, `landscape`, `comprehend(s)`, `swift`, `robustness`. Matching is exact token equality after lowercasing (no stemming beyond the listed inflections; no POS filter). The reported metric is $\text{swr_strict} = \#\{\text{hits}\}/N_{\text{tokens}}$. Related variants `swr_per_1k` and `swr_content` (hits among non-stopword tokens) are exploratory only. SWR rise is treated as a **stylistic association**, not proof of LLM authorship.

Exploratory: Maas, Shannon, hapax, SDs.

Robustness: placebo grid (all strata); intervention-date grid (Oct 2022–Mar 2023); sampling-frame ITS; MDE_{80} and simulation power; HAC lags 1–4; Rademacher wild bootstrap p and 95% CIs ($B = 1999$); uncontrolled ITS (no tokens/ $\log N$); news seasonal FE sensitivity (fails).

Results

Descriptive change

Table 4 should appear about here.

Descriptive pre/post means (crawl_full).

Domain	Period	MATTR	MTLD	Strict SWR
News	Pre	0.799	92.4	0.0002
News	Post	0.795	89.8	0.0003
Blogs	Pre	0.779	84.4	0.0003
Blogs	Post	0.747	72.4	0.0005
Forums	Pre	0.636	61.9	0.0002
Forums	Post	0.557	50.0	0.0001
Commerce	Pre	0.752	82.9	0.0001
Commerce	Post	0.748	76.8	0.0003

Raw drops are largest in blogs and forums; ITS asks whether they exceed pre-trends.

Option A placebo-eligible associational tests (not confirmatory discovery)

Consistent placebo rule. Any stratum–outcome with placebo-grid $p < .05$ at a fake shock is demoted from confirmatory (same standard as news). Failures include blogs MATTR (fake 2022-05, $p \approx .022$), commerce MATTR (fake 2021-01), and forums MTLD (fake 2022-01). Commerce is additionally excluded as a convenience sample. The remnant placebo-eligible associational tests are therefore **{blogs MTLD, forums MATTR}** only (2 tests; BH-FDR as descriptive multiplicity control after placebo screening—**not** a pre-registered confirmatory primary).

Table 5 should appear about here.

Placebo-eligible associational Posts (Option A). Both cells underpowered—inconclusive, not evidence of absence. Not a confirmatory discovery family.

Domain	Outcome	Post β	p	p_{FDR}	Wild p	Pow@5%
Blogs	MTLD	−0.45	.96	.96	.97	.23
Forums	MATTR	−0.011	.87	.96	.92	.34

Table 6 should appear about here.

Diagnostic (non-confirmatory) Posts at November 2022, including demoted blogs MATTR. Wild 95% CIs from unrestricted residual wild bootstrap ($B = 1999$).

Domain	Outcome	Post β	HAC p	Wild CI ₉₅	Role
Blogs	MATTR	−0.005	.70	[−0.039, 0.027]	Diagnostic (placebo fail)
Commerce [†]	MATTR	−0.012	.060	[−0.035, 0.010]	Diagnostic (convenience)
News	MATTR	+0.010	.18	[−0.011, 0.032]	Diagnostic (placebo fail)

[†]Convenience slice; not confirmatory. All listed CIs cover 0.

No informative powered null remains. Option A yields 0/2 FDR rejections among underpowered remnant cells. Blogs MATTR must **not** be sold as a powered null under breakpoint identification: it fails the placebo standard applied to news. Uncontrolled ITS (no tokens/log N) likewise yields wild CIs covering zero for primary Posts.

Borderline placebo survivors (disclose). Remnant cells clear $p < .05$ but are close to the threshold: blogs MTLD at fake 2022-01 has $p \approx .065$; forums MATTR at fake 2022-05 has $p \approx .087$. Placebo tests across fake dates $\times$ strata $\times$ outcomes are **not** multiplicity-adjusted; a stricter family-wise rule could demote further. Remnant Posts are fragile associational descriptions, not identification successes.

News diagnostic only

Table 7 should appear about here.

News estimates at November 2022 (not confirmatory). Placebos fail; month FE does not rescue.

Outcome	Post β	p	$T \times \text{Post } p$
MATTR	+0.010	.18	.58
MTLD	+1.58	.73	.72
Strict SWR	+0.0001	.24	.043

Mechanism of placebo failure. Fake 2021-01 coincides with Nov 2020→Jan 2021: MATTR 0.815 → 0.786 ($\Delta = -0.029$); MTLD 116.8 → 86.6 ($\Delta = -30$). Fake 2022-05 coincides with a one-off spike Mar→May 2022: MATTR +0.040; MTLD +23.8—not May seasonality. Month FE: Jan 2021 MATTR softens to $p \approx .10$ but MTLD stays $p \approx 4 \times 10^{-6}$; May 2022 remains $p < .01$. News breakpoint inference is unreliable in this panel; we report Table 7 only as description.

Power note

MDE_{80} for blogs/commerce MATTR is 2.5–4.8% of pre-mean; for forums MATTR it is $\approx 32\%$. The ≈ 100 -doc operational cell (vs design 300) is the binding constraint for forums precision.

Secondary SWR (suggestive only)

Table 8 should appear about here.

Secondary SWR (illustrative; full secondary FDR family in supplement sensitivity_fdr.json). FDR-fragile; commerce confined to convenience slice; news diagnostic.

Domain	Coef	β	p	p_{FDR}
News (diagnostic)	$T \times \text{Post}$	$+4.0 \times 10^{-5}$	.043	.17

Domain	Coef	β	p	p_{FDR}
Commerce [†]	$T \times \text{Post}$	$+4.5 \times 10^{-5}$	.009	.076

[†]Convenience commerce slice.

Neither survives FDR at $q = .05$. Directionally consistent with Fitterer et al. , but not confirmatory imprint proofs.

Exploratory H2, sensitivity, sampling

Pooled Post $\times$ low-gatekeeping is null. Forums $\times$ Post on MATTR is locally significant uncorrected ($p \approx .010$) and remains exploratory. Intervention-date sensitivity and sampling-frame ITS (no November 2022 share/token/publish-rate break) are unchanged in substance from prior analyses.

Discussion

What the design can and cannot claim

Under a consistent placebo rule (Option A), confirmatory breakpoint identification for mean MATTR/MTLD is not supported in this panel: the only placebo-eligible associational cells are underpowered (0/2 FDR). Blogs MATTR—the prior powered-null candidate—fails fake May 2022 placebos and is diagnostic only. News and commerce remain non-confirmatory. Style markers (SWR) remain FDR-fragile associations. We do not establish a causal ChatGPT effect.

This complements Fitterer et al. and does not refute Sourati et al. variance findings: we test mean type-richness on crawl strata, not writing-complexity variance.

Limitations

$T = 30$ is thin for asymptotic HAC (wild bootstrap mitigates). Operational $n \approx 100/\text{cell}$ (wall-clock constrained vs design 300) drives forums underpower. News is diagnostic-only after failed placebos. Commerce is a convenience host/path slice ($\approx 49\%$ path-only documents). Domain labels are rule-based; a 80-host definition-consistency audit found 100% agreement, but path-only commerce remains the weakest link and the audit is not an independent content taxonomy. English filtering uses langdetect with a heuristic fallback. For undated documents, the panel indexes **crawl month, which is not a publication date**. Replication in the accompanying kit is **metrics-locked** (re-runs ITS/placebos/FDR/bootstrap from locked monthly/document panels; full WARC re-fetch is optional). AI detectors are excluded from the DV.

Conclusion

1. **Option A remnant associational tests:** only blogs MTLD and forums MATTR remain; 0/2 FDR; both underpowered (inconclusive); **not** confirmatory discovery labels.
2. **No informative powered null:** blogs MATTR demoted after fake May 2022 placebo failure ($p \approx .022$), matching the news demotion standard.
3. **Commerce** is outside confirmatory inference (convenience sample).
4. **News diagnostic only:** placebo failures from outlier months; month FE does not rescue.
5. **Wild-bootstrap 95% CIs** for Post cover zero under controlled and uncontrolled ITS for primary stratum–outcomes examined.
6. **SWR:** FDR-fragile; commerce caveated.

Short version: *With a consistent placebo rule, November 2022 breakpoint identification for mean MATTR/MTLD is not placebo-robust in this Common Crawl panel; the prior blogs MATTR powered-null claim is withdrawn; LLM-style imprint remains cautious; no causal ChatGPT effect is identified.*

Abbreviations

BH-FDR

Benjamini–Hochberg false discovery rate

CI

Confidence interval

HAC

Heteroskedasticity- and autocorrelation-consistent (Newey–West)

ITS

Interrupted time series

LLM

Large language model

MATTR

Moving-average type–token ratio

MDE

Minimum detectable effect

MTLD

Measure of textual lexical diversity

SWR

Style-word rate (LLM-associated vocabulary markers)

Declarations

Ethics approval and consent to participate

Not applicable. This study analyses publicly available web documents from Common Crawl and does not involve human participants, identifiable personal data, or interventions requiring ethics approval.

Consent for publication

Not applicable.

Availability of data and materials

The study uses English web documents sampled from publicly available Common Crawl (CC-NEWS and CC-MAIN) crawls (2020–2024). Derived analysis artifacts—locked monthly and document-level metrics panels, analysis scripts, and expected outputs for the metrics-locked replication kit—are available from the corresponding author upon reasonable request. Raw WARC re-collection from Common Crawl is optional and not required to reproduce the reported ITS, placebo, FDR, and wild-bootstrap results.

Competing interests

The author declares that there are no competing interests.

Funding

No funding was received for this study.

Authors' contributions

M.A. conceived the study, collected and analysed the data, wrote the manuscript, and approved the final version.

Acknowledgements

Metrics-locked replication kit (scripts 03–09, locked panels, expected outputs) accompanies this revision. Full WARC re-collection is optional (P1). Pilot ($N = 3,940$) archived in the parent project.

B. Dalaman, A. F. Kalay, N. Kettlewell, *Bridging language barriers: The impact of large language models on academic writing*, IZA Discussion Paper No. 18215 (2025). <https://docs.iza.org/dp18215.pdf>

S. Fitterer, D. Gangl, J. Ulbrich, Testing English news articles for lexical homogenization due to widespread use of large language models, in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1239–1245, Vienna, Austria (2025). DOI: 10.18653/v1/2025.acl-srw.95. <https://aclanthology.org/2025.acl-srw.95/>

Graphite, *AI now writes as many online articles as humans*, industry report (non-peer-reviewed) (2025/2026). <https://graphite.io/five-percent/research/ai-now-writes-as-many-online-articles-as-humans-do>

Y. Guo, G. Shang, M. Vazirgiannis, C. Clavel, The curious decline of linguistic diversity: Training language models on synthetic text, in *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3589–3604 (2024). DOI: 10.18653/v1/2024.findings-naacl.228. <https://aclanthology.org/2024.findings-naacl.228/>

P. Reviriego, J. Conde, E. Merino-Gómez, G. Martínez, J. A. Hernández, Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans, *Machine Learning with Applications* 18 (2024) 100602. DOI: 10.1016/j.mlwa.2024.100602. <https://doi.org/10.1016/j.mlwa.2024.100602>

Zhivar Sourati, F. Karimi-Malekabadi, M. Ozcan, C. McDaniel, A. Ziabari, J. Trager, A. N. Tak, M. Chen, F. Morstatter, M. Dehghani, *The shrinking landscape of linguistic diversity in the age of large language models*, *Nature Human Behaviour* (published 24 August 2026). <https://doi.org/10.1038/s41562-026-02550-0>

H. Yakura et al., Empirical evidence of large language model’s influence on human spoken communication (2024). <https://arxiv.org/html/2409.01754>