



PROMPTSLOVE RESEARCH · REPORT 1.0

The Creator-Identity Illusion: A Within-Author Study of Engagement on Threads

12,045 posts on Meta's Threads, in which twenty of twenty-three widely taught engagement tactics fail to survive holding the author fixed, tested under an adversarial multi-agent verification protocol



Ramanpal Singh

Promptslove · promptslove.com

admin@promptslove.com

Citation and Rights

CITE THIS REPORT AS

Singh, R. (2026). The Creator-Identity Illusion: A Within-Author Study of Engagement on Threads. Promptslove Research Report 1.0. promptslove.com.

BIBTEX

```
@techreport{singh2026threads,  
  title = {{The Creator-Identity Illusion: A Within-Author Study of Engagement on  
Threads}},  
  author = {Singh, Ramanpal},  
  institution = {Promptslove},  
  year = {2026},  
  type = {Research Report},  
  number = {1.0},  
  url = {https://promptslove.com}  
}
```

Author. Ramanpal Singh, Promptslove (promptslove.com). Correspondence to ad'min@promptslove.com.

Data availability. The analysed corpus (12,045 posts), the top-750 extract, the 59-account profile table and the twenty-three tactic tests are released with this report and listed in Appendix D. Data was collected from public posts through the platform's own interface from an authenticated session; no credential is distributed and no private content was accessed.

Conflicts of interest. The author operates promptslove.com, a commercial publisher of AI prompt and automation products, and posts on Threads in the subject area studied. The study analyses public posts about AI tools, including categories in which the author has a commercial interest. No sponsor influenced the design, analysis or reporting. Several findings are commercially inconvenient — including that the make-money-with-AI lane is the worst-performing subject measured, and that prompt-craft content sits below the baseline — and are reported unchanged.

Use of automated systems. Data collection, feature extraction, statistical analysis, adversarial verification and a first draft of the reporting text were produced with the assistance of large language model agents under the author's direction and review. The verification protocol of Chapter 10 was applied to the analysis and is reported with its failures; residual error remains the author's responsibility.

Corrections. An earlier, population-level version of this analysis was published online and reported several effects that the within-author test in Chapter 7 subsequently overturned. That version carries a correction notice directing readers to this report. Withdrawn claims are listed in Appendix E.

Rights. © 2026 Promptslove. This report may be quoted and cited with attribution. Post text quoted in the report remains the property of its authors and is reproduced in short extracts for research commentary.

Preface

This report started with a spreadsheet and an argument. Working on products in the AI tooling space, the author kept being handed the same advice about building an audience on Threads: lead with a short hook, use bullet points, never open on a question, avoid the word “AI”, post at a particular hour. The advice was confident and widely repeated. When evidence was offered it took a consistent form — a screenshot of a post that did well, and a list of the things it did.

The obvious response was to collect enough posts to check. That part was straightforward, and the first version of this analysis duly confirmed most of the advice: bulleted posts really do earn more likes than unbulleted ones, short first lines really do out-perform long ones, and the numbers were comfortably significant. Those findings were published. They were also, it turned out, an artefact of who was writing.

The correction came from a question that should have been asked first: what happens when the same person does it and doesn't do it? The corpus contained hundreds of creators who had done both, and comparing each against themselves dissolved twenty of twenty-three effects, including two the author had already put in print. The bulleted-list effect went from a clean 1.53-fold advantage to a point estimate on the wrong side of no effect. That reversal is the reason this document exists, and it is why the earlier version now carries a correction notice pointing here.

The method that produced the correction is not sophisticated. It is the oldest idea in experimental design — compare like with like — applied to data that was never designed to support it. What makes it worth a report is how much it destroyed, and how little was left: three verified actions out of twenty-three tested, all three about what is attached to a post rather than how the post is written.

Two further habits shaped the work and are visible throughout. Everything was run twice, once by agents asked to find what the data showed and again by agents asked to attack what the first pass claimed; seven of nine headline findings did not survive that exchange, and Chapter 10 reports them with the objection that felled each. And every error found in the study's own working — a scanning defect that silently discarded posts, a storage failure that made the corpus shrink, an entire analytical frame built on a denominator that measured the author's own search terms — is documented rather than quietly repaired, because those errors are more instructive than several of the findings they contaminated.

A reader who wants the practical answer should read Chapter 7 and Section 11.2 and can stop there. A reader who wants to know whether to believe it should read Chapters 3, 4 and 10, which between them state everything this corpus cannot do. A reader who disagrees with a conclusion should find, in every case, the sample size it rests on and what was already said against it.

Ramanpal Singh
Promptslove · promptslove.com

Abstract

Background. Guidance on growing an audience on Meta's Threads circulates as untested folklore, and the evidence usually offered for it — comparing posts that used a tactic against posts that did not — cannot separate the tactic from the person who chose it, because creators are not randomly assigned to techniques.

Method. We assembled an observational corpus of 12,045 unique Threads posts from 9,922 accounts, collected from a logged-in session across approximately 700 search queries in seven passes, each post carrying the platform's own like, reply, repost and quote counters. Rate comparisons use the 4,247 posts at least three days old, whose median is 38 likes. Twenty-three commonly recommended tactics were estimated twice: at population level, and by a within-author paired estimator that restricts attention to creators who both used and omitted the tactic and compares each creator against themselves, with 95 per cent percentile bootstrap intervals over authors and two-sided sign tests. Nine independent language-model analysts then mined the corpus and nine further agents attempted to refute their headline claims by re-deriving them from the raw records.

Results. Twenty of twenty-three tactics fail to survive the within-author test. Bulleted lists move from a 1.53-fold population advantage to 0.80 within author; numbered lists from 2.05 to 0.88; short first lines from 1.64 to 1.14; and an apparent 0.63-fold penalty for using AI vocabulary becomes 1.12 with an interval spanning no effect. Three tactics survive: attaching any media (1.98×, CI 1.33–4.78, 95 authors), posting a carousel (1.73×, CI 1.18–3.18, 68 authors) and, negatively, including an outbound link in the post body (0.30×, CI 0.13–0.71, 41 authors). Format also determines the composition of engagement, with carousels returning 10.2 reposts per hundred likes against 2.2 replies while plain text inverts the ratio. Language is the largest association measured — Thai 6.38×, Chinese 2.95×, Korean 2.62× against 0.82× for Latin script — with the Chinese result replicating inside all four collection passes in which it appears. Attention is concentrated (the top decile takes 79.0 per cent of likes) but not held: 88.7 per cent of accounts appear once, and the 424 top-decile posts come from 391 distinct accounts. Adversarial verification refuted seven of nine headline findings, with creator identity, collection provenance, circular definition and group composition accounting for all seven.

Conclusions. In observational engagement data the gap between a between-author and a within-author comparison is not a refinement but, for most content tactics, the entire effect. The practical residue is small and specific: what is attached to a post survives testing, how the post is written mostly does not, and an operator's own account is the only sample that can answer the question they are actually asking.

Keywords: Threads · social media analytics · virality · within-subject design · fixed effects · confounding · Simpson's paradox · observational study · sampling bias · multi-agent verification · creator economy

Table of Contents

1. Introduction	13
1.1 Why Threads, and why now	13
1.2 Research questions	13
1.3 Contributions	14
1.4 What this study cannot claim	14
1.5 Structure	15
2. Related Work	16
2.1 What makes content spread	16
2.2 Confounding, aggregation and the ecological fallacy	17
2.3 Sampling and validity in social-media data	17
2.4 Ranking as an unobserved mediator	18
2.5 Language-model agents as analysts	18
2.6 Reporting standards and researcher degrees of freedom	18
2.7 Position of this work	19
3. Methodology	20
3.1 Design	20
3.2 Collection	20
3.3 The maturity rule and the analysis set	21
3.4 Estimators	22
3.5 Confound control	22
3.6 Adversarial verification	23
3.7 What this design cannot identify	23
4. The Corpus	25
4.1 Composition	25
4.2 Collection passes and why they matter	26

4.3 The maturity rule	27
4.4 What a search-surfaced sample is	27
4.5 What the corpus cannot measure	28
4.6 Ethics and data handling	28
5. The Shape of Engagement	30
5.1 The distribution	30
5.2 Concentration	31
5.3 What follows for measurement	32
6. Format, Media and the Currency of Engagement	33
6.1 Media type and volume	33
6.2 The currency: reposts against replies	33
6.3 Where the median and the tail disagree	35
6.4 Outbound links	35
6.5 Verification and length	36
7. The Within-Author Test	37
7.1 The identification problem	37
7.2 The estimator	37
7.3 Result	38
7.4 The scale of the illusion	41
7.5 What a null does not mean	43
7.6 Timing, tested the same way	44
8. Subject, Tool and Language	45
8.1 Subject lanes	45
8.2 The tool map	47
8.3 Language markets	49
9. The Creator Landscape	51

9.1 A transient field	51
9.2 Followers predict less than expected	52
9.3 Repeatability cannot be established here	53
9.4 The top decile at the account level	54
10. Adversarial Verification	55
10.1 The protocol	55
10.2 Outcome	56
10.3 The four confound families	57
10.4 What the refuters got wrong	58
10.5 Errors found in the study's own working	58
11. Discussion	60
11.1 Implications for analysis	60
11.2 Implications for practitioners	60
11.3 Why the illusion is so persistent	62
11.4 Threats to validity	62
12. Conclusion	64
12.1 Answers to the research questions	64
12.2 The general result	65
12.3 Future work	65
12.4 Closing	66
References	67
Appendix A. Collection Harness	70
A.1 Mechanism	70
A.2 Three defects	70
Appendix B. Query Design	72

Appendix C. Feature Definitions	73
Appendix D. Reproduction	75
Appendix E. Withdrawn Findings	76

List of Figures

Figure 3.1 — The study pipeline. The first four stages are collection; the last three are analysis.	21
Figure 4.1 — Median likes by collection pass, analysis set. A post's pass predicts its like count more strongly than most of its content features do.	26
Figure 5.1 — The like distribution across the analysis set, plotted on a logarithmic axis, with the hit rate at four thresholds beneath.	30
Figure 6.1 — Reposts and replies per hundred likes by format, posts above 100 likes. Skew is reposts divided by replies.	34
Figure 7.1 — Within-author effect of each tactic, ordered by point estimate. Intervals are 95 per cent percentile bootstraps over authors. Three of twenty-three exclude 1.00.	39
Figure 7.2 — From the population estimate to the within-author estimate. The length of each arrow is the portion of the apparent effect attributable to who was using the tactic.	42
Figure 8.1 — Median likes by subject lane, computed inside the AI collection pass against its median of 30 likes.	45
Figure 8.2 — Share of voice against attention per post, seventeen tools, inside the AI collection pass. The horizontal axis is logarithmic.	47
Figure 8.3 — Median likes by language segment against the 38-like corpus baseline. Segments are assigned by Unicode script range.	49
Figure 10.1 — The verification protocol and its outcome. Four confound families account for all seven refutations.	56

List of Tables

Table 3.1 — Confounds and controls. The right-hand column names the chapter in which the control is exercised and, where relevant, the finding it overturned.	23
Table 4.1 — Corpus summary. The analysis set is the subset used for every rate comparison in this report; the maturity rule is defined in Section 4.3. Median and percentile figures are computed on the analysis set, because they are meaningless on a population that mixes finished posts with posts published hours before collection.	25
Table 5.1 — Percentiles and threshold hit rates, analysis set (n = 4,247). The "one post in" column is the reciprocal of the hit rate and is the practically useful form: it converts a probability into an expected number of attempts.	31
Table 6.1 — Media type performance, analysis set (n = 4,247, baseline median 38 likes). Lift is the group median divided by the baseline. The divergence between median and mean in the text row is the point of the table and is discussed below.	33
Table 6.2 — Engagement composition by format, posts above 100 likes in the analysis set (n = 1,502). Each cell is the median of the per-post ratio within the group, not a ratio of the group's medians; the two differ and the former is the correct summary of a typical post.	34
Table 6.3 — Median against tail behaviour by format. The mean-to-median ratio is a crude dispersion index; higher values indicate that the format's average is carried by a small number of very large posts.	35
Table 7.1 — The full within-author battery. "Population" is the group median divided by the 38-like baseline, the quantity a conventional analysis reports. "Within-author" is the median paired ratio of Section 7.2. "Authors" counts creators who did the tactic in some posts and not in others. Verdict is SURVIVES where the interval excludes 1.00 and UNPROVEN otherwise.	40
Table 8.1 — Subject lane performance inside the AI collection pass (n = 2,804, median 30 likes). Authors counts distinct accounts contributing to the lane; the busiest-author share is the proportion of the lane's posts written by its single most active account, and is reported because a lane dominated by one voice would be a different kind of object from a lane that is genuinely diffuse.	46
Table 8.2 — Tools named in twenty or more posts inside the AI collection pass (n = 2,804, median 30 likes). Share of voice is the percentage of that pass's posts naming the tool. Samples are small and the ordering between adjacent rows carries little weight; the spread across the table carries a great deal.	48
Table 8.3 — Language segments, analysis set (n = 4,247, baseline 38 likes). The plain-text share is included because it bears on whether the segment differences are really format differences in disguise.	49
Table 8.4 — Chinese-language posts against all other posts, computed separately within each collection pass. A collection artefact would not replicate across four independently constructed passes; a genuine audience difference would.	50

Table 9.1 — Concentration within subject lanes, AI collection pass. The busiest-author share is the proportion of a lane's posts written by its most prolific account. A lane owned by an incumbent would show a large value here; none does.	52
Table 9.2 — Selected accounts with follower data, ordered by followers. Reach per follower is the account's median likes divided by its follower count, expressed as a percentage. The column is included to show its own instability rather than to propose it as a metric.	53
Table 10.1 — Verification outcomes by dimension. "Corrected claim" is the defensible version the refuter was able to support, not the analyst's original wording.	57
Table 11.1 — The verified practitioner set. Effect is the within-author ratio from Table 7.1. Everything not in this table is either unproven at this sample size or, in the case of subject and language, an association measured under a weaker design.	61
Table B.1 — Query families, in the order they were issued. Yield is the approximate number of previously unseen posts returned per query at the time that family was run; it declines as the corpus saturates, not because the later families are less productive in absolute terms.	72
Table C.1 — Operational definitions of the twenty-three tested tactics. "First line" means the text preceding the first newline character. Patterns are applied case-insensitively unless stated.	73
Table D.1 — Released artefacts.	75
Table E.1 — Claims produced and withdrawn during the study, with the control that removed them.	76

1. Introduction

Advice about how to grow an audience on Threads is abundant, confident and almost entirely untested. Post carousels. Keep the first line short. Open with a command, never a question. Use bullet points. Avoid the word "AI". Post at a particular hour. Each of these circulates as settled knowledge, and each is usually supported in the same way: by pointing at posts that did well and noting what they have in common.

That form of evidence has a specific defect. Posts are not randomly assigned to techniques — creators choose them — and the creators who write in tight, well-formatted, disciplined prose are disproportionately the ones who have been publishing for years and have an audience already. A comparison between their posts and a newcomer's measures the technique and the person together, and cannot separate them. The advice that results describes what successful accounts look like, which is not the same thing as what would happen if you changed your own post.

This study collects the data needed to make the separation, and finds that it matters enormously. Across 12,045 Threads posts, twenty-three commonly recommended tactics were measured twice: once in the conventional way, comparing posts that used the tactic against posts that did not, and once by restricting attention to creators who did both and comparing each creator against themselves. **Twenty of the twenty-three effects disappear under the second test.** Three survive, and all three concern what is attached to a post rather than how the post is written.

1.1 Why Threads, and why now

Threads launched in July 2023 as Meta's text-first companion to Instagram, reaching one hundred million sign-ups within five days [Meta, 2023]. It is now a substantial platform with an unusually young ranking system and an unusually thin research literature: empirical work on Threads is limited to a small number of studies of its early migration dynamics and its federation with Mastodon [Jeong et al., 2024; Jeong et al., 2025], with almost nothing on what drives engagement inside it.

That gap has practical consequences. A large advisory industry has grown around the platform, and its claims are transferred wholesale from X and Instagram without testing whether they hold. Threads differs from both in ways that plausibly matter: it has no public impression counter on third-party posts, its search returns a shallow slice of results, its audience skews toward markets that are under-served in English-language commentary, and its ranking system has been revised repeatedly since launch.

1.2 Research questions

The study addresses five questions, and Chapter 12 answers each in a sentence.

RQ1. How is engagement distributed across posts that Threads surfaces, and what is the appropriate baseline against which an individual post should be judged?

RQ2. Which post-level tactics are associated with higher engagement at the population level, and which of those associations survive when creator identity is held fixed?

RQ3. Does the choice of format change not only the volume of engagement but its composition — the balance between reposts, which carry a post to strangers, and replies, which deepen an existing audience?

RQ4. Which subjects, products and language markets are associated with higher engagement, and how much of that association survives control for how the data were collected?

RQ5. What does the creator landscape look like — is attention concentrated in a stable set of established accounts, and does follower count predict the reach of a given post?

1.3 Contributions

The study makes four contributions, in descending order of how transferable they are.

A methodological result. In social-engagement data, the difference between a population-level comparison and a within-author comparison is not a refinement; it is frequently the difference between an effect and no effect. Twenty of twenty-three tactics reverse or collapse. This has been demonstrated before for wording effects on Twitter using author- and topic-controlled pairs [Tan et al., 2014]; the present study extends the approach to the full battery of formatting and structural tactics that practitioner advice actually consists of, and reports the scale of the correction.

An empirical account of Threads. To the author's knowledge this is the first substantial public analysis of what drives engagement on Threads. It reports the like distribution, the format hierarchy, the composition of engagement by format, subject-lane and language-market differences, and the structure of the creator landscape, each with sample sizes and intervals.

A verification protocol and its yield. Because the analysis was conducted with substantial assistance from language-model agents — fluent producers of plausible and wrong analysis [Hu et al., 2024] — a second population of agents was tasked with refuting the first's conclusions. Seven of nine headline findings did not survive. The protocol, its outcome and the two cases where the refuters were themselves wrong are reported in Chapter 10.

A dataset. The 12,045-post corpus with full engagement counters, the top 750 posts by likes, and the tactic-test table are released with this report.

1.4 What this study cannot claim

Three limits should be fixed in the reader's mind before any result is read, and they are restated wherever they bite.

The corpus is search-surfaced. Every post in it was returned by Threads' own search, so the sample over-represents content the ranking system already favoured and contains no observable failures. Absolute rates — the share of posts clearing a thousand likes, for instance — are inflated. Comparisons between groups inside the corpus are far less affected, because the surfacing bias applies to both arms.

The corpus was assembled in seven passes whose engagement levels differ elevenfold, so any comparison between groups drawn unevenly from different passes is partly a measure of the query list. Every claim in Chapters 5 to 9 is computed either within a single author, within a single pass, or both.

Nothing here is causal. Within-author variation is not random assignment: a creator who sometimes attaches an image may do so exactly when they have something worth showing. The surviving results are strong observational evidence and should be read as such.

1.5 Structure

Chapter 2 places the work in the literature on virality, confounding and social-data sampling. Chapter 3 gives the methodology and Chapter 4 the corpus, including an extended account of what a scraped, search-surfaced sample can and cannot support. Chapters 5 to 9 are results: the engagement distribution, format and engagement currency, the within-author battery, subject and language markets, and the creator landscape. Chapter 10 documents the adversarial verification protocol and the seven findings it destroyed. Chapter 11 discusses implications for practitioners and for analysts, and Chapter 12 answers the five research questions. Readers in a hurry should read Chapter 7, Chapter 10 and Section 11.2.

The central claim of this report is not that any particular tactic works or fails. It is that the standard evidence for all of them measures the author rather than the technique — and that when you remove the author, most of the advice goes with them.

2. Related Work

This study sits at the intersection of four literatures: empirical work on what makes online content spread, the statistical treatment of confounding in aggregated data, the methodology of social-media sampling, and the emerging practice of using language-model agents to conduct analysis. This chapter takes each in turn and states precisely what the present work borrows and where it departs.

2.1 What makes content spread

The modern empirical literature on virality is largely a literature of population-level regression on content features. [Berger & Milkman, 2012] established the template, showing across nearly seven thousand New York Times articles that emotionally arousing and positive content is more likely to be shared. On Twitter, [Suh et al., 2010] analysed 74 million tweets and found that URLs and hashtags are associated with higher retweet rates, while follower and followee counts dominate account-level variation. [Cheng et al., 2014] reframed the problem as cascade prediction, showing that the future growth of a cascade is substantially predictable from early structural signals.

These are careful studies and their findings are, within their designs, correct. The design they share is a comparison across authors: features are measured on posts, and posts by different people are pooled. The present study takes one of their headline results — that hashtags and links are associated with more sharing — and shows that on Threads the association does not survive holding the author fixed (hashtags 1.12, interval 0.79 to 1.68), while the link association not only survives but reverses in sign relative to [Suh et al., 2010], at 0.30 with an interval from 0.13 to 0.71. Whether this reflects a platform difference, a decade of ranking-system evolution, or the difference between the two designs cannot be settled here; the divergence is reported rather than explained away.

The closest methodological precursor is [Tan et al., 2014], which isolates the effect of word'ing on message propagation by constructing pairs of tweets that share both topic and author, and which found meaningful wording effects under that control. That paper's design is the one this study generalises: where Tan and colleagues built controlled pairs to test word'ing, the present work applies author fixed effects to the full battery of formatting, structural and vocabulary tactics that practitioner advice comprises, and reports how many survive. The contrast in outcome is instructive — careful wording effects survived their control; the formatting folklore examined here mostly does not.

Adjacent to this is the influencer-marketing literature, which measures the same outcomes with a commercial frame. [Hughes et al., 2019] shows that influencer-level characteristics dominate campaign engagement, and [Leung et al., 2022] surveys how engagement is

benchmarked in practice. Both are consistent with this study's central finding, from the opposite direction: if the creator dominates outcomes, then any post-level comparison that does not hold the creator fixed is measuring the creator.

2.2 Confounding, aggregation and the ecological fallacy

The statistical phenomenon underlying this study's results is old and well characterised. [Simpson, 1951] described how an association present in aggregated contingency tables can reverse when the data are partitioned; [Robinson, 1950] had already shown that correlations computed on group aggregates need not hold for the individuals composing them, an error that became known as the ecological fallacy. [Kievit et al., 2013] gives a practical guide to detecting the reversal in behavioural data and supplies the between- versus within-cluster diagnostics this study applies.

The econometric treatment is equally settled. [Mundlak, 1978] decomposes a pooled estimate into within- and between-unit components and shows the conditions under which the pooled coefficient is a biased estimate of the within-unit effect; [Angrist & Pischke, 2009] gives the standard modern account of why fixed-effects designs eliminate all time-invariant unit-level confounding. [Charness et al., 2012] makes the same point for experimental design, contrasting what between-subject and within-subject comparisons can identify.

The contribution here is not theoretical. It is to demonstrate that in one substantively important domain — practitioner guidance on a major social platform — the gap between the two estimates is not a technical refinement but the whole of the effect for twenty of twenty-three claims tested.

2.3 Sampling and validity in social-media data

A body of methodological work establishes that social-media datasets are rarely samples of what researchers assume. [Morstatter et al., 2013] compared Twitter's Streaming API against the full Firehose and found systematic, topic-dependent divergence. [Tufekci, 2014] catalogued the resulting threats to representativeness and validity, including the single-platform and single-hashtag conveniences that shape most published corpora. [Ruths & Pfeffer, 2014] warned that platform-specific sampling and filtering produce datasets whose apparent behavioural effects are artefacts of collection. [Olteanu et al., 2019] provides the comprehensive taxonomy, and its categories map directly onto this study's two structural limitations: search-surfacing bias, and what it terms population bias arising from the researcher's own query design.

The present study contributes a specific and, as far as the author is aware, under-documented instance of the latter: collection-pass confounding. Because a scraped corpus is typically assembled over multiple sessions with different query sets, and because different queries surface strata with radically different engagement levels, the pass in which a

post was collected can predict its engagement better than most of its content does. In this corpus the spread between passes is elevenfold. Section 4.2 documents the effect and Chapter 10 shows that it invalidated one of the study's own headline findings.

2.4 Ranking as an unobserved mediator

Engagement measured through a ranked feed is a joint product of the content and the ranking system that distributed it. The algorithm-audit literature formalises this: [Sandvig et al., 2014] set out the methods for studying platform algorithms from outside, [Bakshy et al., 2015] demonstrated on Facebook that ranking materially reshapes what users see and act on, and [Metaxa et al., 2021] surveys the current state of external audit design.

This study does not audit Threads' ranking system and cannot separate content effects from ranking effects. It inherits the standard consequence: every reported association is an association with engagement as distributed by the platform, and a tactic that changes ranking treatment is indistinguishable here from one that changes reader behaviour. Where this matters most is the outbound-link result of Section 6.4, which is commonly attributed to deliberate suppression of off-platform links; the present data establish the association and are silent on its mechanism.

2.5 Language-model agents as analysts

The analysis in this report was substantially conducted by language-model agents, which introduces a failure mode the older literature does not address. [Hu et al., 2024] benchmarks agents on end-to-end data-analysis tasks and finds competence alongside measurable error rates. [Huang et al., 2024] shows that intrinsic self-correction — asking a model to check its own reasoning without external feedback — can degrade rather than improve accuracy, which rules out the most convenient verification strategy. [Du et al., 2024] shows that having multiple model instances debate and challenge one another improves factuality, and [Silberzahn et al., 2018] demonstrates empirically that independent human analysts given identical data reach materially different conclusions.

The protocol in Chapter 10 follows directly from these four results: independent analysts rather than a single one, cross-agent refutation against the raw data rather than self-review, and full disclosure of what the refutation destroyed.

2.6 Reporting standards and researcher degrees of freedom

Finally, the study's decision to report its own refuted findings follows the replication literature. [Simmons et al., 2011] established that undisclosed flexibility in data collection and analysis inflates false-positive rates far above their nominal level, and [Gelman & Loken, 2014] showed that this holds even without deliberate p-hacking, because analytic choices contingent on the observed data create an implicit multiple-comparisons problem — the garden of forking paths.

An observational study of this kind is full of such forks: which posts count as mature, which keyword defines a subject lane, which threshold separates a thin sample from a reportable one. Chapter 10's inventory of refuted claims, and Section 3.4's statement of the reporting floor and multiplicity correction, are this study's response. They do not eliminate the problem; they make its extent visible.

2.7 Position of this work

In summary: this study adopts the outcome measures of the virality literature, the identification strategy of the fixed-effects literature, the scepticism of the social-data sampling literature, and the verification posture of the multi-agent and replication literatures. Its novelty is empirical rather than theoretical — a first substantial account of engagement on Threads, and a demonstration that the standard analytical approach in this area produces conclusions that mostly do not survive the most basic control available.

3. Methodology

This chapter describes how the corpus was collected, how effects were estimated, and what the design can and cannot identify. It is written to be reimplementable: a reader with access to a Threads account and a scriptable browser should be able to reproduce the collection, and a reader with the released data files should be able to reproduce every number in Chapters 5 to 9 from the estimator definitions given in Section 3.4.

3.1 Design

The study is observational and cross-sectional. No content was published for the purpose of the research, no account was manipulated, and no randomisation was possible. It follows that every effect reported here is an association, and that the strength of the design lies entirely in which alternative explanations it can eliminate by construction.

The design eliminates one class of alternative explanation directly, by making each creator their own control group wherever a creator varied the tactic under test. This is the panel-data logic of a fixed-effects specification applied to a corpus that was not collected as a panel: within an author, everything stable about the author — audience size and quality, subject specialism, language, skill, and whatever accumulated standing the ranking system assigns the account — is differenced away. The estimator is defined formally in Section 3.4 and applied in Chapter 7.

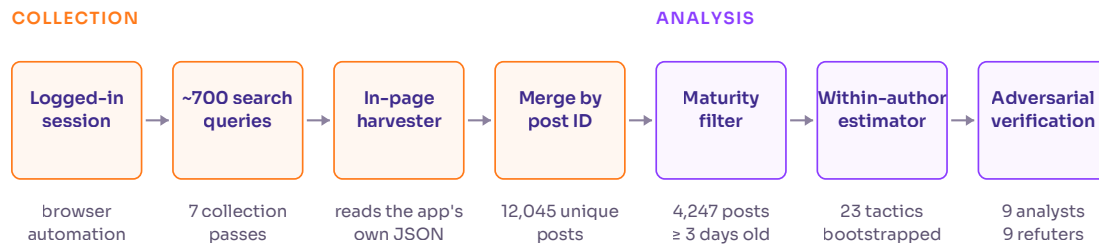
Where within-author comparison is impossible, the design falls back to stratification by data-collection pass, which eliminates the second-largest confound in a scraped corpus (Section 4.2) but leaves creator identity uncontrolled. Chapter 8 uses this weaker design and says so at every claim.

3.2 Collection

Threads exposes no public research API. Data was therefore collected through the platform's own web interface from an authenticated session, driven by browser automation, at a request rate comparable to ordinary human browsing.

The mechanism reads the data the application already ships to the client. Each server-rendered page embeds a JSON payload containing the post objects the page will display, and subsequent interactions fetch further payloads over the application's internal GraphQL endpoint. A small harvester script was injected into each page to walk both sources recursively, extract every object carrying a post identifier together with a like count, and record the fields listed in Section 4.1. **All engagement figures are therefore the platform's own counters at the moment of collection, not estimates inferred from rendered markup.**

STUDY PIPELINE: COLLECTION, MERGE, ESTIMATION, ADVERSARIAL VERIFICATION



Every engagement figure is the platform's own counter, read from the JSON the application ships with each page, not an estimate. The maturity filter exists because a post collected hours after publication has not finished accruing likes; the within-author estimator exists because creator identity dominates every raw comparison.

Figure 3.1 — The study pipeline. The first four stages are collection; the last three are analysis.

Approximately 700 distinct queries were issued across seven passes, combining: topic terms in the AI and technology space; adjacent professional lanes such as design, marketing and careers; general-interest terms; and equivalents in thirteen non-English languages. Both available result orderings — the platform's "top" and "recent" filters — were used. Search returns approximately twenty results per query and does not paginate for an automated session, so breadth of querying rather than depth of scrolling was the only available route to scale.

Records were merged by post identifier across passes, retaining the first-seen record and refreshing its engagement counters from any later pass that observed the same post. Appendix A documents the harvester, the three defects found in it during the study, and the compact-index design adopted after the browser storage limit was reached.

3.3 The maturity rule and the analysis set

Posts continue to accrue engagement after publication, so a corpus that mixes recent and settled posts confounds content with exposure duration. Every rate comparison in this report is restricted to posts at least three days old at collection, an analysis set of 4,247 posts. The threshold was selected by inspecting the accrual profile rather than by convention; Section 4.3 gives the reasoning and the reason age is treated as a nuisance variable throughout rather than as a subject of inference.

3.4 Estimators

Baseline and lift. The analysis-set median is 38 likes. A group's lift is its median divided by that baseline. Medians are used throughout in preference to means because the outcome is heavy-tailed (Section 5.1) and a single outlying post moves a group mean more than any effect measured in this study. Ratios of medians are reported rather than a standardised mean difference for the same reason; where a rank-based effect size is wanted, the stochastic-superiority measure of [Vargha & Delaney, 2000] is the appropriate companion statistic.

The within-author ratio. For a binary tactic, let A be the set of authors with at least one analysis-set post using it and at least one not using it. For each author compute the median likes in each condition, form the ratio $(m_1 + 1) \div (m_0 + 1)$, and report the median of that ratio over A . The added constant prevents division by zero and biases the statistic toward the null, making the test conservative.

Uncertainty. Intervals are 95 per cent percentile bootstraps over authors, 800 resamples, following [Efron, 1979] and the interval-construction guidance of [DiCiccio & Efron, 1996]. Resampling authors rather than posts is deliberate: the unit of independence is the account, since one prolific author's posts are not independent draws.

Significance. Paired comparisons carry a two-sided sign test [Dixon & Mood, 1946] on the number of authors whose ratio exceeds one. It is chosen over a paired t-test because the ratio distribution is skewed and heavy-tailed, and over a signed-rank test because it makes no symmetry assumption about the differences. It discards magnitude information, which is why the carousel row in Table 7.1 shows an interval excluding 1.00 alongside a sign test of 0.068; both are reported rather than choosing whichever is more favourable. Unpaired group comparisons use the Mann–Whitney U test [Mann & Whitney, 1947].

Multiplicity. Twenty-three tactics were tested in the primary battery. Under a Benjamini–Hochberg procedure at a false-discovery rate of 0.05 [Benjamini & Hochberg, 1995], two of the three survivors — attached media at $p < 0.001$ and outbound links at $p = 0.004$ — retain significance; the carousel at $p = 0.068$ does not, and is reported as surviving on its interval while failing the corrected sign test. This is stated explicitly rather than resolved by selecting the more favourable criterion.

Reporting floor. No group with fewer than twenty posts, and no paired comparison with fewer than fifteen authors, is reported as a finding. Where such groups are shown they are labelled thin and carry no weight in any conclusion.

3.5 Confound control

Four confounds recur and each has a specific remedy, applied consistently.

Table 3.1 — Confounds and controls. The right-hand column names the chapter in which the control is exercised and, where relevant, the finding it overturned.

Confound	Mechanism	Control applied	Where
Creator identity	Tactics correlate with author skill and audience	Within-author paired ratio	Ch. 7 — overturned 20 of 23 tactics
Collection provenance	Passes differ in engagement level elevenfold	Compute within pass; report pass by pass	Ch. 8 — overturned the brand penalty
Post age	Engagement accrues; old posts are survivors	Maturity floor; age-bucket stratification	Ch. 4, Ch. 6 — halved the link penalty
Composition	Groups differ on more than the test variable	Stratify by media type before comparing	Ch. 6 — link posts are mostly text posts

The pattern these controls exhibit is a form of Simpson's paradox [Simpson, 1951]: an association present in aggregate data reverses or disappears when the data are partitioned by a variable correlated with both the treatment and the outcome. Section 7.4 gives the clearest instance, in which a clean 1.53-fold association for bulleted lists becomes a point estimate of 0.80 once authors are held fixed.

3.6 Adversarial verification

Because the analysis was conducted with substantial assistance from language-model agents, and because such agents are fluent producers of plausible and wrong analysis, a verification stage was built into the design rather than added afterwards. Nine analyst agents each mined one dimension of the corpus; nine refuter agents were then instructed to destroy the first nine's headline claims by recomputing them independently and hunting for confounds, with an explicit instruction to default to refuted where evidence was thin.

Chapter 10 reports the protocol and its outcome in full, including the two cases in which the refuters themselves were wrong. The within-author battery of Chapter 7 was constructed after that exchange, specifically to adjudicate the disagreements it produced.

3.7 What this design cannot identify

Three limits are structural and no amount of additional data of this kind removes them.

The design cannot establish causation. Within-author variation is not random assignment: an author who sometimes attaches media may do so precisely when they have something worth showing, so the estimator may credit the attachment for the quality of the underlying material. The surviving results of Chapter 7 should be read as strong observational evidence, not as causal estimates.

The design cannot observe failure. A search-surfaced corpus contains no posts that were never surfaced, so no statement about base rates across all posting is available (Section 4.4).

The design cannot observe reach. Threads exposes no impression count on third-party posts, so "reach" throughout this report means likes, and any mechanism that operates on impressions without moving likes is invisible to it.

4. The Corpus

This chapter describes what was collected, how the collection was structured, and — at greater length than is customary — what the resulting sample cannot be used for. The corpus is the foundation of every number in Chapters 5 through 9, and it carries two structural properties that shape the whole analysis: it is a search-surfaced sample rather than a random one, and it was assembled in seven passes whose engagement levels differ by an order of magnitude.

4.1 Composition

The corpus comprises 12,045 unique Threads posts published between 6 July 2023 and 24 August 2026, authored by 9,922 distinct accounts, carrying 2,546,878 likes in total. Each record holds the platform's own counters for likes, replies, reposts and quotes, together with the media type, a verification flag, the presence of an outbound link preview, the publication timestamp, the post's canonical URL and its full text. [Table 4.1](#) summarises the file.

Table 4.1 — Corpus summary. The analysis set is the subset used for every rate comparison in this report; the maturity rule is defined in Section 4.3. Median and percentile figures are computed on the analysis set, because they are meaningless on a population that mixes finished posts with posts published hours before collection.

Quantity	Full corpus	Analysis set (≥ 3 days old)
Posts	12,045	4,247
Distinct authors	9,922	3,397
Total likes	2,546,878	1,704,431
Median likes	—	38
90th percentile likes	—	846
Largest single post	81,351	74,589
Posts carrying an outbound link	—	152 (3.6%)
Posts from verified accounts	—	994 (23.4%)
Authors appearing exactly once	—	3,014 (88.7%)

Media type is assigned by precedence from the attachment structure the application returns: a post carrying two or more images is a `carousel`, one carrying a video is `video`, one carrying a single image is `image`, and a post with no attachment is `text`. The distribu'

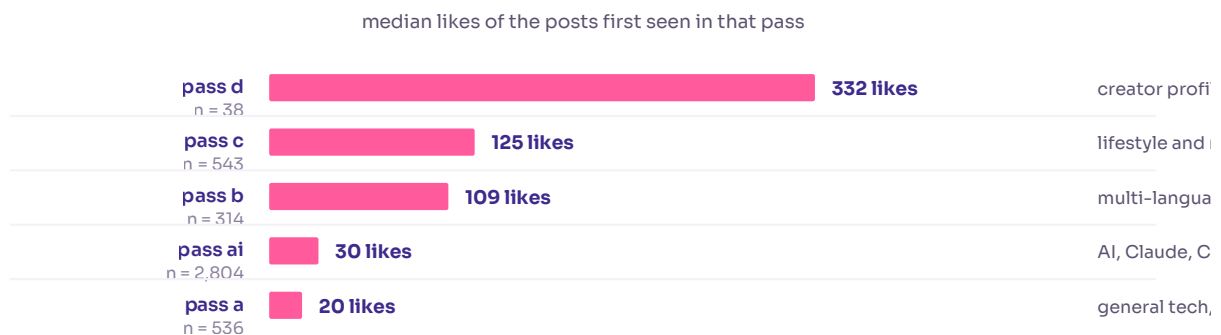
tion across the analysis set is 41.6 per cent text, 23.9 per cent image, 23.2 per cent carousel and 11.3 per cent video. Language is not a field the platform exposes; it is inferred from Uni' code script ranges, a method that separates Chinese, Japanese, Korean, Thai, Arabic and Devanagari cleanly but cannot distinguish Spanish from Portuguese, and which therefore reports a single Latin-script segment rather than pretending to a precision it does not have.

4.2 Collection passes and why they matter

The corpus was not collected in one sweep. It accumulated over seven passes, each driven by a different slice of the roughly 700-query list described in Section 3.2: an initial pass on AI, Claude, ChatGPT and prompting; a second on general technology, development and design; further passes on multi-language topics, lifestyle breadth and creator profiles; and two final passes on recency-filtered feeds. Posts were merged by post ID, with the earliest-seen record retained and its engagement counters refreshed from any later pass that saw the same post.

The passes are not interchangeable populations. [Figure 4.1](#) shows the median likes of the posts first seen in each, and the spread is eleven-fold.

WHY A MERGED CORPUS MUST BE ANALYSED PASS BY PASS



A post's collection pass predicts its like count better than most of its content does — an eleven-fold spread

Figure 4.1 — Median likes by collection pass, analysis set. A post's pass predicts its like count more strongly than most of its content features do.

The mechanism is not mysterious. Different queries surface different strata of the platform: a search for prompt engineering returns the practitioner mid-tail, while a search for a common Chinese lifestyle term returns posts from a segment with far higher typical engagement. Because query choice determined which posts entered the corpus, and query choice varied by pass, **any comparison between two groups that are drawn in different proportions from different passes is partly a measurement of the query list rather than of Threads.**

Methodological note

How this constraint is honoured. Three rules follow, and they are applied throughout Chapters 5 to 9. First, where a comparison can be made within a single author it is, because an author's posts are distributed across passes in a way that is unrelated to the tactic being tested. Second, where a within-author comparison is impossible — language, for instance, because authors do not alternate languages — the comparison is computed separately in side each pass and reported pass by pass, so the reader can see whether the effect replicates. Third, no claim about the supply of any content category is made anywhere in this report, because supply here is a property of the query list. Section 4.5 returns to this point, which cost the study several findings.

4.3 The maturity rule

Engagement accrues over time, so a corpus that mixes a post published in 2024 with one published four hours before collection will attribute to content what is actually a difference in exposure duration. All rate comparisons in this report are therefore restricted to posts at least three days old at the time of collection, which reduces the working set from 12,045 to 4,247.

Three days is a compromise between contamination and statistical power, and it was chosen from the data rather than by convention. Median likes rise from 4 in the first day to 10 by day three and 13 by day five; between day five and day thirty the median roughly doubles again, but the ordering of content categories is already stable by day three. A stricter rule of seven days would remove a further 890 posts for a marginal gain in stability.

An important caveat attaches to this. Median likes continue to climb across age bands well beyond any plausible accrual window — from 12 likes at 3–7 days to 119 at 91–365 days — and then fall to 59 beyond a year. Real like counts cannot decrease, so this is not an accrual curve. It is a survivorship curve: older posts appear in a search-surfaced corpus only if they are still being surfaced, which selects for the ones that performed. Age is therefore treated as a nuisance variable to be controlled, never as a finding.

4.4 What a search-surfaced sample is

Every post here entered the corpus because Threads' own search returned it for one of the queries. This has a precise consequence: the sample over-represents content that the ranking system already favoured, and it contains no observable population of failures. The corpus can rank one performing post against another; it cannot estimate what proportion of posts using a given tactic sank without trace, because those posts were never retrievable.

Every claim in this report should therefore be read with the silent prefix among posts that Threads surfaced. That prefix weakens the absolute rates in Chapter 5 considerably — the true share of posts clearing a thousand likes across all of Threads is certainly far below the 8.7 per cent measured here — while leaving the comparisons largely intact, since the surfacing bias applies to both arms of any within-corpus contrast.

4.5 What the corpus cannot measure

Four quantities are simply absent, and no amount of analysis recovers them.

Impressions and reach. Threads exposes no view count on posts by other authors. Every statement about "reach" in this literature is, in this dataset, a statement about likes. Where the word appears in Chapters 5 to 9 it is used loosely and the underlying measure is always stated.

Saves and shares to private channels. A post that is screenshotted or sent in a direct message registers nothing. For instructional content — which the AI lane produces heavily — this is likely to be a substantial fraction of its true value, and its absence biases the analysis against reference-style posts.

Follower counts at scale. Follower data was captured for only 59 creators, from profile pages visited during collection, and those creators were selected partly because their posts had performed. Any follower-stratified claim in this report rests on that small, non-random sample and is flagged accordingly. This is the single largest methodological gap relative to comparable studies on other platforms, which typically have follower counts for the whole corpus.

Supply. As established in Section 4.2, the number of posts in a category here is the number my queries retrieved. The study originally produced a set of "white space" findings ranking categories by demand relative to supply; adversarial verification established that the supply half of every such ratio was measuring the query list. Those findings were withdrawn rather than qualified, and Chapter 8 reports demand alone.

A corpus assembled by search is a sample of what a ranking system chose to show, taken through a keyhole whose shape the analyst cut. It supports comparisons between things it surfaced. It supports no statement at all about what it did not.

4.6 Ethics and data handling

All posts analysed are public, retrieved through the platform's own interface from an authenticated session belonging to the author, at a request rate comparable to ordinary human browsing. No private, restricted or deleted content was accessed, no attempt was made to identify any individual beyond the public handle attached to a public post, and no

personal data beyond public profile fields was retained. Post text is reproduced in this report only where it is quoted as evidence for a specific claim, in short extracts, with attribution to the public handle — the standard for research commentary on public speech.

The derived dataset is distributed with the report as three files: the full post table, the top 750 posts by likes, and the 59-creator profile table. The collection code is described in Appendix A in sufficient detail to reimplement, and is not distributed as a runnable artefact.

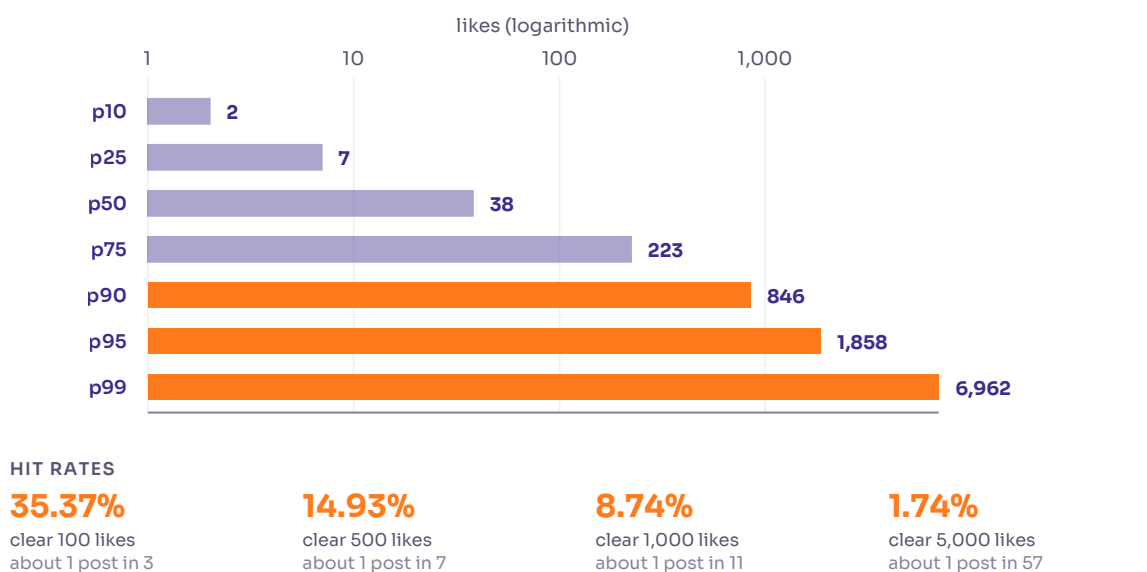
5. The Shape of Engagement

Before any tactic can be evaluated, the distribution it operates in has to be described. This chapter establishes three facts that govern everything after it: engagement on Threads is extremely concentrated, the concentration is not held by a stable elite, and the correct unit of judgement for an individual post is therefore a percentile rather than an average.

5.1 The distribution

Across the 4,247 posts in the analysis set, the median post earns 38 likes and the mean earns 401. When a mean is an order of magnitude above its median, the mean is describing a tail rather than a typical case, and reporting it as "average engagement" would mislead every reader who does not check. [Figure 5.1](#) gives the percentile ladder.

HOW LIKES ARE DISTRIBUTED ACROSS THE 4,247 MATURE POSTS



Half of all mature posts never reach 38 likes while the top tenth take 79 per cent of every like awarded. Percentiles, not averages, are the correct scoreboard for an operator: a 223-like post is a 75th-percentile post.

Figure 5.1 — The like distribution across the analysis set, plotted on a logarithmic axis, with the hit rate at four thresholds beneath.

The ladder spans three and a half orders of magnitude between its ends. A quarter of surfaced posts never reach 7 likes. Half never reach 38. The 90th percentile sits at 846 and the 99th at 6,962, so the gap between a good post and a great one is larger than the gap between a poor post and a good one.

Table 5.1 — Percentiles and threshold hit rates, analysis set (n = 4,247). The "one post in" column is the reciprocal of the hit rate and is the practically useful form: it converts a probability into an expected number of attempts.

Percentile	Likes	Threshold	Share of posts clearing it	One post in
p10	2	≥ 100 likes	35.4%	3
p25	7	≥ 500 likes	14.9%	7
p50	38	≥ 1,000 likes	8.7%	11
p75	223	≥ 5,000 likes	1.7%	57
p90	846	—	—	—
p95	1,858	—	—	—
p99	6,962	—	—	—

These hit rates are inflated by the search-surfacing bias described in Section 4.4 and should be read as rates within surfaced content, not across all posting. Their internal structure is nonetheless informative: the ratio between successive thresholds is roughly constant, which is the signature of a heavy-tailed distribution rather than one with a characteristic scale.

5.2 Concentration

The top decile of the analysis set — 424 posts, cut at 849 likes — accounts for **79.0 per cent of every like in the set**. The top one per cent accounts for 34.3 per cent. An operator's outcome over a quarter is therefore dominated by the two or three posts that broke out, and a strategy that reliably raises the median while never producing a tail event will underperform one that occasionally produces tail events.

This is not a novel observation about social platforms in general. What is specific, and what matters for anyone deciding whether to compete here, is the second fact: the concentration is not held by a stable elite. Those 424 top-decile posts were written by 391 distinct accounts. Roughly three quarters of those accounts appear exactly once in the entire analysis set. The tail is not a club; it is a lottery with a very large number of one-time winners.

Methodological note

Two readings of the same number. The concentration statistic is often deployed to argue that platforms are winner-take-all and that new entrants are wasting their time. The authorship statistic points the other way: if 391 different accounts produced the top 424 posts, and 88.7 per cent of accounts in the analysis set appear only once, then the barrier to a breakout post is evidently not accumulated audience. Both readings are supported by this corpus. They are reconciled in Chapter 9, which shows that the durable asset is not the breakout post but what follows it.

5.3 What follows for measurement

Three consequences run through the rest of the report.

Medians, not means. Every effect in Chapters 6 to 9 is reported as a ratio of medians. With this distribution a mean-based effect size is a report on which arm happened to contain the largest outlier — a single 74,589-like post moves a group mean of a few hundred posts by more than any content effect measured here.

Intervals, not point estimates. Because the underlying variable is heavy-tailed and the within-author subsamples are small, every headline effect carries a bootstrap interval, and effects whose intervals span 1.00 are reported as unproven rather than as small. Chapter 7 depends on this distinction entirely.

Hit rate as a second dependent variable. A tactic can raise the median while lowering the chance of a tail event, or the reverse. Where the two diverge the report says so; the divergence is itself a finding, and Section 6.3 contains the clearest instance of it.

Judge a post against the ladder, not against an average. At 223 likes a post sits in the top quartile of everything the platform surfaced; read as a shortfall against a mean of 401, the same post looks like a failure.

6. Format, Media and the Currency of Engagement

This chapter reports the container: whether a post carries an attachment, what kind, and what that choice does both to the volume of engagement and to its composition. Format is the largest content-level effect in the corpus, and — unusually for this study — it is one of the few that survives the tests in Chapter 7. It is also the clearest case in the data of a variable that changes not just how much engagement a post attracts but what kind.

6.1 Media type and volume

[Table 6.1](#) gives the population-level picture across the analysis set.

Table 6.1 — Media type performance, analysis set (n = 4,247, baseline median 38 likes). Lift is the group median divided by the baseline. The divergence between median and mean in the text row is the point of the table and is discussed below.

Media	n	Share %	Median likes	Lift	Mean likes	Clear 500 likes
Carousel	985	23.2	102	2.68×	600.7	22.6%
Video	478	11.3	52	1.37×	521.2	19.7%
Image	1,016	23.9	40	1.05×	387.3	11.5%
Text	1,768	41.6	19	0.50×	398.1	11.3%

Three features deserve comment. First, the most-used format is the worst-performing one on the median: plain text is 41.6 per cent of the analysis set and sits at half the baseline. Second, the spread between best and worst median is 5.4-fold, which is larger than any language-independent content effect measured elsewhere in this study. Third, and least obvious, **the text row's mean is not low**. At 398 likes it is within a quarter of the carousel's mean while its median is a fifth of the carousel's. Plain text is a high-variance format: it usually fails and occasionally produces the largest posts in the corpus, five of the top ten among them.

The practical reading is not "never post text". It is that the two formats occupy different risk positions. A carousel raises the floor; a text post buys a lottery ticket whose expected value is carried almost entirely by its tail.

6.2 The currency: reposts against replies

Volume is only half of what format decides. The composition of engagement differs so sharply between formats that they are better understood as different instruments than as better and worse versions of the same one. [Figure 6.1](#) and [Table 6.2](#) report reposts and replies per hundred likes, computed on posts that cleared 100 likes so that the ratios are stable.

WHAT A HUNDRED LIKES TURNS INTO, BY FORMAT

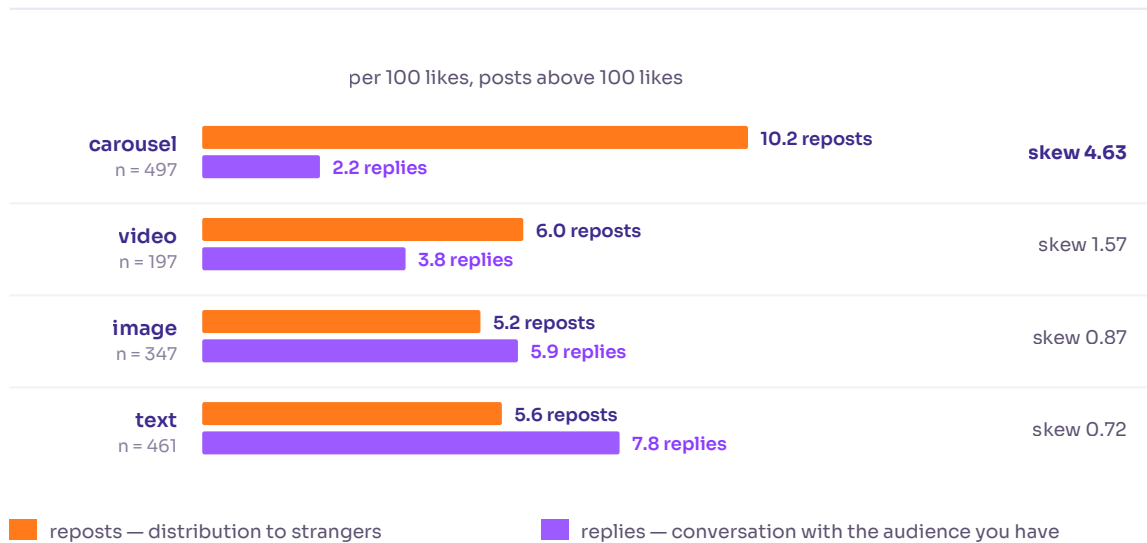


Figure 6.1 — Reposts and replies per hundred likes by format, posts above 100 likes. Skew is reposts divided by replies.

Table 6.2 — Engagement composition by format, posts above 100 likes in the analysis set (n = 1,502). Each cell is the median of the per-post ratio within the group, not a ratio of the group's medians; the two differ and the former is the correct summary of a typical post.

Media	n	Reposts / 100 likes	Replies / 100 likes	Skew (reposts ÷ replies)	Instrument
Carou- sel	497	10.2	2.2	4.63	distribution
Video	197	6.0	3.8	1.57	distribution, weak- er
Image	347	5.2	5.9	0.87	balanced
Text	461	5.6	7.8	0.72	conversation

The skew spans 6.4-fold from carousel to text. A repost places a post in front of an audience that does not follow its author, so it is the mechanism by which strangers arrive; a reply is a conversation with an audience that has already arrived. On this evidence the carousel is an acquisition instrument and the plain text post is a retention instrument, and an account that runs only one of them is doing only half a job.

Methodological note

How far this survives a stricter test. The population comparison in Table 6.2 could in principle be a creator effect: perhaps accounts that make carousels are simply the kind of account that gets reposted. Ten authors in the analysis set cleared 100 likes on both a carousel and a text post, which permits a within-author check. Seven of the ten had the higher repost ratio on their carousel, with within-author medians of 13.1 reposts per hundred likes on carousels against 6.4 on text. Ten authors is far too thin to establish a magnitude; it is enough to say the direction is not an artefact of who posts what, and not enough to say more than that.

6.3 Where the median and the tail disagree

Section 6.1 noted that plain text has a poor median and a competitive mean. Table 6.3 makes the pattern explicit by reporting the two dependent variables side by side.

Table 6.3 — Median against tail behaviour by format. The mean-to-median ratio is a crude dispersion index; higher values indicate that the format's average is carried by a small number of very large posts.

Media	Median likes	Mean likes	Mean ÷ median	Clear 500	Reading
Carousel	102	600.7	5.9	22.6%	raises the floor and the ceiling
Video	52	521.2	10.0	19.7%	moderate floor, strong ceiling
Image	40	387.3	9.7	11.5%	unremarkable on both
Text	19	398.1	21.0	11.3%	low floor, intact ceiling

The carousel is the only format that improves the floor and the ceiling together. Video's dispersion index of 10.0 marks it as a format whose average is a tail phenomenon — a reading confirmed in Chapter 7, where video's apparent advantage disappears entirely under a within-author test while the carousel's does not.

6.4 Outbound links

Posts carrying an outbound link preview are 3.6 per cent of the analysis set and run a median of 7 likes against 40 for posts without one, a raw ratio of 0.18. This is the single largest negative association in the corpus, and it is also the one most exposed to a composition objection: link posts are disproportionately plain text (74 per cent, against 41.6 per cent of the corpus) and disproportionately young.

Stratifying by media type and age bucket moves the estimate to 0.42 with a bootstrap interval of 0.26 to 0.70 — that is, a penalty of roughly 58 per cent rather than 82 per cent, with the interval admitting a penalty as small as 30 per cent. The within-author test in Chapter 7 gives 0.30 with an interval of 0.13 to 0.71 across the 41 authors who posted both ways. All

three estimates agree on a substantial penalty; they disagree on its size by a factor of two. The honest summary is that an outbound link in the post body is associated with losing somewhere between a third and three quarters of a post's likes, and that the effect survives every control the corpus supports.

Format is the largest content lever measured here, and the only one that survives a within-author test. It is also the one most commonly discussed as though it were a matter of taste.

6.5 Verification and length

Two secondary associations complete the chapter. Posts from verified accounts run a median of 72 likes against 30 for unverified — a ratio of 2.4 — but 70.4 per cent of top-decile posts come from unverified accounts, so verification is better understood as a correlate of account maturity than as a lever. Post length shows a monotone negative association at population level, from a median of 55 likes below 120 characters to 15 above 500; that association does not survive the within-author test, and Chapter 7 reports it among the casual ties.

7. The Within-Author Test

This is the central chapter of the study. It applies one idea — make each creator their own control group — to twenty-three tactics that population-level analysis rates as effective or harmful, and reports that twenty of them do not survive. The chapter is organised around the estimator, the result, and an extended discussion of what a null result does and does not license, because the last of those is where this kind of analysis is most often over-read.

7.1 The identification problem

Chapter 6 reported that posts with attached media out-earn plain text posts. That comparison, and every comparison like it, has a structural weakness: the analyst chose the posts, but the platform's creators chose the tactic. Posts are not randomly assigned to techniques. The accounts that write in tight bulleted lists are disproportionately the accounts that have been publishing for years, that have accumulated an audience, that photograph well and that have already learned what their readers respond to.

When such an account's bulleted post is compared against a stranger's unformatted one, the comparison contains the tactic and everything else that differs between the two authors. Formally, with a post-level outcome y , a tactic indicator x and an unobserved author quality a , the population estimate of the effect of x is biased by the covariance between x and a . Nothing about sample size fixes this: collecting a hundred thousand posts estimates the confounded quantity more precisely.

The corpus offers a direct remedy. Many creators appear in it more than once, and many of those varied the tactic across their own posts. Restricting attention to those creators and comparing each one against themselves differences out a entirely, along with every other stable property of the account: audience size, audience quality, topic specialism, language, posting habits, photographic skill and the platform's accumulated opinion of the account.

7.2 The estimator

For a binary tactic x , let A be the set of authors with at least one analysis-set post where $x = 1$ and at least one where $x = 0$. For each author $a \in A$ compute the median likes in each condition, m_{a1} and m_{a0} , and form the ratio

$$r_a = (m_{a1} + 1) \div (m_{a0} + 1)$$

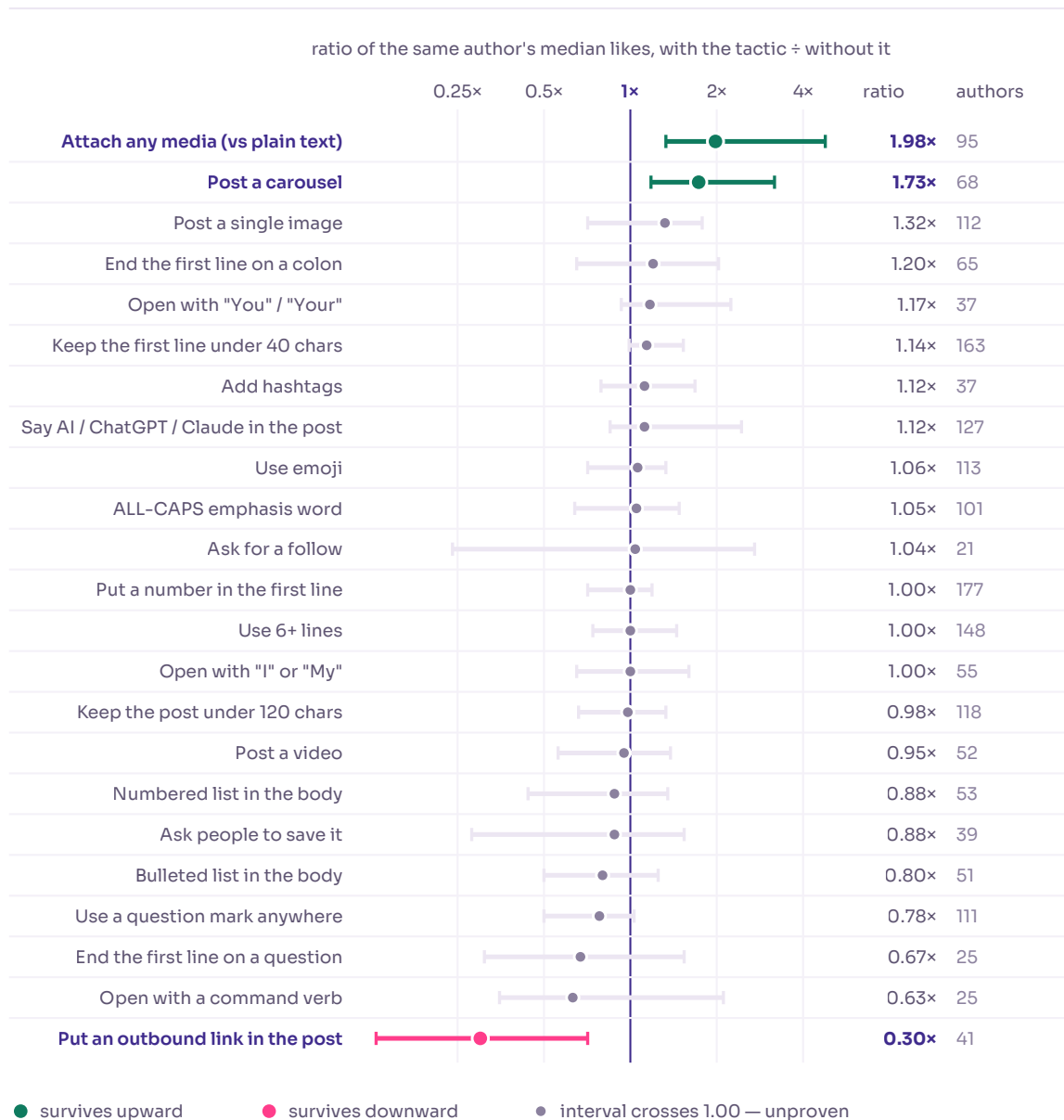
The reported effect is the median of r_a over A , with a 95 per cent percentile bootstrap interval from 800 resamples of authors, and a two-sided sign test on the number of authors whose ratio exceeds 1. The added constant guards against division by zero for authors whose median in one condition is zero; it biases ratios toward 1 for very low-engagement authors, which makes the test conservative rather than permissive.

Three properties of this estimator matter for reading the results. It weights every author equally rather than every post, so a prolific account cannot dominate. It is a paired test, so its power depends on the number of authors who varied the tactic, not on the number of posts. And it is silent by construction on any tactic that authors do not vary: language can't not be tested this way, because creators do not alternate languages between posts, and it is handled separately in Chapter 8.

7.3 Result

[Figure 7.1](#) reports all twenty-three tactics with sufficient paired authors, and [Table 7.1](#) gives the numbers.

DO THE TACTICS SURVIVE WHEN THE SAME AUTHOR DROPS THEM? (23 TESTED)



Each row compares one tactic against itself inside the same account: for every author who both used and skipped the tactic, the ratio of their median likes in the two conditions, summarised as the median ratio across authors with a 95 per cent bootstrap interval (800 resamples). Three of the twenty-three clear 1.00.

Figure 7.1 — Within-author effect of each tactic, ordered by point estimate. Intervals are 95 per cent percentile bootstraps over authors. Three of twenty-three exclude 1.00.

Table 7.1 — The full within-author battery. "Population" is the group median divided by the 38-like baseline, the quantity a conventional analysis reports. "Within-author" is the median paired ratio of Section 7.2. "Authors" counts creators who did the tactic in some posts and not in others. Verdict is SURVIVES where the interval excludes 1.00 and UNPROVEN otherwise.

Tactic	Posts	Popula- tion	Au- thors	Improved %	Within-au- thor	95% CI	p	Verdict
Attach any media (vs plain text)	2,479	1.55×	95	68	1.98×	1.33–4.78	<0.001	SUR-VIVES
Post a carousel	985	2.68×	68	62	1.73×	1.18–3.18	0.068	SUR-VIVES
Post a single image	1,016	1.05×	112	54	1.32×	0.71–1.78	0.395	un-proven
End the first line on a colon	251	1.24×	65	57	1.20×	0.65–2.03	0.321	un-proven
Open with "You" / "Your"	131	0.74×	37	51	1.17×	0.93–2.24	1.000	un-proven
Keep the first line under 40 characters	1,538	1.64×	163	55	1.14×	0.99–1.53	0.273	un-proven
Add hashtags	664	0.92×	37	54	1.12×	0.79–1.68	0.743	un-proven
Say AI / ChatGPT / Claude in the post	1,575	0.63×	127	52	1.12×	0.85–2.44	0.723	un-proven
Use emoji	1,699	1.05×	113	51	1.06×	0.71–1.33	0.851	un-proven
ALL-CAPS emphasis word	554	0.89×	101	51	1.05×	0.64–1.48	0.842	un-proven
Ask for a follow	98	1.04×	21	52	1.04×	0.24–2.71	1.000	un-proven
Put a number in the first line	1,174	1.26×	177	49	1.00×	0.71–1.19	0.764	un-proven
Use six or more lines	1,625	1.00×	148	49	1.00×	0.74–1.45	0.935	un-proven
Open with "I" or "My"	297	0.66×	55	49	1.00×	0.65–1.60	1.000	un-proven
Keep the post under 120 characters	1,331	1.45×	118	49	0.98×	0.66–1.33	0.927	un-proven
Post a video	478	1.37×	52	50	0.95×	0.56–1.38	1.000	un-proven
Numbered list in the body	273	2.05×	53	43	0.88×	0.44–1.35	0.410	un-proven
Ask people to save it	183	0.89×	39	44	0.88×	0.28–1.54	0.522	un-proven
Bulleted list in the body	169	1.53×	51	43	0.80×	0.50–1.25	0.401	un-proven
Use a question mark anywhere	845	0.47×	111	42	0.78×	0.50–1.03	0.128	un-proven

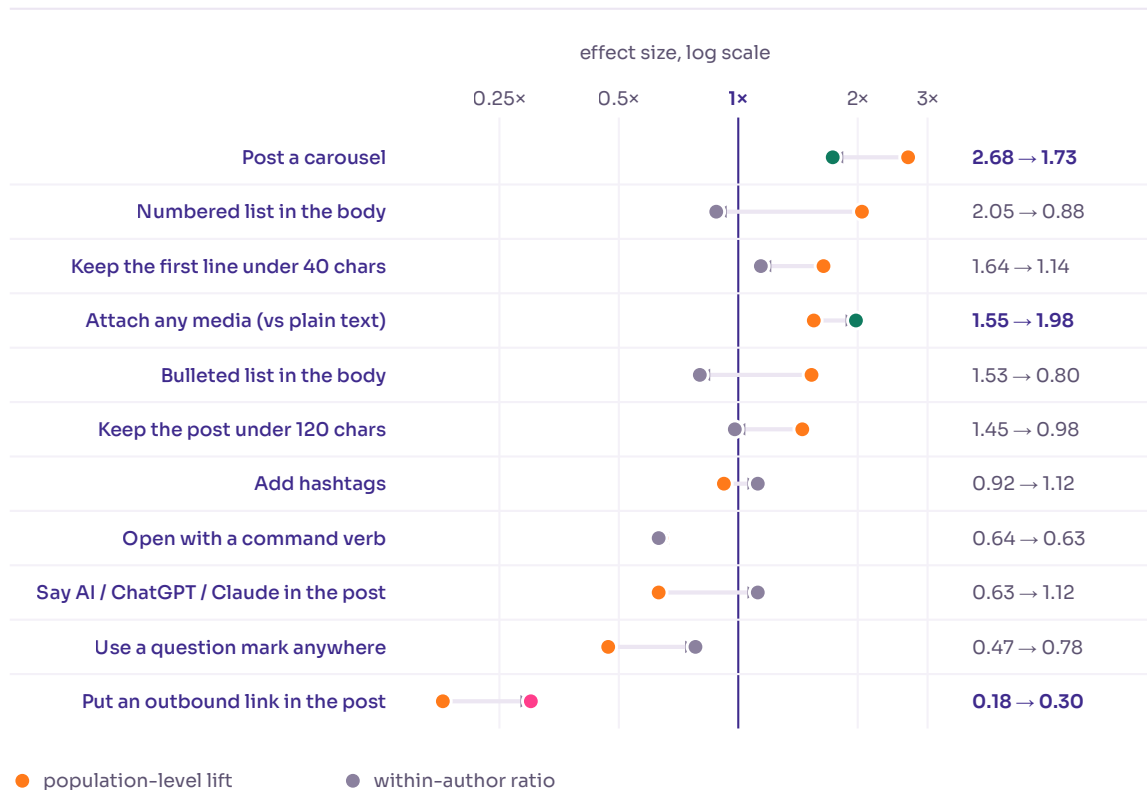
Tactic	Posts	Population	Authors	Improved %	Within-author	95% CI	p	Verdict
End the first line on a question	286	0.39×	25	40	0.67×	0.31–1.54	0.424	un-proven
Open with a command verb	62	0.64×	25	44	0.63×	0.35–2.11	0.690	un-proven
Put an outbound link in the post	152	0.18×	41	27	0.30×	0.13–0.71	0.004	SUR-VIVES

Three tactics clear the bar. Attaching any media to a post is associated with roughly doubling the same author's median, with 68 per cent of the ninety-five paired authors improving and a sign test below 0.001. Posting a carousel specifically returns 1.73 times the same author's non-carousel median across sixty-eight authors, on an interval whose lower bound of 1.18 sits above 1.00 while its sign test at 0.068 does not quite clear conventional significance — a discrepancy that reflects the sign test discarding magnitude information, and one that should temper confidence in the precise multiplier rather than in the direction. Including an outbound link is associated with losing about seventy per cent of the same author's median, with only 27 per cent of the forty-one paired authors doing better when they included one.

7.4 The scale of the illusion

The interesting quantity is not the survivors but the distance travelled by the casualties. [Figure 7.2](#) plots the population estimate and the within-author estimate for eleven representative tactics.

WHAT THE POPULATION SAYS, AND WHAT THE SAME AUTHOR SAYS



Arrows run from the effect a naive population comparison reports to the effect the same authors show when they drop the tactic. Eight of the eleven collapse toward 1.00; only the media and link rows hold their distance from it. The gap between the two dots is the size of the creator-identity illusion.

Figure 7.2 — From the population estimate to the within-author estimate. The length of each arrow is the portion of the apparent effect attributable to who was using the tactic.

Numbered lists move from 2.05 to 0.88. Emoji, for which a positive association with brand engagement is reported in the marketing literature [McShane et al., 2021], move from 1.05 to 1.06 — that is, to nothing at all. Bulleted lists move from 1.53 to 0.80 — that is, from a substantial apparent benefit to a point estimate on the wrong side of 1.00. Short first lines move from 1.64 to 1.14. The AI vocabulary penalty, which a population reading puts at 0.63 and which would license the advice stop saying AI, moves to 1.12 with an interval from 0.85 to 2.44: no detectable effect in either direction.

The pattern is consistent and it has a simple interpretation. These are markers of an experienced writer. Accounts that use structured formatting, short openings and disciplined language are accounts that have practised, and their posts do better for reasons the formatting merely signals.

Methodological note

A worked example. The bulleted-list result is worth following through. Across the analysis set, posts with a bulleted list earn a median of 58 likes against 38 overall — a clean, significant, entirely real 1.53× association that any conventional analysis would report as a finding, and that the first version of this research did report. Fifty-one authors in the corpus wrote both bulleted and unbulleted posts. Among them, the median author does 0.80 times as well when they use bullets, and only 43 per cent do better. Both numbers are correct. They answer different questions: the first asks what kind of post tends to do well, the second asks what happens if you change your post. Only the second is advice.

7.5 What a null does not mean

Twenty tactics returned intervals spanning 1.00. That is a failure of evidence, not a demonstration of absence, and the distinction is load-bearing.

The test is underpowered by construction. It can only use authors who appear more than once in a search-surfaced corpus and who varied the tactic — between 21 and 177 authors per row, with a median of 68. At those sizes, and with a heavy-tailed outcome, an effect of ten or twenty per cent is invisible. Several rows are consistent with a real and useful effect: single images at 1.32 (0.71–1.78), colon-led openings at 1.20 (0.65–2.03), short first lines at 1.14 (0.99–1.53) whose lower bound sits a hair below 1.00.

What the chapter establishes is therefore narrower and more useful than "these tactics do not work". It is that **the evidence usually offered for them does not survive the most basic control**, and that the burden of proof has shifted. A practitioner allocating scarce attention should spend it on the three tactics that cleared the bar before spending it on the twenty that did not, and should treat any confident claim about the remaining twenty — including claims made in the earlier, population-level version of this research — as unproven.

A second limitation applies to the survivors as much as the casualties. Within-author variation is not random assignment. An author who sometimes attaches media and sometimes does not may attach media precisely when they have something worth showing, in which case the estimator credits the format for the quality of the underlying material. This study cannot separate the two, and readers should hold the media results as strong observational evidence rather than as a causal estimate.

Twenty of twenty-three tactics measure the author rather than the technique. The three that survive are all about what is attached to the post, not how the post is written.

7.6 Timing, tested the same way

Posting time is treated separately because it is continuous rather than binary, but the same logic applies and the result is the same. The raw data appear to support a strong recommendation: the best three-hour window in the analysis set runs at a median of 53.5 likes against 32 in the worst, a 1.67-fold spread, and the best single hour reaches 69 against 23, a threefold spread.

Scoring each post against its own creator's median instead, across the 147 creators with three or more analysis-set posts, all eight three-hour windows fall between the 47.4th and 52.3rd percentile of their author's distribution, and every interval spans no effect. In a variance decomposition on log likes, creator identity accounts for 46.7 per cent, post age for 19.4 per cent, hour of day for 5.1 per cent across twenty-four free parameters, a three-hour window for 2.7 per cent and weekday for 1.2 per cent. The apparent timing effect is creator mix: some kinds of account post at some hours.

Posting cadence behaves identically. Across 157 creators with four or more posts, the rank correlation between an author's post count and their median likes is 0.068, and it falls to -0.012 once the number of posts the corpus happens to contain for that author is controlled.

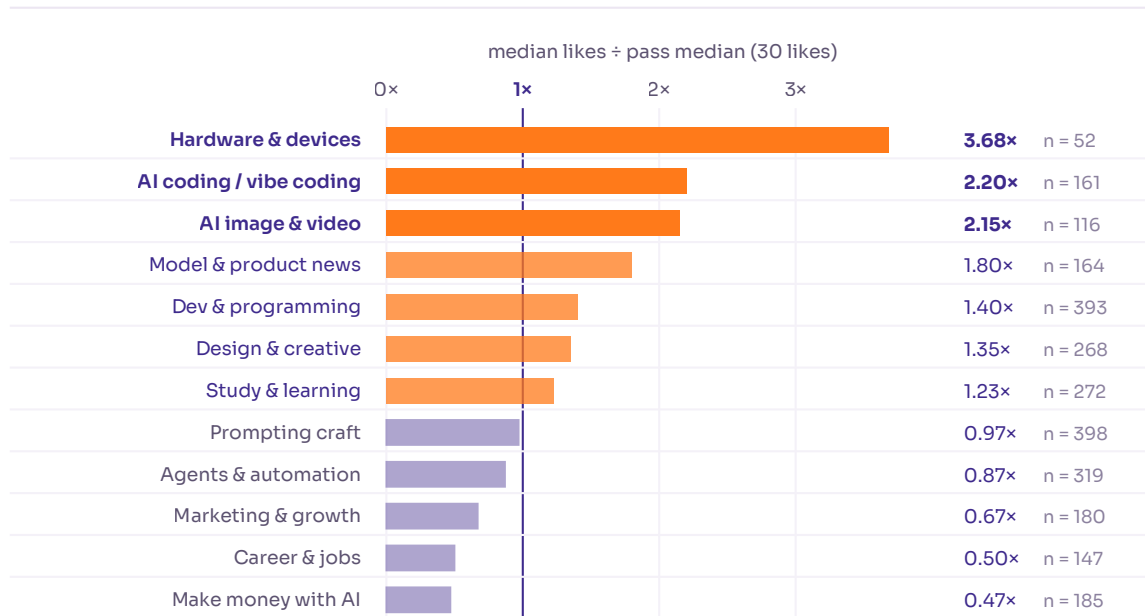
8. Subject, Tool and Language

Chapter 7 established that the writing tactics an operator can vary post to post mostly do not survive a within-author test. This chapter turns to the choices that sit above the post — what to write about, which products to attach the work to, and which language market to serve. These cannot be tested within an author, because creators do not switch subject or language between posts in any quantity. They are tested instead within collection pass, which removes the provenance confound of Section 4.2 while leaving other confounds in place. The chapter is correspondingly more cautious in its claims than Chapter 7 and more explicit about what remains uncontrolled.

8.1 Subject lanes

Twelve subject lanes were defined by keyword pattern and scored inside the AI collection pass alone ($n = 2,804$, median 30 likes), so that no part of the comparison reflects which query found the post. Lanes are not mutually exclusive; a post about building a tool with Cursor appears in both the coding and the tooling lane. [Figure 8.1](#) and [Table 8.1](#) report the result.

MEDIAN LIKES BY SUBJECT LANE, INSIDE ONE COLLECTION PASS



All twelve lanes measured inside the AI collection pass ($n = 2,804$, median 30 likes) so that no part of the comparison is contaminated by which search terms found the post. Lanes where a post can carry visible evidence — hardware, coding, image and video — outrank lanes that can only describe a method.

Figure 8.1 — Median likes by subject lane, computed inside the AI collection pass against its median of 30 likes.

Table 8.1 — Subject lane performance inside the AI collection pass (n = 2,804, median 30 likes). Authors counts distinct accounts contributing to the lane; the busiest-author share is the proportion of the lane's posts written by its single most active account, and is reported because a lane dominated by one voice would be a different kind of object from a lane that is genuinely diffuse.

Lane	n	Median likes	Lift	Clear 1k	Authors	Busiest author
Hardware and devices	52	110	3.68×	17.3%	48	5.8%
AI coding and vibe coding	161	66	2.20×	11.8%	129	3.1%
AI image and video	116	64	2.15×	12.1%	94	4.3%
Model and product news	164	54	1.80×	9.1%	134	3.7%
Development and programming	393	42	1.40×	7.4%	319	1.5%
Design and creative	268	40	1.35×	6.7%	247	2.2%
Study and learning	272	37	1.23×	5.9%	230	2.6%
Prompting craft	398	29	0.97×	4.5%	300	2.8%
Agents and automation	319	26	0.87×	4.1%	260	2.5%
Marketing and growth	180	20	0.67×	3.3%	156	2.8%
Career and jobs	147	15	0.50×	2.7%	130	6.1%
Make money with AI	185	14	0.47×	2.2%	158	3.2%

The spread from top to bottom is 7.8-fold, larger than any tactic effect in Chapter 7. The ordering has a legible structure: lanes in which a post can carry visible evidence — a device, a screen recording of code, a generated image — occupy the top three positions, while lanes that can only describe a method or promise an outcome occupy the bottom three. The worst-performing lane in the corpus is the one that promises money.

That structure is consistent with the surviving result of Chapter 7, which is that attaching something to look at is the one writing-level choice that reliably matters. A lane is, in part, a constraint on what can be attached.

Methodological note

What this table is not. It measures demand and nothing else. It is tempting to divide each lane's engagement by its share of posts and call the result an opportunity score — the study did exactly that in an earlier pass, and produced a confident ranking of "white space" lanes. That ranking was withdrawn during verification, because the share of posts in a lane is the share of my query list devoted to it. I searched prompting far more often than hardware, so hardware appears scarce by construction. No inference about supply, competition or saturation can be drawn from this corpus, and none is drawn here.

8.2 The tool map

Which products a post names is a strategic choice for a creator in this niche, and the corpus separates the named tools sharply. Table 8.2 and Figure 8.2 report every tool named in at least twenty posts within the AI pass.

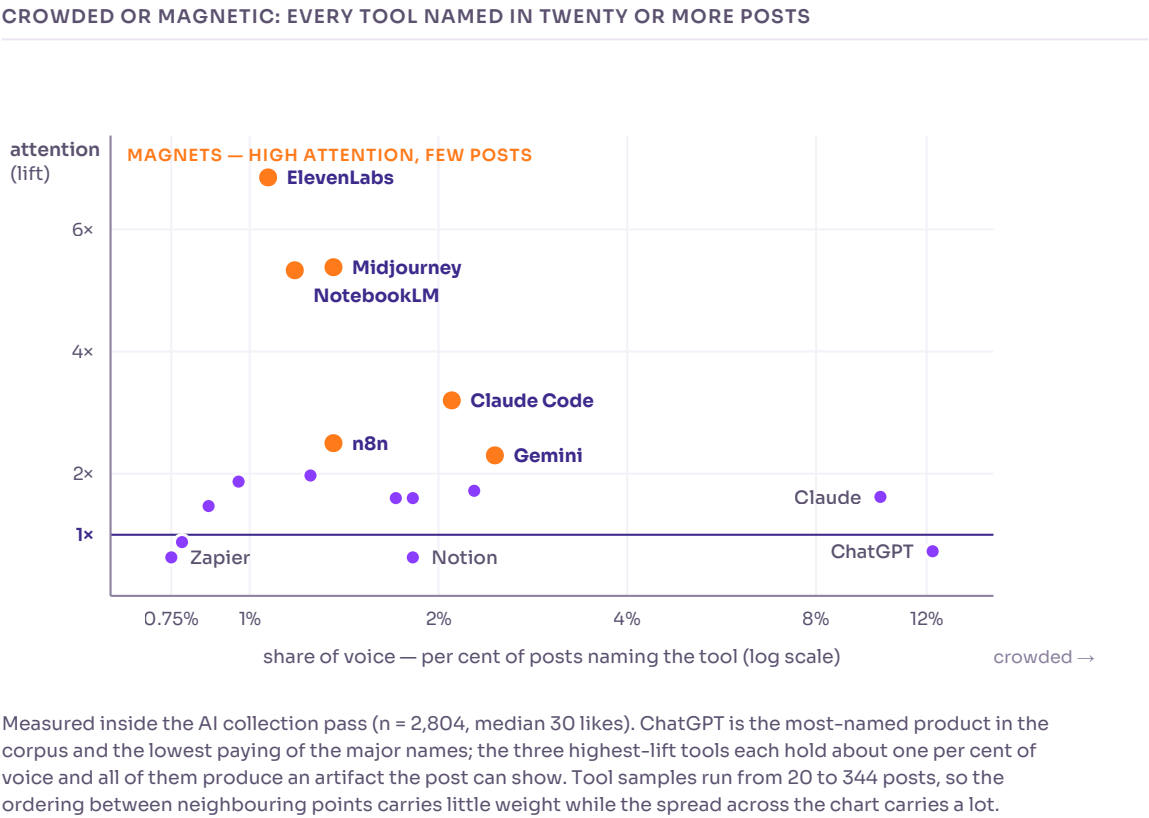


Figure 8.2 — Share of voice against attention per post, seventeen tools, inside the AI collection pass. The horizontal axis is logarithmic.

Table 8.2 — Tools named in twenty or more posts inside the AI collection pass (n = 2,804, median 30 likes). Share of voice is the percentage of that pass's posts naming the tool. Samples are small and the ordering between adjacent rows carries little weight; the spread across the table carries a great deal.

Tool	Posts	Share of voice	Median likes	Lift	Clear 1k
ElevenLabs	30	1.07%	206	6.85×	16.7%
Midjourney	38	1.36%	162	5.38×	18.4%
NotebookLM	33	1.18%	160	5.33×	9.1%
Claude Code	59	2.10%	96	3.20×	10.2%
n8n	38	1.36%	75	2.50×	7.9%
Gemini	69	2.46%	69	2.30×	13.0%
Cursor	35	1.25%	59	1.97×	11.4%
Copilot	27	0.96%	56	1.87×	0.0%
Perplexity	64	2.28%	52	1.72×	10.9%
Claude	284	10.13%	48	1.62×	6.7%
OpenAI	51	1.82%	48	1.60×	3.9%
Canva	48	1.71%	48	1.60×	6.2%
Excel	24	0.86%	44	1.47×	8.3%
Figma	22	0.78%	26	0.88×	4.5%
ChatGPT	344	12.27%	22	0.73×	6.4%
Zapier	21	0.75%	19	0.63×	14.3%
Notion	51	1.82%	19	0.63×	7.8%

ChatGPT is the most-named product in the corpus and the lowest-performing of the major names: 12.3 per cent of the pass's posts and a lift of 0.73. Claude sits at 10.1 per cent share and 1.62; the narrower term Claude Code, at a fifth of that share, returns 3.20. The three highest-lift tools — ElevenLabs, Midjourney and NotebookLM — each hold roughly one per cent of voice, and all three produce an artefact a post can display: a synthesised voice, an image, a generated document.

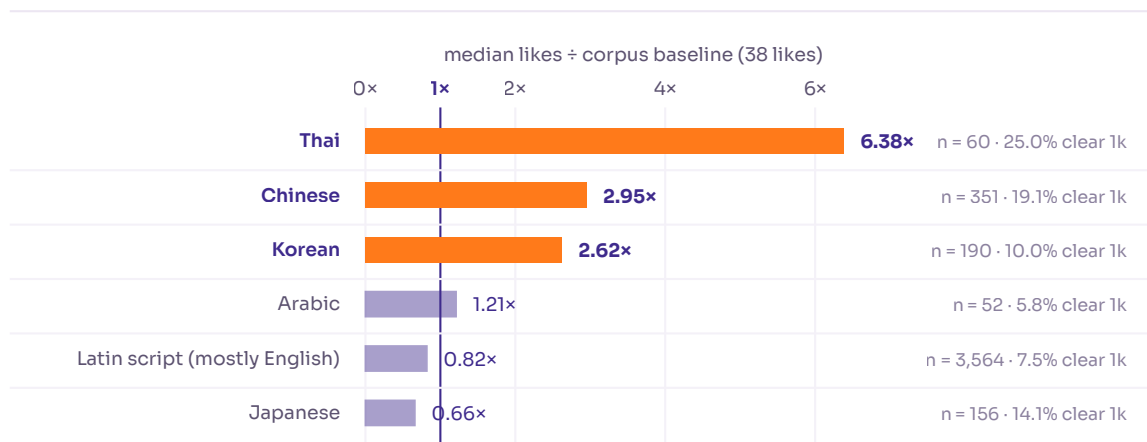
Two competing explanations for the ChatGPT result are worth separating, because the corpus can distinguish them only partially. The first is saturation: so much has been said about ChatGPT that a marginal post adds nothing. The second is composition: the name may be used disproportionately by newer accounts writing general-interest content. The within-author test of Chapter 7 speaks to this indirectly — mentioning AI vocabulary at all showed

no within-author effect (1.12, interval 0.85 to 2.44) — which suggests the vocabulary itself is not the mechanism and that the specificity of the claim being made matters more than the brand named in it.

8.3 Language markets

Language is the largest single association in this study, and also the one whose interpretation requires the most care. It cannot be tested within an author. It is therefore tested within collection pass, which is a weaker control but a real one. [Figure 8.3](#) and [Table 8.3](#) report the segments.

MEDIAN LIKES BY LANGUAGE SEGMENT, WHOLE CORPUS



The three highest segments are not an artifact of how the corpus was collected: Chinese out-earns the rest inside each of the four collection passes it appears in. Nor is it a format effect — Chinese posts are more text-heavy than the corpus average, so they win while using the format that loses everywhere else.

Figure 8.3 — Median likes by language segment against the 38-like corpus baseline. Segments are assigned by Unicode script range.

Table 8.3 — Language segments, analysis set (n = 4,247, baseline 38 likes). The plain-text share is included because it bears on whether the segment differences are really format differences in disguise.

Segment	n	Median likes	Lift	Clear 1k	Plain text %
Thai	60	242	6.38×	25.0%	12%
Chinese	351	112	2.95×	19.1%	50%
Korean	190	100	2.62×	10.0%	53%
Arabic	52	46	1.21×	5.8%	19%
Latin script (mostly English)	3,564	31	0.82×	7.5%	41%
Japanese	156	25	0.66×	14.1%	61%

The headline that the study's first analysis pass produced — non-English posts run at 2.26 times English — is an average that conceals four counter-examples and should not be used. Japanese sits below the Latin-script segment on the median while having twice its rate of thousand-like posts, which is a different distribution rather than a better one. Arabic is barely above baseline. The effect belongs to three named markets, not to the property of not being English.

The Chinese result is the one that survives the most scrutiny. [Table 8.4](#) recomputes it separately inside each collection pass.

Table 8.4 — Chinese-language posts against all other posts, computed separately within each collection pass. A collection artefact would not replicate across four independently constructed passes; a genuine audience difference would.

Pass	Chinese posts	Median	Other posts	Median	Ratio
ai	134	71	2,670	29	2.45×
a	78	82	458	18	4.46×
b	86	116	228	106	1.09×
c	52	228	491	116	1.97×

The ratio exceeds 1 in all four passes. The segment is 95 per cent Traditional script, which locates it in Taiwan and Hong Kong rather than the mainland. And it is not a format artefact: Chinese posts are more text-heavy than the corpus average, 50 per cent against 41 per cent, so they out-perform while using the format that under-performs everywhere else.

What remains uncontrolled is topic. The Chinese posts entered this corpus through different searches than the English ones, so some part of the gap may be subject matter rather than audience. The verification agent assigned to this dimension made exactly that objection and could not resolve it either; within the Chinese segment, AI-related posts run 1.76 times non-AI ones on an interval spanning 1.00, which is too thin to settle the question. The defensible claim is therefore: **Traditional-Chinese-language Threads shows systematically higher engagement per post than the Latin-script segment, replicated across four independent collection passes and not explained by format, with topic remaining a live alternative explanation.**

The best writing tactic this study can verify is worth about twice. The choice of language market is associated with two to six times. Audience selection dominates craft at this stage of the platform's life.

9. The Creator Landscape

Chapters 6 to 8 treated posts as the unit of analysis. This chapter changes the unit to the account, and asks three questions an operator entering the niche would want answered: who holds the attention, whether follower count predicts reach, and whether the high performers are repeatable or lucky. The answers are, in order: nobody, barely, and the corpus cannot tell — a set of findings that is unusually favourable to new entrants and unusually hostile to the archetype-building that this literature tends to produce.

9.1 A transient field

The analysis set contains 4,247 posts from 3,397 distinct accounts. Of those accounts, **3,014 — 88.7 per cent — appear exactly once**, and their single posts constitute 71.0 per cent of the set. Only 93 accounts contribute four or more posts.

This shape has a mundane component and a substantive one. The mundane component is that search surfaces a wide, shallow slice: a query returns twenty posts from twenty different accounts, so a corpus assembled from many queries will naturally be author-diffuse. The substantive component survives that caveat, because it concerns which accounts the surfacing favours. If a small set of established accounts dominated Threads' distribution in this niche, repeated searching would keep returning them. It does not.

Table 9.1 makes the point at the lane level.

Table 9.1 — Concentration within subject lanes, AI collection pass. The busiest-author share is the proportion of a lane's posts written by its most prolific account. A lane owned by an incumbent would show a large value here; none does.

Lane	Posts	Distinct authors	Posts per author	Busiest author share
Development and programming	393	319	1.23	1.5%
Design and creative	268	247	1.08	2.2%
Agents and automation	319	260	1.23	2.5%
Study and learning	272	230	1.18	2.6%
Prompting craft	398	300	1.33	2.8%
Marketing and growth	180	156	1.15	2.8%
AI coding and vibe coding	161	129	1.25	3.1%
Make money with AI	185	158	1.17	3.2%
Model and product news	164	134	1.22	3.7%
AI image and video	116	94	1.23	4.3%
Hardware and devices	52	48	1.08	5.8%
Career and jobs	147	130	1.19	6.1%

The largest share any account holds of any lane is 6.1 per cent, and in the two largest lanes it is under three per cent. Whatever else is difficult about entering this niche, displacing an incumbent is not part of it.

9.2 Followers predict less than expected

Follower counts were captured for 59 accounts, of which 37 have four or more analysis-set posts. This is a small and non-random sample — the accounts were visited because their posts had surfaced — and every statement in this section inherits that limitation. Within it, the rank correlation between follower count and an account's median likes is 0.219.

Table 9.2 shows the pattern that produces so weak a correlation.

Table 9.2 — Selected accounts with follower data, ordered by followers. Reach per follower is the account's median likes divided by its follower count, expressed as a percentage. The column is included to show its own instability rather than to propose it as a metric.

Account	Followers	Posts in set	Median likes	Best post	Reach per follower
@yourchatgptguide	1,582,898	10	226	1,194	0.01%
@evolving.ai	528,991	9	773	3,032	0.15%
@chatgptmastery	450,594	9	477	4,208	0.11%
@raminnasibov	247,228	7	1,660	6,315	0.67%
@unlock.ai.tools	18,948	9	637	2,139	3.36%
@artificial.corner	5,823	51	155	3,026	2.66%
@ihtesham.ai	1,505	9	634	3,640	42.13%
@iman.aimuni	119	4	702	2,783	589.08%
@rheaabano	20	4	4,655	18,617	23,275%

The account with 1.58 million followers has a lower median than an account with 1,505, and the largest single post among these creators came from an account with twenty followers. Read naively this suggests small accounts out-perform large ones, and the study's first analysis pass said so.

Methodological note

Why that reading is wrong, and what survives it. These small accounts are present in the corpus because a post of theirs performed well enough to be surfaced by search. Accounts of the same size whose posts sank were never collected, so the small-account rows are a sample of winners while the large-account rows are closer to a sample of typical output. The comparison is therefore invalid as stated, and the "small accounts out-reach big ones" finding is withdrawn. What survives is a weaker and still useful claim: **an account with twenty followers produced a post with 18,617 likes**, which establishes that the distribution mechanism does not gate reach on accumulated audience. Follower count sets a floor, not a ceiling.

9.3 Repeatability cannot be established here

An obvious next question is whether the high performers are repeatable — whether some accounts have found a system — and it is the question this corpus is least able to answer. The study's first pass attempted it, classifying accounts into archetypes and reporting that "metronomes" with a high ratio of median to best post carried 8.1 times baseline while "lottery winners" carried 0.74 times. Verification destroyed the finding on two independent grounds, and both are instructive.

The first is circularity. The metronome class was defined partly by requiring a lift above 2.0, so reporting that metronomes have high lift restates the filter rather than discovering anything about consistency.

The second is a sampling artefact. The consistency ratio — an account's median divided by its best post — is mechanically driven by how many of that account's posts the scrape happened to surface. The rank correlation between an account's post count in the corpus and its consistency ratio is -0.355 : the median ratio is 0.279 at four posts and 0.059 at eight or more. An account that appears twice will look consistent almost by construction, and an account that appears fifty times will not.

The honest position is that **this corpus cannot distinguish a repeatable performer from a lucky one**, because it does not observe any account's full output. Establishing repeatability requires account-complete panel data — every post by a fixed set of accounts over a fixed window — which is a different collection design and is proposed in Section 12.3 as the most valuable follow-up study.

9.4 The top decile at the account level

Section 5.2 established that the top decile takes 79 per cent of the likes. Resolving that decile to accounts gives the chapter's closing fact. The 424 top-decile posts come from 391 distinct accounts; roughly three quarters of those accounts appear exactly once in the analysis set; and 70.4 per cent of top-decile posts come from unverified accounts.

The composition of the decile differs from the corpus in exactly one structural respect worth reporting: 37.0 per cent of top-decile posts are carousels against 23.5 per cent of the corpus, a ratio of 1.57. Every other structural feature tested — numerals, emoji, hashtags, question marks, length, line count — separates the decile from the rest by less than the sampling noise, or not at all. That single surviving separator is the same one that survived the within-author test in Chapter 7, which is the strongest internal consistency check this study produces.

The tail is not a club. Three hundred and ninety-one accounts produced the top four hundred and twenty-four posts, and three quarters of them appear once. What separates those posts from the rest is not who wrote them; it is, measurably, whether there was something to look at.

10. Adversarial Verification

Every finding in this report passed through a protocol designed to destroy it. This chapter documents that protocol, reports what it destroyed, and argues that the destruction is the most transferable contribution the study makes. Seven of nine headline findings did not survive. They are reported here in full, with the objection that felled each, because a study that reports only its survivors gives the reader no way to calibrate how hard it tried.

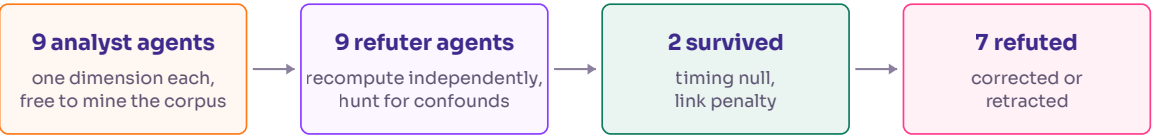
10.1 The protocol

The analysis was run twice by independent language-model agents operating on the same corpus but with opposed incentives.

In the first pass, nine analyst agents were each assigned one dimension — subject demand, tool mentions, the creator landscape, engagement composition, mined hook phrases, top-decile anatomy, language, timing, and negative patterns. Each was given the corpus path, the schema, the maturity rule, a minimum sample size of twenty per reported bucket, and an instruction to report medians with sample sizes and to name confounds. Each returned a structured object containing a headline claim, supporting findings with sample sizes, tables and caveats.

In the second pass, nine refuter agents were each given one analyst's headline claim and supporting findings and instructed to refute it: to recompute the underlying numbers with their own code rather than reusing the analyst's approach, to check whether the maturity filter had been applied, to test whether the effect was explained by a confound — creator identity, media type, subject, language, account size or collection provenance — and to default to refuted where evidence was thin. Each returned a boolean verdict, an assess' ment, the confound it found, and a corrected version of the claim.

NINE ANALYSES, NINE INDEPENDENT ATTEMPTS TO REFUTE THEM



WHAT THE REFUTERS FOUND

- **Creator identity** — the same tactic inside one author shows no effect
- **Collection provenance** — groups drawn from different passes are not comparable
- **Circular definition** — an archetype defined by high performance reports high performance
- **Composition** — link posts are mostly text posts, and text is the weak format

Figure 10.1 — The verification protocol and its outcome. Four confound families account for all seven refutations.

The protocol consumed 1.44 million tokens across eighteen agents and produced no errors. Its cost was roughly three times that of the original analysis, which is the correct order of magnitude for the value returned.

10.2 Outcome

Table 10.1 lists all nine verdicts.

Table 10.1 — Verification outcomes by dimension. "Corrected claim" is the defensible version the refuter was able to support, not the analyst's original wording.

Dimension	Verdict	Confound identified	Corrected position
Subject demand	re-futed	Sampling frame	Engagement lift survives; the "white space" inference does not, because supply is the query list
Tool mentions	re-futed	Collection provenance	No general penalty for naming a tool; the 40% gap was a mix of scrape passes
Creator landscape	re-futed	Circular definition	Consistency archetypes are an artefact of sample depth and cannot be established here
Engagement composition	re-futed	Topic, held format constant	Format skew is real and robust; it is not the only driver, and topic moves it 3.7-fold
Mined hook phrases	re-futed	Creator identity	The AI-vocabulary penalty is entirely between accounts; within authors it vanishes
Top-decile anatomy	re-futed	Status proxied by a paid badge	Carousel separation survives every control; the account-size reading does not
Language	re-futed	Segment aggregation	The effect belongs to Chinese, Thai and Korean, not to "non-English"
Timing	survived	Creator identity, post age	Timing is a non-lever, and controlling for confounds strengthens the null
Negative patterns	survived	Composition, format	Link penalty is real; magnitude falls from 82% to roughly 58% after stratification

10.3 The four confound families

The seven refutations reduce to four recurring mechanisms, and naming them is more useful than cataloguing the individual errors.

Creator identity. The dominant mechanism, and the one that motivated Chapter 7. A tactic correlated with author quality inherits that quality in any between-account comparison. The refuter working on mined hook phrases established this most sharply: the raw gap between first lines that name AI and those that do not replicates strongly (median 22 against 38, Mann–Whitney $z = -6.33$), but restricting to accounts with two or more posts in the set erases it (54 against 57), and the entire raw penalty sits inside single-post accounts.

Collection provenance. Described in Section 4.2 and discovered by the refuter assigned to tool mentions, who reconstructed which part-file first contained each post ID and found five passes with medians from 20 to 332 likes. Because brand-mention rates tracked the pass almost perfectly, the apparent brand penalty was a pass-composition effect. This is the confound most specific to scraped corpora and the one least discussed in the literature the study consulted.

Circular definition. The creator-archetype finding defined its favourable class partly by high performance and then reported that the class performed highly. This error is easy to make when categories are constructed during analysis rather than before it.

Composition. Groups that differ on the variable of interest often differ on others. Link-car'rying posts are 74 per cent plain text against 41.6 per cent of the corpus, and their median age is 20 days against 51, so the raw 0.18 ratio blends a link effect with a format effect and an age effect. Stratifying by media type and age moves it to 0.42.

10.4 What the refuters got wrong

Symmetry requires reporting that the refuters were not uniformly right. Two of their objections were themselves overstated, and the within-author battery of Chapter 7 — run after the protocol, precisely to adjudicate — settled both.

The negative-patterns analyst reported hashtags as "the most under-diagnosed antipattern", at 0.56 after format and age adjustment. The within-author test puts hashtags at 1.12 with an interval from 0.79 to 1.68: no detectable effect. The analyst over-claimed and the refuter did not catch it.

Conversely, the anatomy refuter argued that the carousel separation might be an account-size artefact, noting correctly that the study had proxied account size with a verification badge that on this platform indicates a paid subscription rather than audience. The objection was sound in principle and wrong in fact: within the sixty-eight accounts that posted both carousels and other formats, the carousel advantage holds at 1.73 with an interval excluding 1.00.

The lesson is not that adversarial verification is unreliable but that it is a filter rather than an oracle. It reliably catches claims that rest on a confound. It does not reliably catch claims that are merely wrong, and it can generate objections that are plausible and false. The arbiter in both cases here was a stricter estimator, not a more forceful argument.

10.5 Errors found in the study's own working

Three errors were caught during the study and are recorded because they are more instructive than the findings they contaminated.

The first was a scanning defect in the collection harness. Threads sometimes serves navigation as an in-application transition, and the harvester's guard against double-scanning a page suppressed the scan entirely on those transitions. Roughly nineteen posts per affected query were silently discarded, and the loss was invisible because the harness reported zero new posts, which is also what a genuinely exhausted query reports. It was caught by noticing that a page displaying twenty posts had yielded none.

The second was a storage-quota failure. At approximately 9,400 posts the browser store exceeded its five-megabyte origin limit, writes began failing silently, and the running count started to decrease as stale snapshots overwrote newer ones. The fix — exporting to disk and replacing the in-browser store with a compact index of post identifiers — is described in Appendix A. The general lesson is that a monotone counter that goes down is the cheapest possible integrity alarm, and that collection harnesses should assert monotonicity explicitly.

The third was the "white space" family of claims described in Section 8.1: a whole analytical frame built on a supply denominator that measured the analyst's own behaviour. It was caught by a refuter, not by the analyst, and it is the clearest instance in this study of an error that no amount of internal consistency checking would have surfaced.

A study that reports only its survivors gives the reader no way to judge how hard it tried to kill them. Seven of nine headline findings did not survive here; the two that did are worth more for the company they no longer keep.

11. Discussion

This chapter draws out what the results mean, first for anyone analysing engagement data and then for anyone actually posting. It ends with a full statement of the threats to validity, ordered by how much damage each would do if it turned out to be operating.

11.1 Implications for analysis

The default comparison in this field is the wrong one. Pooling posts across authors and comparing feature groups is the standard method, it is what most published virality analysis does, and on this corpus it produced twenty findings out of twenty-three that a within-author test cannot reproduce. The problem is not sample size — a larger pooled corpus estimates the confounded quantity more precisely — and it is not statistical technique. It is identification. Any dataset in which the treatment is chosen by a unit that also has a persistent quality level will exhibit this, and social platforms are the paradigm case.

Provenance is a first-class variable in scraped data. Section 4.2 documents an elevenfold spread in median engagement across collection passes of the same corpus. Any study that assembles data over multiple sessions with varying query sets — which is most scraped social-media research — should record the pass with each record and check its findings within pass. This is cheap to do and, in this study, it was the difference between reporting and retracting a headline finding.

Supply cannot be measured from a query-driven corpus. The share of a corpus occupied by a topic is the share of the analyst's query list occupied by that topic. Any metric that divides engagement by category volume — opportunity scores, saturation indices, white-space maps — is measuring the analyst. This study produced a full set of such claims before catching the error, and reports demand alone.

Null results deserve their own reporting standard. Twenty of the twenty-three tactics returned intervals spanning no effect. Reporting these as "does not work" would be as much an over-claim as the original advice. They are reported as unproven, with the interval shown, because several are consistent with a real effect that this design lacks the power to detect — single images at 1.32 with an interval from 0.71 to 1.78 being the clearest case.

11.2 Implications for practitioners

The practical content of this study is unusually small, which is itself the finding. Three tactics survived, and they can be stated in three sentences.

Table 11.1 — The verified practitioner set. Effect is the within-author ratio from Table 7.1. Everything not in this table is either unproven at this sample size or, in the case of subject and language, an association measured under a weaker design.

Action	Effect	95% CI	Au- thors	Confidence
Attach media rather than posting plain text	1.98×	1.33– 4.78	95	High — clears both criteria
Prefer a carousel where the material allows	1.73×	1.18– 3.18	68	Moderate — interval clears, sign test does not
Keep outbound links out of the post body	0.30×	0.13– 0.71	41	High — largest verified effect in the study

Beyond those three, the study supports four weaker recommendations, each flagged with the design that produced it.

Choose the format for the job, not for the score. Carousels return 10.2 reposts per hundred likes against 2.2 replies; plain text inverts it at 5.6 against 7.8 (Section 6.2). Reposts introduce strangers, replies deepen relationships with an audience already present. An account that posts only carousels accumulates an audience it never speaks to; one that posts only text speaks to an audience that stops growing. This is a population-level result with a ten-author within-author check pointing the same way — direction supported, magnitude unsettled.

Prefer subjects that can carry visible evidence. Within a single collection pass, hardware (3.68×), AI coding (2.20×) and AI image and video (2.15×) lead; prompting (0.97×), agents (0.87×) and make-money-with-AI (0.47×) trail. The ordering is consistent with the one verified post-level result — attach something to look at — and a lane is partly a constraint on what can be attached. This is not controlled for creator identity and should be held loosely.

Language market is the largest association in the study. Thai, Chinese and Korean segments run at 2.6 to 6.4 times the corpus baseline, and the Chinese result replicates inside all four collection passes it appears in and is not explained by format. Topic remains a live alternative explanation (Section 8.3). If it holds, audience selection is worth several times what any writing tactic is worth — which is an argument for serving an under-covered market properly rather than for another round of hook optimisation.

Measure against yourself. Creator identity explained 46.7 per cent of the variance in log likes; hour of posting explained 2.7 per cent. Comparing a post to someone else's is close to uninformative about your own decision. The protocol that produced this study's results is available to any operator: vary one thing deliberately for a fortnight, then compare the median of your posts with it against the median of your posts without it. At twenty posts, differences under roughly 1.5-fold are indistinguishable from noise.

Methodological note

What an operator should stop doing. The clearest practical consequence is negative. Effort currently spent on hook formulae, bullet formatting, emoji policy, hashtag strategy, character counts and posting times is, on this evidence, effort spent on variables that do not move the outcome detectably within an account. That effort has an opportunity cost: it is not being spent on producing something worth attaching, which is the one thing that did survive.

11.3 Why the illusion is so persistent

It is worth asking why advice of this kind survives despite being untested, and the data suggest three mutually reinforcing mechanisms.

The first is that the correlations are real. Bulleted posts do earn more likes than unbulleted ones across the corpus — 1.53 times as many. Anyone who checks the aggregate will confirm the advice, and the check feels rigorous.

The second is the direction of teaching. Advice flows from accounts that have succeeded, and those accounts are describing their own behaviour accurately. They do write short first lines and structure their posts. The error is not in their observation; it is in the inference that these habits transfer, when they are markers of a practice rather than its cause.

The third is that the platform's tail structure supplies constant confirmation. With 8.7 per cent of surfaced posts clearing a thousand likes, any tactic adopted enthusiastically will produce a breakout within a few weeks by chance alone, and that post becomes the evidence.

11.4 Threats to validity

Ordered by the damage each would do if it is operating.

Selection into treatment within author (severe if operating). The within-author estimator removes stable author characteristics but not the author's own choice of when to use a tactic. If creators attach media precisely when they have unusually good material, the media results credit the container for the contents. This cannot be tested with observational data and is the single most important caveat on the study's three surviving findings. It would take a field experiment — the same creator publishing matched material in both forms — to settle, and Section 12.3 proposes exactly that.

Search-surfacing and survivorship (severe for rates, mild for comparisons). No failed posts are observable. Every absolute rate in Chapter 5 is inflated relative to all posting, and every account-level statistic in Chapter 9 is drawn from accounts that surfaced at least once. This is why the small-account finding was withdrawn in Section 9.2.

Low power in the within-author battery (moderate). Between 21 and 177 authors per tactic. Effects of 10 to 20 per cent are invisible. The twenty nulls should be read as failures of evidence, not demonstrations of absence.

Topic-language entanglement (moderate). The language segments were reached through different queries than the English posts, so subject matter is not separated from audience in Section 8.3. Within-pass replication makes a pure collection artefact unlikely but does not isolate the mechanism.

Unobserved outcomes (moderate). No impressions, no saves, no direct-message shares. Instructional content, which the AI lane produces heavily, is plausibly under-measured by a like-based outcome, and the study's format conclusions would shift if saves were observable.

Single platform, single window (moderate). One platform, one topic neighbourhood, one collection window in August 2026, on a ranking system that is revised frequently. The method should generalise; the specific numbers should be assumed to decay.

Automated analysis (mitigated but present). The analysis was conducted largely by language-model agents. The verification protocol in Chapter 10 caught seven headline errors, and two further errors that the refuters missed were caught by the within-author battery. It would be unreasonable to assume none remain.

The most useful output of this study is a smaller and more honest set of instructions than the one it set out to test — three verified actions, four qualified ones, and a method for checking the rest against your own account.

12. Conclusion

This study set out to test whether the advice that circulates about growing an audience on Threads survives contact with data, and found that most of it does not — not because the underlying correlations are false, but because they describe the people using the tactics rather than the tactics themselves.

12.1 Answers to the research questions

RQ1 — How is engagement distributed, and what is the right baseline? It is heavy-tailed. Among the 4,247 mature posts, the median earns 38 likes while the mean earns 401; the top decile takes 79.0 per cent of all likes and the top one per cent takes 34.3 per cent. A post should be judged against the percentile ladder — 223 likes is the 75th percentile, 846 the 90th — and never against an average that no post experiences.

RQ2 — Which tactics survive control for creator identity? Three of twenty-three. Attaching any media returns 1.98 times the same author's median (interval 1.33–4.78, 95 authors); posting a carousel returns 1.73 (1.18–3.18, 68 authors); including an outbound link returns 0.30 (0.13–0.71, 41 authors). The remaining twenty — including bulleted lists, numbered lists, short first lines, command openings, question marks, emoji, hashtags, ALL-CAPS, character counts, follow requests and AI vocabulary — return intervals spanning no effect. Several population-level effects reverse: bulleted lists move from 1.53 to 0.80, numbered lists from 2.05 to 0.88.

RQ3 — Does format change the composition of engagement? Yes, and by more than it changes the volume. On posts above 100 likes, carousels return 10.2 reposts per hundred likes against 2.2 replies, a skew of 4.63; plain text returns 5.6 against 7.8, a skew of 0.72. The 6.4-fold spread means the two formats are different instruments: one recruits strangers, the other deepens an existing audience.

RQ4 — Which subjects, tools and languages perform, and how much survives control for collection? Within a single collection pass, subject lanes span 7.8-fold from hardware (3.68×) to make-money-with-AI (0.47×), with lanes that can display visible evidence at the top. Among named tools, ChatGPT holds the largest share of voice (12.3%) and the lowest lift of the major names (0.73×), while ElevenLabs, Midjourney and NotebookLM each hold about one per cent of voice at 5.3 to 6.9 times. Language is the largest association measured: Thai at 6.38×, Chinese at 2.95× and Korean at 2.62× against Latin script at 0.82×, with the Chinese result replicating inside all four passes and not explained by format. Topic remains unseparated from language.

RQ5 — What does the creator landscape look like? Diffuse and transient. Of 3,397 accounts in the analysis set, 88.7 per cent appear exactly once and contribute 71.0 per cent of the posts; no account holds more than 6.1 per cent of any subject lane; the 424 top-decile

posts come from 391 distinct accounts, roughly three quarters of which appear only once. Follower count correlates with median likes at 0.219 among the 37 accounts where it is observed, and an account with twenty followers produced a post with 18,617 likes. Follower count appears to set a floor rather than a ceiling — though the account-level sample is small and selected on performance, so this is the study's weakest claim rather than its strongest.

12.2 The general result

Stated without reference to Threads: **in observational engagement data, the difference between a between-author and a within-author comparison is not a methodological refinement but, for most content tactics, the entire effect.** Twenty of twenty-three claims that any conventional analysis would report as findings did not survive making each creator their own control group. Seven of nine headline findings from this study's own first analytical pass did not survive an adversarial re-derivation.

Both numbers are uncomfortably high, and both were produced by applying controls that are available to anyone with the same data. The implication for this literature is not that its published findings are wrong — several used stronger designs than the practitioner advice examined here — but that the informal analysis which dominates public discussion of social platforms is producing conclusions at a rate its evidence does not support.

12.3 Future work

Three studies would materially advance what is known here, in order of value.

A field experiment on attachment. The three surviving findings all concern what is attached to a post, and all three are vulnerable to the same objection: creators may attach media exactly when they have good material. A cooperating set of creators publishing matched content in both forms, with assignment randomised, would settle in weeks what no amount of observational data can.

Account-complete panel collection. The corpus cannot distinguish a repeatable performer from a lucky one, because it never observes any account's full output (Section 9.3). Collecting every post from a fixed panel of a few hundred accounts over six months would support the questions this study had to abandon: whether consistency exists, what a creator's engagement autocorrelation looks like, and whether a breakout post has any durable effect on the posts that follow it.

A cross-platform replication of the within-author correction. If twenty of twenty-three tactics collapse on Threads, the same test on X, Instagram and LinkedIn would establish whether the correction factor is a property of this platform or of the genre of advice. The design transfers unchanged.

12.4 Closing

The study began as an attempt to find out which tactics work, and ended by mostly establishing which questions the available evidence cannot answer. That is a less satisfying result than a list of techniques, and a more useful one: the three surviving actions are worth more than the twenty discarded, precisely because they were made to survive something.

The method matters more than any of the numbers. Hold the author fixed. Record where the data came from. Report what died. Anyone with a spreadsheet of their own posts can run the same test on the advice they are being given, and on this evidence they should.

References

Thirty-five works, each checked to resolve to the cited publication. In-text citations give author and year in square brackets.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. <https://press.princeton.edu/books/paperback/9780691120355/mostly-harmless-econometrics>

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132. <https://doi.org/10.1126/science.aaa1160>

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2), 192–205. <https://doi.org/10.1509/jmr.10.0353>

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1–8. <https://doi.org/10.1016/j.jebo.2011.08.009>

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)* (pp. 925–936). ACM. <https://doi.org/10.1145/2566486.2567997>

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189–228. <https://doi.org/10.1214/ss/1032280214>

Dixon, W. J., & Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41(236), 557–566. <https://doi.org/10.1080/01621459.1946.10501898>

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. arXiv:2305.14325. <https://arxiv.org/abs/2305.14325>

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26. <https://doi.org/10.1214/aos/1176344552>

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460–465. <https://doi.org/10.1511/2014.111.460>

Hu, X., Zhao, Z., Wei, S., Chai, Z., Ma, Q., Wang, G., Wang, X., Su, J., Xu, J., Zhu, M., Cheng, Y., Yuan, J., Li, J., Kuang, K., Yang, Y., Yang, H., & Wu, F. (2024). InfiAgent-DABench: Evaluating agents on data analysis tasks. arXiv:2401.05507. <https://arxiv.org/abs/2401.05507>

- Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. arXiv:2310.01798. <https://arxiv.org/abs/2310.01798>
- Hughes, C., Swaminathan, V., & Brooks, G. (2019). Driving brand engagement through online social influencers: An empirical investigation of sponsored blogging campaigns. *Journal of Marketing*, 83(5), 78-96. <https://doi.org/10.1177/0022242919854374>
- Jeong, U., Beigi, A., Tahir, A., Tang, S. X., Bernard, H. R., & Liu, H. (2025). Fediverse sharing: Cross-platform interaction dynamics between Threads and Mastodon users. In *Proceedings of the 2025 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '25)*. arXiv:2502.17926. <https://arxiv.org/abs/2502.17926>
- Jeong, U., Nirmal, A., Jha, K., Tang, S. X., Bernard, H. R., & Liu, H. (2024). User migration across multiple social media platforms. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM24)*. arXiv:2309.12613. <https://arxiv.org/abs/2309.12613>
- Kievit, R. A., Frankenhuys, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology*, 4, 513. <https://doi.org/10.3389/fpsyg.2013.00513>
- Leung, F. F., Gu, F. F., & Palmatier, R. W. (2022). Online influencer marketing. *Journal of the Academy of Marketing Science*, 50(2), 226-251. <https://doi.org/10.1007/s11747-021-00829-4>
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50-60. <https://doi.org/10.1214/aoms/1177730491>
- McShane, L., Pancer, E., Poole, M., & Deng, Q. (2021). Emoji, playfulness, and brand engagement on Twitter. *Journal of Interactive Marketing*, 53, 96-110. <https://doi.org/10.1016/j.intmar.2020.06.002>
- Meta. (2023, July 5). Introducing Threads: A new way to share with text. Meta Newsroom. <https://about.fb.com/news/2023/07/introducing-threads-new-app-text-sharing/>
- Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends in Human-Computer Interaction*, 14(4), 272-344. <https://doi.org/10.1561/11000000083>
- Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*. arXiv:1306.5204. <https://arxiv.org/abs/1306.5204>
- Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1), 69-85. <https://doi.org/10.2307/1913646>

- Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 13. <https://doi.org/10.3389/fdata.2019.00013>
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. <https://doi.org/10.2307/2087176>
- Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213), 1063–1064. <https://doi.org/10.1126/science.346.6213.1063>
- Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on Internet platforms. Paper presented at the Data and Discrimination preconference, 64th Annual Meeting of the International Communication Association, Seattle, WA. <https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3), 337–356. <https://doi.org/10.1177/2515245917747646>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2), 238–241. <https://doi.org/10.1111/j.2517-6161.1951.tb00088.x>
- Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing (SocialCom)* (pp. 177–184). IEEE. <https://doi.org/10.1109/SocialCom.2010.33>
- Tan, C., Lee, L., & Pang, B. (2014). The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 175–185). Association for Computational Linguistics. <https://doi.org/10.3115/v1/P14-1017>
- Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*. arXiv:1403.7400. <https://arxiv.org/abs/1403.7400>
- Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25(2), 101–132. <https://doi.org/10.3102/10769986025002101>

Appendix A. Collection Harness

This appendix documents the collection mechanism in enough detail to reimplement, together with the three defects found in it during the study. The defects are included because each produced a characteristic failure that another researcher building a similar harness is likely to meet.

A.1 Mechanism

Threads renders each page with a JSON payload embedded in the document, containing the post objects that page will display, and fetches further payloads over an internal GraphQL endpoint as the reader interacts. The harness reads both. A small script is injected into each loaded page which walks every embedded `application/json` block and every intercepted network response recursively, to a depth limit of sixty, collecting each object that carries a post identifier alongside a like count. For each such object it records the identifier, the author handle and verification flag, the four engagement counters, the media type derived from attachment structure, the link-preview flag, the publication timestamp, the post code and the text.

Records accumulate in browser storage keyed by post identifier, so re-encountering a post is idempotent. Batches are exported to disk as CSV by constructing a form and submitting it to a local receiver process — the route chosen after the browser's file-download path was found to block repeated automated saves from the same origin.

A.2 Three defects

Suppressed scans on in-application navigation. The harness guarded against scanning a page twice by setting a flag after the first scan. Threads serves some navigations as in-application transitions that preserve the JavaScript context, so the flag persisted across what was logically a new page and suppressed the scan entirely. The failure was silent: the harness reported zero new posts, which is also what a genuinely exhausted query reports. It was caught by observing a page that displayed twenty posts and yielded none, and fixed by resetting the flag before every scan. Approximately nineteen posts per affected query were lost before the fix.

Storage-quota exhaustion. At roughly 9,400 posts the accumulated store exceeded the five-megabyte origin limit for browser local storage. Writes began failing, the failures were caught and counted rather than raised, and the running total started to decrease as pages loaded stale snapshots and wrote them back. The remedy was to export the full store to disk and replace it with a compact index: a comma-separated string of the 9,394 known post identifiers occupying 183 KB, against which the walker tests membership before storing anything. Subsequent passes therefore stored only genuinely new records.

Renderer stalls from over-eager persistence. An early version serialised the entire store to local storage on every intercepted network response. At a few thousand posts this meant repeatedly serialising megabytes inside a network callback, which froze the page. Fixed by throttling persistence to at most one write every three seconds, with a forced flush after each page scan.

Methodological note

The transferable lesson. All three defects were silent. None raised an error; each degraded collection in a way that looked like an ordinary empty result. A collection harness should assert its own invariants — that a page displaying posts yields posts, and that a cumulative counter never decreases — and stop when they fail, because the alternative is discovering a month later that part of the corpus is missing for reasons that cannot be reconstructed.

Appendix B. Query Design

Approximately 700 distinct queries were issued across seven passes. The query list was not fixed in advance; it was extended during collection as yields fell, which is a source of the provenance confound documented in Section 4.2 and a reason no supply claim is made anywhere in this report.

Table B.1 — Query families, in the order they were issued. Yield is the approximate number of previously unseen posts returned per query at the time that family was run; it declines as the corpus saturates, not because the later families are less productive in absolute terms.

Pass	Family	Examples	Approx. queries	Yield/ query
ai	Core AI and tooling	claude code, prompt engineering, ai agents, chatgpt prompts	150	15–19
ai	Creator and marketing	threads algorithm, grow on threads, ai marketing	40	13–17
a	General technology	typescript, devops, cybersecurity, ui design	90	10–17
b	Non-English AI terms	生成AI, 인공지능, inteligência artificial, yapay zeka	70	15–20
c	Non-English general	仕事, 일상, trabalho, حياة, نورا	120	13–20
c	Tag search	#ai, #aitools, #webdev, #promptengineering	25	8–12
d	Creator profiles	direct profile visits for 59 accounts	59	2–8
e, f	Recency-filtered feeds	the same core terms under the "recent" ordering	150	8–15

Two structural properties of Threads search shaped the design. It returns approximately twenty results per query and does not paginate for an automated session, so scale had to come from query breadth rather than depth. And its two orderings, "top" and "recent", return substantially different sets, so high-value terms were issued under both.

Appendix C. Feature Definitions

Every tactic tested in Chapter 7 is a deterministic function of the stored record. The definitions are given here so that a reader can reproduce or dispute them.

Table C.1 — Operational definitions of the twenty-three tested tactics. "First line" means the text preceding the first newline character. Patterns are applied case-insensitively unless stated.

Tactic	Definition
Attach any media	media type is not <code>text</code>
Post a carousel	media type is <code>carousel</code> (two or more images)
Post a single image	media type is <code>image</code>
Post a video	media type is <code>video</code>
Outbound link in the post	the platform attached a link preview to the post
End the first line on a question	first line, stripped, ends with "?"
Question mark anywhere	"?" occurs in the post text
Add hashtags	"#" occurs in the post text
Use emoji	any character in U+1F300–U+1FAFF or U+2600–U+27BF
Numbered list in the body	a line begins with a digit followed by "." or ")"
Bulleted list in the body	a line begins with <code>*</code> , <code>-</code> or <code>*</code>
ALL-CAPS emphasis word	a run of four or more consecutive capital letters
End the first line on a colon	first line, stripped, ends with ":"
Number in the first line	any digit occurs in the first line
First line under 40 characters	length of the first line < 40
Post under 120 characters	total text length < 120
500 or more characters	total text length ≥ 500
Six or more lines	newline count + 1 ≥ 6
AI vocabulary	word-boundary match on ai, chatgpt, gpt, claude, gemini, llm, prompt, openai
Ask for a follow	"follow" immediately followed by me, for or an @handle
Ask people to save it	"save this" or "bookmark"
Open with "I" or "My"	first line begins with "i ", "i" or "my "
Open with "You" / "Your"	first line begins with you, your or "if you"
Open with a command verb	first line begins with stop, start, save, steal, copy, try, use, read, bookmark, delete or forget

Two definitional choices are worth flagging as contestable. Media type is assigned by precedence, so a post carrying both a video and text is `video`; a scheme that treated them as separate dimensions would produce different format tables. And the link flag records a platform-generated link preview rather than the presence of a URL string in the text, which is the stricter of the two available definitions and, in a comparable study on another platform, the one that reversed the sign of the effect.

Appendix D. Reproduction

The released files are sufficient to reproduce every number in Chapters 5 to 9 without re-collecting anything.

Table D.1 — Released artefacts.

File	Contents	Rows
<code>threads_dataset_12k.csv</code>	Every post: identifiers, all four engagement counters, media type, link flag, timestamp, URL, character and line counts, full text	12,045
<code>threads_most_viral_top750.csv</code>	The 750 highest-liked posts, same schema	750
<code>threads_creators.csv</code>	Profile data for the 59 accounts whose pages were visited: handle, followers, verification, name, biography	59
<code>within_author_tests.csv</code>	The twenty-three tactic tests: population lift, paired author count, improvement rate, within-author ratio, interval, sign-test p, verdict	23
<code>market_map_tables.csv</code>	Lane, tool, language and percentile tables from Chapters 5 and 8	41

The within-author estimator is defined completely in Section 3.4 and is roughly twenty lines of code: group the analysis set by author, split each author's posts on the tactic, take the median of each side, form the ratio with the additive constant, and take the median of the ratios with a percentile bootstrap over authors. Collection-pass membership, needed to re-produce the stratified analyses of Chapter 8, is recoverable by checking which of the seven part files first contains each post identifier.

Two caveats on reproduction. Engagement counters are as at collection and will have risen; recomputing from a fresh scrape will produce different absolute values and, in all likelihood, similar ratios. And the corpus reflects Threads' ranking behaviour in August 2026, which is subject to change without notice.

Appendix E. Withdrawn Findings

Eight claims produced during the study are recorded here as withdrawn, with the reason. They are listed because a reader assessing the reliability of what remains is entitled to see what was discarded and why, and because several are claims that circulate widely and would otherwise have been reported here as confirmed.

Table E.1 — Claims produced and withdrawn during the study, with the control that removed them.

Claim as originally produced	Removed by	Status
Bulleted lists double reach (2.03× baseline)	Within-author test	Withdrawn — 0.80× (0.50–1.25) within author
Imperative openings lift 1.88×	Within-author test	Withdrawn — 0.63× (0.35–2.11), interval spans no effect
Never open on a question (0.33× baseline)	Within-author test	Softened — 0.78× (0.50–1.03), directional only
Short posts triple the median	Within-author test	Withdrawn — 0.98× (0.66–1.33)
Hashtags halve reach after adjustment	Within-author test	Withdrawn — 1.12× (0.79–1.68), no detectable effect
Saying "AI" costs 40–44% of reach	Creator identity, provenance	Retracted — effect sits entirely in single-post accounts
Small accounts out-reach large ones	Survivorship	Retracted — small accounts entered the corpus by having a hit
Hardware is the biggest "white space"	Sampling frame	Half retracted — engagement lift stands; scarcity claim does not
Consistent creators ("metronomes") carry 8.1× baseline	Circular definition, sample depth	Retracted — class defined partly by high performance
Post at 00:00 UTC — 3× spread across hours	Creator identity	Retracted — all windows within 0.474–0.523 of author percentile

Two of these — the bulleted-list and question-opening claims — were published in an earlier, population-level version of this research before the within-author battery was run. That version has been corrected in place with a notice directing readers here.



Colophon

Set in Sora. Figures drawn as vector SVG directly from the study's own aggregates, so no figure can drift from the text it illustrates. Typeset from HTML and CSS Paged Media with WeasyPrint; the table of contents, list of figures and list of tables carry resolved page references rather than manual ones.

Report 1.0 · 24 August 2026 · Ramanpal Singh · Promptslove · promptslove.com · admin@promptslove.com