

Who Wrote the Answer? Tracing Sanctioned Authors' Words from Wikipedia into LLM Output

A. de Brunner and L. Lolova

Anomaly AI Limited

DRAFT: UNDER REVIEW

Abstract

Search-augmented language models answer in Wikipedia's words: about a third of the answers we collect carry a run of eight or more consecutive words copied straight from an article. Wikipedia blocks the accounts it finds editing in bad faith, but a block typically removes the editor and not the edit, so what a sanctioned account wrote goes on answering questions long after the account is gone.

We trace that text end to end across 62 articles from a single contentious topic area, attributing all 683,276 surviving prose words to the accounts that added them and stratifying those accounts by strength of community evidence. Text written by flagged authors outlasts its share of the work: the 995 accounts on the three tier lists hold 38.8% of the surviving wikitext against 19.1% of bytes added across all revisions, a $2.03\times$ enrichment, and the effect is largest for the accounts Wikipedia itself sanctioned: $2.19\times$ for Tiers 2 and 3 alone.

We then measure propagation into 1,200 search-enabled answers from three models, under two question designs that between them cover both what we can trace and what users actually ask: one generated from corpus passages of known authorship, the other drawn from Google autocomplete. Matching runs of eight or more identical words, we compare each pool's share of quoted words with its share of article words. The two shares track: accounts blocked at least once hold 52.2% of the corpus prose and supply 53.6% and 58.5% of pasted words under the two designs, a lift of $1.08\times$ over all 1,200 answers.

Wikipedia's sanctions often do not travel with the text those accounts wrote. At verbatim granularity these models reproduce surviving article prose at the rate it occupies the encyclopaedia, whoever wrote it.

1 Introduction

Wikipedia is a consensus-based, community-moderated encyclopaedia, and increasingly a core training and retrieval source for large language models. Search-augmented models answer in Wikipedia's own words: about a third of answers carry a run of eight or more consecutive words copied from an article (§5). Whoever wrote those words on Wikipedia is, in effect, writing part of the answer. Wikipedia blocks and bans the accounts it finds editing in bad faith, but a block typically removes the editor, not the edit. More than half the prose standing in the articles we study was written by accounts blocked at least once, a label that runs from a single expired sanction to a standing ban, and when models quote those articles in their answers, that prose makes up just as much of what they quote.

The provenance, propagation and representation of content from atypical or malicious actors remain little studied. Work on problematic Wikipedia editing, whether vandalism detection, sockpuppet identification or undisclosed paid editing, stops at the encyclopaedia. Its outcome is whether an edit was reverted or an account blocked. But most edits are never removed, including most of what sanctioned accounts wrote, and nobody has followed that text further. The record of who wrote which words sits in the revision history and appears nowhere in a model's answer.

We ask two questions. **Detection:** can accounts with unusual editing histories be picked out from public edit history alone, with rules fixed in advance and no block records consulted? **Transmission:** when models quote an article, does the authorship of what they quote match the authorship of the article? Detection decides whether provenance can

be traced forwards instead of reconstructed after the fact. Transmission decides whether a sanction ever reaches the reader.

We work on 62 articles related to Israel and Palestine, chosen because many sit under **extended-confirmed protection (ECP)**: Wikipedia’s rule that only accounts with at least **500 edits** and **30 days** of registration may edit them. That gate, and a dense sanction history, is why the topic is usable here. Editors sit on a scale ordered by strength of evidence. Tiers 2 and 3 restate Wikipedia’s own record: past block episodes, and sanctions still in force. Tier 1 is computed from edit history alone: BURST-95 flags accounts that compressed most of their editing into one month, ECP fast pivot those that crossed the 500-edit / 30-day gate and moved into the protected topic within days. All 683,276 prose words standing in those articles are traced to the account that added them.

For detection, each Tier 1 rule is checked against block history, compared with accounts of similar editing volume, since block rates rise with activity. For transmission, we take 1,200 answers from three models, half from questions written against the corpus and half from Google autocomplete, match runs of eight or more identical words, and compare a group’s share of the quoted words with its share of the article words. If the two shares match, nothing in the pipeline is filtering by author.

A flag is not a finding of misconduct: about 60% of flagged accounts have never been blocked. We do not fact-check any of the statements made by these authors, and we do not argue that models should filter by author. Search was forced on, so this measures retrieval and leaves memorisation untested, in three commercial products as they behaved on one day, on one topic, in one language.

§3 sets out the tiers. §4 gives the attribution method, the two selectors and the query protocol. §5.1–5.3 validate the selectors upstream; §5.4 reports what models paste. §6 takes up what travels when the pipeline never sees who wrote the text.

2 Related work

2.1 Word-level authorship attribution

WikiWho (Flöck & Acosta, 2014) reconstructs which revision and which account introduced every token in a revisioned page, and TokTrack (Flöck, Erdogan & Acosta, 2017) publishes that provenance for all of English Wikipedia. The authorship table used here is built from WikiWho output over the 62-article corpus. We add nothing to attribution itself, only

what it carries: each token labelled with its author’s sanction status and selector flags, and where it goes, matched against text generated outside Wikipedia. Token provenance has been used to describe the encyclopaedia; here it is used to follow the encyclopaedia’s prose out of it.

2.2 Persistence as endorsement

Adler et al. (2008) measure an editor’s contribution by how long their text lasts, on the argument that surviving text has been tacitly accepted, and Halfaker et al. (2009) make the complementary case that reversion is the community’s clearest act of rejection. Both treat persistence as a proxy for community judgement. On the corpus studied here the proxy is weaker than it looks. Of bytes those 995 accounts ever added, 15.5% is still standing against 7.6% for the corpus as a whole (§6.3), and more than half the prose now standing was written by accounts blocked at least once. Neither fact shows that the community endorsed that text, and neither shows that removal failed: most of what anyone writes is eventually cut, so whoever’s text sticks a little more ends up over-represented among what remains (§6.3), and blocks are often for conduct that need not concern the surviving text at all (§3.1). Longevity records what escaped removal, which is not the same as what the community approved, and the distinction becomes consequential once that text is read by people who never see a revision history.

2.3 Detecting problematic editors

Kumar, West & Leskovec (2016) show hoaxes persisting for years before anyone catches them, the precedent for content that outlives scrutiny. Later work identifies editors from content and behaviour, with Joshi et al. (2020) on undisclosed paid editing and Raszewski & de Kock (2025) on sockpuppetry, while ORES (Halfaker & Geiger, 2020) is what Wikipedia runs in production. All of it stops at the account. Judged as detectors our Tier 1 rules are poor, since roughly 60% of flagged accounts have never been blocked, but a detector needs precision because it triggers an intervention, where a provenance instrument needs only to be computable from public data, fixed before the outcome, and checkable by hand for any single account. Ajmani, Vincent & Chancellor (2023) sit closest to our second selector, studying the same extended-confirmed gate: they ask how it reshapes the editor population, we ask what an account’s timing at it predicts.

2.4 What language models reproduce

Carlini et al. (2023) and Nasr et al. (2023) quantify and extract memorised training data; training-data attribution asks which documents caused an output (Chang et al., 2024), and corpus inspection asks what is in the training set at all (Elazar et al., 2024). Search is forced on in every call here, so retrieval is the channel we measure, and the unit throughout that literature is the document where ours is the author of individual words inside one. Brazilek et al. (2026) is the nearest neighbour, tracing Wikipedia advocacy editing into model behaviour, but differs on three axes: the channel is training-time where ours is retrieval, the models are small open-weight where ours are deployed products, and the population is a self-identifying coalition where ours comes from sanction records and rules fixed in advance.

2.5 Attribution in generative search

Liu, Zhang & Liang (2023) find that citations often fail to support the sentences they attach to; ALCE (Gao et al., 2023) benchmarks citation generation; Yang (2025) characterises which sources get cited. All take the source as the unit. A Wikipedia article is one source with one URL and thousands of authors, so source-level attribution is invariant to the distinction measured here: a system can cite Wikipedia impeccably and still say nothing about who wrote the sentence it took. That invariance is what we mean by authorship-blind, and it is why an effect of this kind could not have surfaced in prior evaluations, however carefully they were run.

3 Anomaly framework

3.1 Tier definitions: classifying editors by strength of community evidence

Whether an editor is "anomalous" is not a binary judgement, and we do not treat it as one. Community evidence about an account ranges from nothing at all, through repeated administrative intervention, to formal technical findings such as confirmed sockpuppetry. We therefore organise editors on a three-tier scale ordered by **strength of evidence**, and stratify every downstream measurement by tier, so that "anomalous" is never collapsed into a single label.

The frozen tier lists hold 995 accounts. The per-tier counts that hold surviving prose and enter downstream measurement, the live sanction re-verification and its two exceptions are in Appendix C.1.

Alongside the tier scale we track a coarser binary label, **ever blocked**: at least one non-copyright block episode in the account's public block log, regardless of whether the block is still in force. Its relationship to the tiers differs by tier. Tier 2 accounts are ever-blocked by construction, their defining feature is *past* block episodes whose sanctions have since expired. Tier 3 accounts are sanctioned *as of the block-log extract*, which in nearly all cases means a local block currently in force (the exceptions are sanctions enforced outside the local block log, chiefly WMF/global actions and standing ArbCom or community bans that persist without an active block). The Tier 2/Tier 3 boundary is sanction status, not conduct severity: an account indefinitely blocked for Tier-2-type conduct sits in Tier 3 until unblocked. Tier 1 carries no block requirement at all: most Tier 1 accounts have no block history (§5.1–5.2), which is what makes block status a valid *validation target* for the Tier 1 selectors, since no selector consumes it as an input. We use ever-blocked as the primary upstream validation label and the headline downstream pool because it is the broadest sanction signal that is independently observable from public data.

3.2 Predicted (Tier 1) versus observed (Tiers 2–3) classification

The tiers differ in kind as well as in degree. Tiers 2 and 3 are **observed**: they restate what Wikipedia's own processes have already recorded in block logs and sanction records. Tier 1 is **predicted**: it is computed from public edit history using rules we fixed in advance, with no access to block logs, revert history, or any judgement about the content an account wrote.

This separation is what makes the framework testable end to end. Because Tier 1 uses no sanction information as input, we can validate it *against* sanction records without circularity (§5.1–5.2). And because all tiers are frozen before any model is queried, downstream reproduction rates (§5.4) cannot leak into tier assignment.

3.3 Selection criteria for Tier 1 metrics

Candidate Tier 1 metrics were screened against four pre-specified criteria:

Yield screens on survival, which is also the quantity the enrichment in §6.3 reports, so the Tier 1 contribution to that figure is not independent of how the selectors were chosen. Tiers 2 and 3 carry no such dependence, their membership coming from Wikipedia's block log. The dependence runs against

Tier	Who decides	What qualifies	Sanction in force?	Interpretation
1 (289 accounts)	Our pre-specified rules	Structural edit-history patterns: BURST-95 or ECP fast pivot (defined in §4.3), union of the two	No, 0 of 289 . Not used in <i>flagging</i> (which sees no sanction data), but enforced at <i>membership</i> : Tier 2/3 members and any currently-blocked account are subtracted from the union	Structural patterns without a standing sanction. 88 of the 289 have expired blocks below the Tier 2 threshold
2 (116 accounts)	Wikipedia administrators	≥ 3 separate block episodes (edit warring, conduct, arbitration enforcement), after excluding bots, reversed blocks, self-requested and compromised-account blocks, copyright blocks, and blocks last applied before 2016	No block in force, 0 of 116 . 17 carry logged topic-area sanctions whose current status is unverified	Repeated intervention, no standing block. Blocks need not relate to the topic area
3 (590 accounts)	Wikipedia + technical evidence	Standing sanctions: CheckUser findings, ArbCom actions, community and WMF bans, related formal enforcement	Yes, 588 of 590 currently blocked . The other 2 are bans enforced without a local block (one WMF/global ban, one ArbCom ban)	Strongest community evidence

Criterion	Question it answers
Defensible	Does "atypical" follow from the measurement alone, with no claim about intent?
Auditable	Can a reviewer verify the flag for one account in one click, from public history?
Independent	Is it measured from account history, not from the content under test and not from sanction records?
Yield	Does it select accounts with <i>surviving</i> corpus text, so there is something to trace downstream?

the reported result: Tier 1 enriches least of the three tiers ($1.89\times$ against $2.11\times$ and $2.25\times$), so pooling it lowers the headline rather than raising it.

Two selectors passed and were frozen for the paper: **burst editing** (temporal concentration of mainspace edits; BURST-95) and **ECP fast pivot** (rapid entry into extended-confirmed-protected topic articles immediately after crossing the eligibility gate).

3.4 Downstream null hypothesis: proportional reproduction ($Y/X \approx 1$)

The framework's downstream test is deliberately simple. For any pool of editors (a tier, or the ever-blocked set), let $\mathbf{X}$ be the pool's share of tokens in the frozen article corpus, and $\mathbf{Y}$ be the pool's share of tokens inside verbatim spans that models reproduce (§4.2). If nothing on the path from article to answer acts on authorship, pasted text should track

surviving text: **$Y/X \approx 1$ for every tier**. Departures from that null have names we use throughout: **suppression** is under-reproduction relative to corpus share (lift < 1 ; a lift of 0.93 is 7% suppression), and **amplification** is over-reproduction (lift > 1 ; a lift of 1.22 is 22% amplification). $Y \ll X$ would mean the retrieval or generation path systematically excludes that tier's prose; $Y \gg X$ would mean it preferentially surfaces it. The null (lift ≈ 1) is itself the substantive claim under test.

Lift measures composition: whose prose ends up in the answer, against whose prose is standing in the article. That holds whatever route the text takes. No pipeline in this study can see authorship, so a departure from 1 has to run through some property of what the pool wrote, such as article position, register or sourcing density. Those are candidate mechanisms for a departure, not rival explanations that would make one spurious. Which property is the question §6.2 takes up, and the answer does not change what share of a model's output a sanctioned account wrote.

3.5 Conservative operationalisation and interpretive scope

Tier 1 criteria are intentionally weaker than a production moderation system would use: thresholds are calibrated so that flagged accounts sit in the extreme tail (top ~5% within activity band for burst), and no content signals are used. Results should be read as a **lower bound** on the prevalence and downstream reach of structurally anomalous editing. Flagging an account is **not** a judgement that the editor did anything wrong; it records that their measurable editing history is statistically unusual relative to peers at similar activity levels.

4 Methods

4.1 Data and cohorts

All upstream measurements use English Wikipedia **MediaWiki History (MWH)**, snapshot **2026-06**, and count only edits to ordinary articles unless stated otherwise. Bots are excluded by username at scan time; the random cohort also drops bot-flagged dump rows and anonymous editors whose name is an IP address.

Frozen corpus. The downstream corpus is 62 mainspace I/P articles, frozen as a slug list before any model was queried. The topic was chosen because extended-confirmed protection (only accounts with 500 edits and 30 days' tenure may edit) gives it a well-defined policy gate (used by the ECP selector, §4.3.2) and a high density of community-sanctioned editing history. Beyond this section the topic is referred to as I/P throughout; nothing in the method depends on which contentious topic area it is.

Selection cannot manufacture the result. The downstream test is a ratio measured *inside* the corpus: **X**, a pool's share of surviving article prose, and **Y**, its share of the prose models quote, are computed over the same 62 articles. Whatever the article choice does to the amount of sanctioned prose present, it does it to X and Y alike. The question the paper asks, whether the prose models quote is drawn in the same proportions as the prose available to be quoted, is invariant to it. There is no article list on which quoted authorship matches article authorship by construction; that correspondence is a property of the path from article to answer.

We make no claim about the *direction* in which the topic area shifts the amount of sanctioned prose, because two mechanisms pull against each other and this study does not separate them. Contested articles attract more sanctioned editors, which adds such prose; they also attract more scrutiny and reverting, which removes it, and the upstream survival

evidence indicates removal is substantial (§6.3). The net is an empirical matter, and what we observe is only the net: 52.2% of surviving prose in this corpus was written by accounts blocked at least once. If heavier scrutiny does suppress it, that figure understates what a quieter topic area would carry, which points the same way as the survival argument below.

What hand assembly does bound is generality. These are I/P articles chosen by a person, so the corpus supports a claim about *these* articles and invites the question of other topics without settling it. Two things limit how far that concern reaches. The set is visibly not curated for contentiousness: the topic's central flashpoint articles sit alongside entries on regional geography, food and individual biography, and article sizes span 891 to 25,002 attributed words. The full list is published with the corpus, so a reader can check the composition directly (Data availability). And the direction of any residual bias is conservative: I/P articles sit under extended-confirmed protection and sustained community review, so prose surviving here has passed a stricter filter than prose surviving on a typical article. Text left standing under this much scrutiny is, if anything, less likely to be reverted elsewhere, which makes survival measured here a floor.

Population gate (span ≥ 30 days). Scored populations include only accounts whose mainspace editing span is at least 30 calendar days, since otherwise a 30-day burst window is saturated by construction. After the gate the Random editor sample has 17,513 scored accounts and I/P has 18,565. The same gate applies to peer denominators in validation.

Activity banding. Editors are bucketed by lifetime mainspace edit count into six bands, 50–200, 200–1k, 1k–5k, 5k–10k, 10k–50k, 50k+, because burst share declines mechanically with total activity. All thresholds are calibrated, and all peer comparisons matched, **within band**.

Block labels. Upstream validation uses two labels from the public block record, each tied to a stated question in §4.3:

- **Ever blocked (non-copyright):** ≥ 1 non-copyright block episode in `block_accounts.csv`, whether or not still in force. Used for BURST-95 validation (§4.3.1) and for ECP Question 2 (§4.3.2).
- **Currently blocked:** block in force at the frozen ECP check (dated API snapshot). Used only for ECP Question 1 (§4.3.2).

Downstream tier assignment (Tiers 2–3, §3.1) is carried per word in the authorship table below. We do not further slice blocks by reason.

Cohort	n	What it is
Random editor sample	20,000	A fixed random sample of English Wikipedia editors (seed 42), each with at least 50 article edits in their career. Used to set the burst cutoffs, to see how common the flags are on ordinary Wikipedia, and to check that findings hold outside the I/P topic.
I/P global	18,839	Editors who touched any of the 62 I/P articles (the article itself or its talk page). From 27,286 such accounts we kept those with at least 50 article edits anywhere on the wiki.
Prolific-editor pool	5,461	An extra pool of high-volume editors used only when setting burst cutoffs in the busiest activity bands, so those cutoffs are not noisy. Never used as a comparison group in results.

4.2 Downstream measurement: verbatim reproduction and word-level authorship

Word-level authorship. For every article in the frozen corpus we compute surviving-prose authorship one word at a time: 683,276 prose tokens across the 61 prose-bearing articles of the 62-article corpus, each mapped to the account whose edit introduced it, with that account’s tier and block labels attached. This table is the source of both the corpus baseline **X** and the attribution of matched spans. Word-level attribution is a different measurement from a coarser proxy, and the difference is material: an earlier line-level survivorship weighting on this same corpus reversed the headline result.

Frozen editor lists. The Tier 1 lists were fixed before the first model query: 683 BURST-95 accounts and 640 ECP fast pivoters, a union of 1,236 with 87 accounts on both. Freezing them in advance is what keeps downstream reproduction out of tier assignment (§3.2).

Two question designs. Two sets of 200 questions were put to the same three models, complementary by construction, each strong where the other is weak, which is what makes their agreement informative. Method 1 is generated from corpus passages of known authorship, which lets it ask whether a model reproduced *the passage we pointed it at* rather than only whether it reproduced the corpus somewhere. That anchor is also its weakness, since the passages over-represent anomalous-tagged authorship, and Method 2 exists to test whether the oversampling matters. Method 2 draws real search-autocomplete queries and has no anchor at all, so it cannot carry a targeted comparison and for the same reason cannot produce a tier result that is an artefact of question design. Both instruments are frozen and checksummed; their construction is in Appendix A.

One gate matters for this study in particular: a question that copies a five-word run from the article would hand the model the text whose reproduction we are measuring. The full funnel is in Appendix A.2 and worked examples from both instruments in Appendix A.3.

Models and querying. Claude Sonnet 4.6, GPT-5.6 Terra and Grok 4.3, with search forced on, no system prompt, the question text as the entire prompt, and one completion per question: 600 answers per design and 1,200 in total, with no errors, empty answers or truncations. Three parameters could not be applied uniformly, each imposed by a vendor constraint: temperature pinning, the five-search cap, and the London location pin. Retrieval volume therefore differs across arms for reasons that have nothing to do with the models, and §6.4 declines to read anything into it. Full protocol: `methods/DOWNSTREAM_QUERY_PROTOCOL.md`.

Overlap rule. A **verbatim span** is a run of at least eight consecutive identical tokenised words shared between an answer and the corpus prose stream. Paraphrase and semantic similarity do not count. Eight words is long enough that shared boilerplate rarely triggers it and short enough to catch genuine quotation, and excluding paraphrase makes every figure we report a lower bound on reproduction.

Comparison. For each pool, **X** is its share of attributed corpus prose tokens, **Y** its share of matched span tokens across answers, and **lift** is Y/X . Lift near 1 is the no-filter null of §3.4. Every matched span token is attributed by word-position lookup into the authorship table, with complete owner coverage. Both shares are computed over the 658,753 tokens that enter the matching stream. The full authorship table holds 683,276; a span can only match text that is in the stream, so **X** must be measured over the same tokens as **Y**. Every later figure uses that denominator without restating it.

4.3 Upstream measurement: Tier 1 selector definitions and validation design

4.3.1 Burst editing (BURST-95)

Definition. For each account, order mainspace revisions by timestamp and index days as `floor(epoch(ts)/86400)`. Let W^* be the maximum number of edits falling in any window of ≤ 30

	Method 1	Method 2
Source	Corpus passages, seeded generation	Google autocomplete via AnswerSocrates, region GB
Passage anchor	Yes, recorded to word positions	None
Article coverage	58 of 62	60 of 62
Known weakness	43.5% of questions target anomalous-tagged passages	On-topic gate retains article vocabulary; no not-answerable control arm
Question only it can answer	Did the model reproduce the passage we pointed it at?	Do the results hold for questions users actually type?

calendar days. Then:

- $\text{burst_30day_share} = W * / n_mainspace_edits$
- $\text{span_days} = \max(\text{day}) - \min(\text{day}) + 1$

BURST-95 = $\text{span} \geq 30$ days AND $\text{burst_30day_share} \geq \text{cutoff}[\text{band}]$. An account on BURST-95 concentrated an extreme share of its lifetime mainspace editing into a single month, relative to editors of similar total volume. Cutoffs are the 95th percentile of burst_30day_share among $\text{span} \geq 30$ editors in the Random editor sample, computed separately for each activity band; the six values are in Appendix B.1.

Setup. Scored on the Random editor sample and the I/P global cohort (§4.1), after the $\text{span} \geq 30$ population gate.

Comparison. One question: **are BURST-95 accounts more often ever-blocked than band-matched peers?** Peers are accounts in the same cohort and edit-count band, also with $\text{span} \geq 30$, who are **not** on BURST-95. Band matching is the fair comparison: BURST-95 accounts are extreme *within* their activity band, and block base rates vary strongly with volume. Results are also broken down by band. We do not treat I/P BURST-95 versus the Random editor sample as a primary estimand: that contrast mixes burst flagging with topic selection.

Limitations. Three limits bind what BURST-95 can claim. First, the flag is built from lifetime mainspace history, so a block can precede the burst month: the design is retrospective, not a test that burst comes before sanction. Second, most flagged accounts were never blocked (63.4% on I/P), so this is a moderate-precision risk marker for provenance work, not a screener and not evidence of wrongdoing. Third, the association with block history is weak above roughly 10,000 lifetime edits, and the flag is silent on *where* editing happened, and speaks only to *when* it was concentrated. We show that extreme temporal concentration and ever-blocked status travel together. We do not show cause.

4.3.2 ECP 500 fast pivot

Definition. I/P articles fall under extended-confirmed protection (ECP): an account needs **500 edits and 30 days of tenure** before it may edit them. Qualifying events are edits to ordinary articles, article talk pages, or user talk pages, counted in time order. The crossing index $n \setminus *$ is the later of the account’s 500th qualifying event and its first event at least 30 days after registration. An I/P edit is a revision to any of the 62 frozen corpus titles. **ECP fast pivot:** the account’s first such edit falls at or after crossing and within 14 days of it. The flag asks whether first engagement with the corpus follows that gate: accumulate qualifying edits elsewhere, then enter the protected topic when eligibility opens. **Pivot latency** is the time from crossing to the account’s first I/P edit among post-crossing entrants. Frozen parameters: 500 edits, 30-day tenure, 14-day pivot window (Appendix B.2).

Setup. Primary analysis uses the I/P global cohort (§4.1), restricted to **ECP-scorable** accounts: at least 500 total edits, a valid registration, an untruncated history extract reaching $n \setminus *$, and neither a bot nor an IP. The pivot target is the frozen 62-article corpus. The Random editor sample is used only as an exploratory check (Appendix C.3).

Comparison. Unlike BURST-95, which answers one question with one label, ECP validation is built around **two questions with two labels**. Accounts whose first I/P edit predates crossing are **topic veterans** and are excluded from Q1.

These are two separate comparisons, not the same result read both ways. Q1 asks whether people who are blocked *today* entered the topic quickly after becoming eligible. Q2 asks whether people who entered quickly have *any* block on their record, past or present. Only Q2 matches the BURST-95 check (does the flag enrich for ever-blocked?). The same design is run on the Random editor sample as an exploratory check, where weak replication is expected by construction since the pivot target is the I/P corpus itself; that check is reported in Appendix C.3.

Limitations. Four limits bind what this selector

	Question	Block label	What we compare
Q1	Do accounts that are blocked <i>now</i> enter the protected topic faster after crossing ECP?	Currently blocked	Fast-pivot rate among currently blocked vs not; and median pivot latency among post-crossing entrants (Q1 without the 14-day cut)
Q2	Do accounts that fast-pivot have more block history than other ECP-scorable editors?	Ever blocked	Ever-blocked rate among fast pivoters vs the ECP-scorable base rate

can claim. First, Q1 uses whether an account is blocked *now*, not whether it was blocked at the time of the pivot. Second, Q2 counts any past block, including blocks that happened *before* the account crossed ECP, so a fast pivot is not shown to precede sanction. Third, we only see edits to the 62 corpus articles, so I/P editing on other pages is missed, and the ECP rule itself was not applied the same way to accounts that registered before 2015: the 500-edit / 30-day gate was not yet a uniform lock on these pages, so an older account could reach that edit count and enter the topic when many articles were still open to newer editors. For those accounts, “crossing then pivoting” is not the same incentive shape the selector assumes. Fourth, and most important: we show association, not cause. An editor may hit the ECP gate, dive straight into a contested article, edit-war, and later be blocked, putting fast pivot and block on the same record because of that whole pattern of engagement, not because crossing the gate caused the sanction, and not because they were gaming the 500-edit rule. We do not separate those stories.

5 Results

5.1 Burst editing and block history

Accounts that pack most of their editing into one month are about one-and-a-half to nearly twice as likely to have a block on their record as otherwise similar editors. The flag fires on about 5% of the Random editor sample by construction, since the cutoffs are that sample’s own 95th percentile within each activity band: 875 of 17,513 scored accounts. The same fixed cutoffs applied to I/P fire on 683 of 18,565 (3.68%). That is not evidence of less bursting in I/P, because the cohorts have different edit-count mixes and the cutoffs are band-relative, so every comparison below holds activity band fixed.

What we test is co-occurrence, not cause. If the flag and block history were unrelated, BURST-95 accounts would be blocked at the same rate as non-flagged peers of similar volume. They are not. On I/P, 36.6% of BURST-95 accounts have ever been blocked (95% CI 33.1–40.3, $n = 683$), against 24.1% of those peers (CI 23.4–24.7, $n = 17,882$), a lift

of $1.52\times$ (Fisher $p \approx 7.7 \times 10^{-13}$). The same rule on the Random editor sample yields 21.6% versus 12.0% ($n = 875$ vs 16,638), $1.79\times$ ($p \approx 1.5 \times 10^{-14}$). The pattern appears on ordinary Wikipedia as well as in the contested topic. We do not claim that bursting produces blocks, or that a block produces bursting: only that the two travel together often enough that the flag is informative for provenance.

Within I/P that co-occurrence is concentrated among lower- and mid-volume editors and fades once lifetime edits run into the thousands (Table 1). Extreme month-concentration lines up with block history among ordinary editors and stops doing so among prolific ones. In the top band it reverses: 27.5% of the 40 flagged accounts above 50k lifetime edits have ever been blocked, against 45.1% of their peers, a lift of $0.61\times$. On 40 accounts that reversal is not itself a finding, but the flag carries no usable signal there and should not be read as if it does. On I/P, 63.4% of BURST-95 accounts were never blocked: enrichment is not the same as identifying wrongdoing.

Table 1. Ever-blocked rates for I/P BURST-95 versus non-flagged I/P peers in the same lifetime edit-count band.

5.2 Fast pivot after the extended-confirmed gate

Among I/P editors who clear Wikipedia’s extended-confirmed gate, both validation questions of §4.3.2 show association. Of 14,106 ECP candidates, 14,007 can be scored and 640 (4.6%) are fast pivoters.

Accounts that are blocked today enter the protected topic faster after becoming eligible. They fast-pivot at 9.25% (113 of 1,221) against 4.12% of non-blocked peers (527 of 12,786), a lift of $2.25\times$ (Fisher $p \approx 2.4 \times 10^{-12}$). The binary flag sits on a continuous timing gap: among those who eventually edit the corpus after crossing, the median wait is **81 days if currently blocked and 202 days if not** (Wilcoxon $p \approx 0$).

Accounts that enter quickly also have more block history. Of the 640 fast pivoters, 40.3% have ever been blocked (95% CI 36.6 to 44.2), against 27.9% among all ECP-scorable I/P accounts, a lift of $1.45\times$ ($p \approx$

Lifetime edits	n (BURST-95)	BURST-95 ever-blocked	Peer ever-blocked	Lift
50–200	53	34.0%	12.1%	2.80 ×
200–1k	139	46.8%	17.9%	2.61 ×
1k–5k	258	34.9%	23.4%	1.49 ×
5k–10k	100	31.0%	28.9%	1.07×
10k–50k	93	37.6%	34.3%	1.10×
50k+	40	27.5%	45.1%	0.61×

3.8×10^{-12}). That is the same style of check as for BURST-95, and roughly the same effect size.

5.3 Comparing the two Tier 1 selectors

On fair ever-blocked baselines the two selectors give essentially the same answer, about 1.5× enrichment on I/P, from almost non-overlapping sets of accounts.

One rule looks at *when* editing was concentrated; the other at *where* an account went next after a policy gate. Their union is 1,236 accounts, and only 87 (7.0%) sit on both lists, so Tier 1 is not one pattern counted twice. What they share is the size of the ever-blocked enrichment; where they differ is generality (burst replicates off-topic; gate-pivot does not) and how much of the corpus each later owns downstream (§5.4, §6.2).

5.4 Verbatim reproduction by editor tier

Across two independent question designs, ever-blocked editors’ prose is reproduced verbatim at essentially its corpus share. The product null holds. $\mathbf{X}$ is identical across methods by construction, since both score against the same corpus and the same authorship table.

Table 2. Pooled over all 1,200 answers, with $\mathbf{X}$, $\mathbf{Y}$ and lift as defined in §4.2 and 1.0 the no-filter null of §3.4. Intervals are 95% percentile bootstrap CIs resampling **answers**, not tokens: one pasted passage contributes hundreds of mutually-agreeing tokens, so a binomial interval over token counts would badly understate the uncertainty. Answers containing no verbatim span remain in the resampled population (10,000 resamples, seed 42; `definitions.downstream.bootstrap_lift_ci`).

The pools come from three different labelling systems and are **not** a partition of the corpus, so the table is grouped by labelling system, and row order carries no ranking. Tiers 1–3 are mutually exclusive: every attributed token belongs to at most one, and together they cover 42.5% of the corpus. The frozen selectors overlap each other and cut across the tiers. Ever blocked cuts across everything.

The last column counts answers containing at least one span from the pool, the evidential weight behind each row: BURST-95 rests on 14. $\mathbf{X}$ is identical across both question designs, which score the same corpus against the same authorship table. The all-tiers row is derived from the underlying token counts, the tiers being mutually exclusive: $X = 279,812 / 658,753 = 42.5\%$, and Y is the same sum over matched span tokens. Its answer count is blank because an answer quoting two tiers would count twice. "Tier 1" names two objects: the mutually exclusive editor tier (23.0% of the corpus, sanctioned-free by construction) and the union of the two frozen selectors (29.7%, overlapping the sanctioned tiers). Both are reported; neither substitutes for the other.

The primary result. Ever-blocked authors hold 52.2% of attributed corpus prose and supply 53.6% of pasted span tokens under Method 1 and 58.5% under Method 2, giving lifts of 1.03× and 1.12×, and $1.08 \times [0.93, 1.22]$ over all 1,200 answers. The pooled interval contains the null and, per §3.4, rules out suppression beyond about 7% (lift below 0.93) and amplification beyond about 22% (lift above 1.22), so this is a null with resolution, bounded by data sufficient to constrain the effect. Neither design shows the suppression a downstream authorship filter would produce. Whatever Wikipedia’s own processes remove upstream, what survives into an article is reproduced without regard to who wrote it.

Ever blocked is not the tier scale, and the two do not give the same answer. Ever blocked is a sanction record, held by 52.2% of corpus prose. The tiers are our evidence scale, and cover 42.5%. They overlap heavily but not by construction: Tier 2 is ever-blocked by definition, Tier 3 nearly always so, and **Tier 1 is independent of block history**, since most Tier 1 accounts have never been sanctioned (§3.1). Reading the same 1,200 answers through the tier scale therefore gives a different figure, and a different conclusion: the three tiers combined are reproduced at **1.25**× [**1.06, 1.44**], which excludes the null.

The gap between 1.08× and 1.25× is Tier 1, at **1.57**× [**1.17, 1.96**]: the one pool both substantially over-reproduced and largely outside the sanction label, since Tier 1 membership excludes Tier 2, Tier 3 and anyone currently blocked. The two sanctioned

	BURST-95	ECP fast pivot
I/P lift, ever blocked (fair baseline)	1.52 × vs band peers	1.45 × vs ECP base
Replicates on Random editor sample?	Yes (1.79× vs band peers)	Weak (32 pivoters; p = 0.39)
I/P-specific by design?	No	Yes
Temporal ordering	Retrospective (lifetime)	Partial (post-crossing)
Flagged n (I/P)	683	640
Never-blocked share	63.4%	59.7%

Pool	X (corpus)	Y (pasted)	lift [95% CI]	answers with span
<i>Editor tiers (mutually exclusive)</i>				
Tier 1: flagged, no standing sanction	23.0%	36.1%	1.57 × [1.17, 1.96]	220 / 1,200
Tier 2: repeat blocks, expired	8.0%	7.9%	0.98 × [0.64, 1.45]	142 / 1,200
Tier 3: standing sanction	11.5%	9.3%	0.81× [0.54, 1.13]	129 / 1,200
All three tiers	42.5%	53.2%	1.25 × [1.06, 1.44]	n/a
<i>Frozen Tier 1 selectors (overlapping)</i>				
BURST-95	1.3%	0.7%	0.58× [0.07, 1.65]	14 / 1,200
ECP fast pivot	28.6%	40.7%	1.42× [1.10, 1.73]	245 / 1,200
Tier 1 union (burst ∪ ECP)	29.7%	41.4%	1.39× [1.09, 1.69]	252 / 1,200
<i>Block label (cuts across both)</i>				
Ever blocked	52.2%	56.1%	1.08 × [0.93, 1.22]	335 / 1,200

tiers do not contribute, sitting at 0.98× and 0.81×. It lifts the tier figure while leaving the ever-blocked figure near 1. The two answer different questions. Does a community sanction reduce how much of an author’s prose models quote? No. Is prose by editors our rules flag reproduced above its corpus share? Yes, modestly, and §6.2 sets out why that says nothing about whether the flags identify misconduct. We report ever blocked as the headline because it is the label independently observable from public records.

Method 2 elicits more copying, but the same structure. The Google-style questions raise the paste rate by roughly half (39.0% vs 26.5%) while leaving every tier-level conclusion in place: direction, ordering, and the identity of the driving selector are unchanged. Two plausible contributors to the volume difference are not separable here: real search questions skew toward lead-paragraph material, the most-syndicated part of any article, whereas Method 1 questions deliberately reach into article bodies and are spread across lead, middle and late positions. The same skew bears on the lifts, and ECP fast pivot does sit higher under Method 2 (1.50×) than Method 1 (1.34×). Measured directly, position does not carry it: corpus prose splits evenly across lead, middle and late thirds, and the largest lead over-representation of any pool is 1.09×. §6.2 takes up what that leaves. That Method 2 agrees despite having **no anomalous-passage oversampling** (§4.2) is the substantive point, because it shows the Tier 1 result is not an artefact of Method 1’s question design.

Tier 1 over-representation is modest, ECP-driven, and resolved only when the designs are pooled. Tier 1 union text is reproduced above

corpus share in both point estimates, but under Method 1 alone the interval includes the null (1.34× [0.92, 1.75]); it separates from 1 under Method 2 (1.45× [1.06, 1.85]) and over the pooled 1,200 answers (Table 2). We therefore state Tier 1 over-representation as a pooled result, not as an independently significant finding in each design. Within Tier 1 the effect belongs almost entirely to ECP fast pivot, and BURST-95 is reproduced at or below corpus rate. The burst figure rests on a 1.3% corpus share and single-digit answer counts, and its intervals show it: uninformative under Method 1 and spanning the null when pooled, so it should be read as *no evidence of over-representation* rather than as an estimate of suppression, including the one arm where it separates from 1.

Whatever drives that excess, it is not sanction status but a property of the accounts our selectors pick out, most of which Wikipedia never acted against.

Tier 3 points below corpus share, but not significantly. Prose by authors under a standing sanction is reproduced below its corpus share in both designs (0.79× and 0.83×, pooled 0.81×), consistently and in the opposite direction from the amplification hypothesis, but every interval includes the null (pooled [0.54, 1.13]). Tier 3 holds 11.5% of corpus prose, so the pool is small enough that a real depletion of this size is not resolvable at this sample. We report the direction and decline to claim under-representation. The pool covers all standing sanctions, sockpuppetry among them: 61.8% of its prose comes from CheckUser or sockpuppetry blocks and the remainder from indefinite conduct blocks, ArbCom and community bans.

Tier 2 sits on the null. Editors with repeat block

history whose sanctions have since expired hold 8.0% of corpus prose and supply 7.9% of pasted tokens, a pooled lift of $0.98\times$, the closest any pool comes to exact proportionality, from 142 of the 1,200 answers. The two designs bracket it from either side ($1.10\times$ and $0.87\times$), which is the pattern expected when a true value near 1 is estimated twice under different question distributions. The pooled interval is wide, $[0.64, 1.45]$, so this is consistent with exact proportionality without demonstrating it; what it rules out is a large effect in either direction.

Model differences are not a controlled comparison. Paste rates order identically in both methods (Sonnet 34.5%/48.5%, Grok 32.5%/49.5%, GPT 12.5%/19.0%), but the arms differ in search cap enforcement, temperature pinning and location pinning for the reasons given in §4.2, so these are provider-and-configuration differences, and say nothing about model behaviour under matched conditions.

6 Discussion

6.1 Proportional reproduction and the absence of authorship in the retrieval path

Ever-blocked editors wrote 52.2% of the prose standing in the 62 corpus articles, and supply 56.1% of the words models quote back verbatim. The ratio of those two shares, which we call lift, is 1.08 with a 95% interval of $[0.93, 1.22]$. The interval contains 1.0, and on the definitions of §3.4 it is narrow enough to rule out suppression beyond about 7% (lift below 0.93) and amplification beyond about 22% (lift above 1.22). This is a null with resolution, bounded by data sufficient to constrain the effect, and that distinction decides whether an absence of effect is a finding or a shrug.

These systems do not make the distinction at all. Prose the community sanctioned and prose it never touched arrive at the same text, because nothing on the path distinguishes them. A Wikipedia article is one source and one retrieval decision written by thousands of accounts, so naming that URL as the provenance of a sentence says almost nothing about who wrote it. The revision history records exactly that, for every word, and it is not part of what a retrieval layer fetches or a generator sees. Every filter in the pipeline acts on the source; none can act on the author, because by the point the text is selected the author is no longer attached to it. That is why the effect holds across three products, so it cannot be a quirk of one ranking function, and why it would not have surfaced in evaluations that take the source as the unit (§2.5).

One thing could hide a real effect behind a ratio near 1. If the ever-blocked set wrote a disproportionate share of the passages models favour, an upward pull from position could offset a genuine downward filter. It does not arise here: corpus prose divides evenly across lead, middle and late thirds, and this pool's share of lead prose is $0.99\times$ its share overall. The null is a null, not two effects cancelling.

A null carries a standing risk that a badly-built instrument produces it by accident, and the defence here is that two instruments produced it whose weaknesses do not overlap. Method 1 can trace every question to the passage it targeted but oversamples anomalous-tagged passages by design; Method 2 draws its questions from real search autocomplete and has no passage anchor at all, so it cannot be biased by passage selection because it selects no passages. They disagree on how much models quote and agree on every tier-level conclusion, which is the pattern expected if the tier result is real and the volume of quotation is a property of the questions asked.

6.2 Divergence between the two selectors upstream and downstream

Tier 1 prose is quoted above its share of the corpus, at a lift of 1.39 $[1.09, 1.69]$ pooled, and almost all of that belongs to ECP fast pivot at 1.42 $[1.10, 1.73]$; BURST-95 contributes almost none of it. The natural reading is that one rule found something downstream and the other failed. The measurements do not support that reading.

The two selectors own very different amounts of the corpus, and the gap is mostly definitional. ECP fast pivot fires on an account's first edit to one of the 62 corpus articles, so every fast pivoter is a corpus author by construction. BURST-95 fires on how an account's lifetime mainspace editing is distributed in time, anywhere on the wiki, and asks nothing about the corpus; cohort membership needs only one edit or talk-page comment on a corpus article. One rule guarantees itself a foothold in the text we measure and the other does not, which is most of why ECP fast pivoters wrote **28.6%** of the surviving prose against BURST-95's **1.3%**.

At 1.3%, the burst pool is too small to read a lift from. Under Method 1 its interval spans from near zero to above 3, and pooled it still crosses the null (§5.4). **The BURST-95 result downstream is an absent measurement**, not evidence that the rule failed.

Upstream, where the unit is the account, the two selectors are comparable: **1.52 $\times$** and **1.45 $\times$** the block rate of their baselines (§5.2, §5.3). BURST-95

earns its place in the Tier 1 union on that account-level validation, and on a property ECP cannot claim: it alone replicates on a general-wiki sample ($1.79\times$ on random20k), with no topic policy and no I/P corpus in the rule. Downstream silence from a 1.3% prose pool does not retract that.

The ECP lift needs a route, because no pipeline here can see authorship (§3.4). The obvious candidate is position: quotation concentrates in leads and definitional sentences, so a pool that wrote a disproportionate share of leads would be over-quoted by a system indifferent to who wrote what. That would not have made the lift spurious. Owning the lead of a contested article is the most consequential position on the page, and it would be a finding about where these accounts worked rather than an artefact to be subtracted. It is also not what happened. The largest lead over-representation of any pool is $1.09\times$, against lifts reaching $1.57\times$, and skew does not track lift across the seven pools: Tier 2 skews toward leads as strongly as ECP and shows none. Register, sourcing density and sentence shape are untested and are the candidates left standing. None of them changes the share itself. A model's answer is composed of prose written by particular accounts in particular proportions, and that composition holds whatever property of the writing routes it there.

6.3 Upstream removal and the surviving residue

Wikipedia does remove and overwrite text: most bytes ever added to these articles are gone (only 7.6% remain). Even so, a large residue of tier-authored text remains. The 995 tier-list accounts wrote 19.1% of bytes ever added and hold 38.8% of the wikitext still standing, about twice their share of churn. Measured on reader-visible prose rather than wikitext, they wrote 42.8% of what is on the page today. The enrichment holds tier by tier and is strongest where it is cleanest: $2.25\times$ for Tier 3 and $2.11\times$ for Tier 2, whose membership comes from the block log, against $1.89\times$ for Tier 1, the one tier whose selectors were screened partly on survival (§3.3).

That gap is not proof that removal failed. If almost everyone's text is cut, anyone whose text sticks a bit more will look over-represented among what is left. Sticky text is not community approval. What matters for models is the residue, the words still on the page, not the removed edits.

Upstream, moderation hits accounts and only partly cleans old text. Downstream, models quote what remains, roughly in proportion to who wrote it (§6.1). Nothing in between re-checks whether the author was sanctioned: a block stops future edits and often

leaves past surviving prose in place, and that is what models read.

Retrieval is not the same as citation, and neither is what we measure. In the Method 2 run, Wikipedia's share of *retrieved* results was similar across all three models (11–16%), while the share surviving into *citations* was not: GPT-5.6 Terra cited Wikipedia on 8 of 200 questions while retrieving it on 181. A study asking "does the model use Wikipedia" would get a materially different answer depending on which half it measured. The verbatim measure used here deliberately reads neither the retrieval log nor the citation list. It scores the answer text itself, which is why it can register reproduction in answers that never cite the source, and why a reader has no way to know that the sentence in front of them came from an article at all.

6.4 Limitations and scope of the measurement

The measurement is narrow by construction.

Search was forced on in every call, so this is retrieval and not memorisation, and nothing here separates the two channels. The models are three commercial products as they behaved on 2026-08-07, on 62 I/P articles in English, and provider-side models change without notice. Search caps, temperature and location pinning could not be applied uniformly across arms, so per-model differences are configuration differences and not evidence about model behaviour. One completion per question leaves within-question variance unmeasured. We measure reproduction and never correctness: nothing here says that any tier's prose is wrong, and the design could not detect it if it were. Roughly 60% of Tier 1 accounts have never been blocked, so a flag is a provenance handle and not a finding of misconduct. The upstream persistence figures rest on net byte deltas and are order-of-magnitude.

Which property of a pool's prose carries its lift is open, and §6.2 treats that as a question about mechanism rather than a limitation on the composition result.

6.5 Implications for moderation

Authorship is public and recoverable from the revision history, and it is discarded at the point where text is selected. The measurement in this paper was only possible because we reconstructed offline, at word level, information the pipeline had already thrown away. Whether that information ought to change what a model says is a separate question, and one no deployed system is currently in a position to ask.

We are not arguing that models should filter by author, and the evidence here would not support it: a Tier 1 flag is a timing pattern, most flagged accounts were never sanctioned, and we make no claim that any tier’s prose is less accurate.

For Wikipedia the gap runs the other way. Sanctioning an account closes off its future edits, and the community can see that it worked. The edits already standing keep travelling, into an audience that never sees a revision history.

6.6 Directions for further work

The mechanism question is the one worth pursuing. Position is ruled out on the writing side, which leaves register, sourcing density and sentence shape untested. Recording the corpus position of every matched span would close the position case completely, and would also show whether models quote leads as heavily as we assume. Beyond that, matching each pool’s sentences against comparable sentences from other accounts would say whether the prose itself is more quotable, which is the closest thing to an explanation this design can reach.

The strongest single test is the same measurement on a corpus chosen differently, another contested topic area or an uncontested one.

Running the same questions with search disabled would separate what these models retrieve from what they have memorised.

7 Conclusion

Accounts with unusual editing histories can be flagged from public edit history alone, using rules fixed in advance (§3.1), though the flags are risk markers of moderate precision rather than screeners. The community’s judgement about those accounts does not survive the journey into a model’s answer, in any of the three systems we tested. Editors on the three tier lists wrote **42.5%** of the attributed prose in the 62 corpus articles and supply **1.25×** that share of the words models quote back: anomalous-authored text is reproduced above its share of the page.

Wikipedia’s enforcement acts on accounts; what reaches a reader is text. A block ends an editor’s next edit and often leaves standing everything they wrote before it, and that standing text is what models quote. The revision history says who wrote it. Nothing that carries the text downstream carries that record.

8 Data availability

Everything needed to audit the design is public in the project repository: the frozen 62-article corpus list, **both question instruments with checksums** (Method 1 md5 93e88305262f617b9a6c81c4a0b1a640; Method 2 md5 fb06da64cd92b2ff70e504e5876ad275), every evidence bundle with its machine-readable numbers, the derivation and analysis scripts, and this manuscript. The Wikipedia inputs are themselves public: the MediaWiki History snapshot (2026-06), the enwiki block logs, and article revision history.

The **1,200 model answers** and the **word-level authorship table** are available from the authors on request. The answers are output generated by third-party commercial models, and redistributing another provider’s model output may be restricted by their API terms; we therefore supply them on request. Because provider-side models change without notice, the stored raw API responses, not a re-run, are the reproducible record of what was measured; the query parameters, and the vendor constraints that could not be applied uniformly across arms, are documented in full in `methods/DOWNSTREAM_QUERY_PROTOCOL.md`.

The appendices follow this section: A (question instruments), B (selector parameters), C (upstream secondary analyses), D (Method 2 replication tables).

A The question instruments

Both instruments are published as evidence bundles, so either design can be audited without access to the run.

A.1 Method 1: generated from corpus passages

200 factual questions generated from corpus passages of known authorship (`gen_questions.py`, seed 42), md5 93e88305262f617b9a6c81c4a0b1a640, roughly three per article across 58 of the 62, each anchored to a passage recorded to word positions.

The anchor is also the instrument’s weakness. Passages were sampled to over-represent anomalous-tagged authorship: 87 of the 200 questions (43.5%) target such passages, against 113 targeting normal-tagged ones. That design tag is **not** the frozen Tier 1 label and does not enter any result; it governed only which passages questions were written about. Method 2 exists to test whether the oversampling matters. Questions are spread across article position

by design: 59 lead, 67 mid, 74 late.

A.2 Method 2: real search-autocomplete queries

200 questions distilled from **58,185** AnswerSocrates autocomplete rows, one search per corpus article, region GB, scraped 2026-08-06, through a seven-gate filter leaving **1,113** survivors, a **5.6×** oversupply from which 200 were selected proportionally across **60 of 62** articles (1–5 each).

Gate	Removes	Rows dropped
0	not a question	44,809
1	exact duplicates	5,242
2	off-topic (<code>on_topic_min</code> 0.5)	2,629
3	copies article text (<code>copy_ngram</code> 5)	503
4	near-duplicates (Jaccard 0.8)	208
5	subject-naming / mis-filed	372
6a	template over-cap (max 2)	3,254
6b	not well-formed (model-judged)	55
	Total dropped	57,072

The funnel reconciles exactly: $58,185 - 57,072 = 1,113$. Gates 0–6a and the selection step are deterministic; gate 6b is model-judged (`claude-sonnet-5`) and cached, so a re-run reproduces the same 200. A Claude model therefore helped build an instrument on which a Claude model is one of three test subjects. The gate judges only whether a question is well formed, it never sees the corpus or any authorship label, and its verdicts are frozen in the published instrument, so it cannot move the paste rate or the authorship composition that §5.1 measures. The frozen instrument is `google_200.csv`, md5 `fb06da64cd92b2ff70e504e5876ad275`, byte-identical to the file sent to all three providers.

Because the oversupply is $5.6\times$, over-rejection is cheap and under-rejection is expensive, so the gates were tuned to discard aggressively.

Method 2 satisfies gate 3, the copied-run rule of §4.2, almost for free, because people do not phrase questions the way an encyclopaedia phrases sentences; Method 1 needs it enforced explicitly.

A.3 Example questions

Five from each instrument, selected by a mechanical rule so the sample cannot be curated after the fact.

Method 1. Taken in question order, keeping each question that introduces an author tier or an article position not yet represented, which yields five spanning all four values of the passage tier field, untiered included, and all three position bands.

Method 2. One question from each autocomplete section contributing more than a single row, plus

a second from the largest section, all from distinct articles, taken in question order. The **comparisons** section is omitted because it contributes one row in 200.

The contrast between the two tables is the point of running both designs. Method 1’s questions are long, precise and unmistakably constructed: they name their subject, pin a date or a quantity, and nobody would type them into a search box. Method 2’s arrive as users wrote them, lower-cased and unpunctuated, and several would be ambiguous without the article they were scraped against. Neither register is the right one to test on, which is why the paper reports both and treats their agreement as the result.

A.4 Known limitations of the Method 2 instrument

Recorded here and left in place, because the oversupply made over-rejection the cheaper error:

- **Mis-filed questions.** Roughly seven of the 200 are filed under an article they are not really about, residue of gate 5. These score as "did not use the article" whatever the model does, biasing the paste rate **downward**.
- **Surviving near-duplicates.** About nine near-duplicate pairs clear the redundancy gates, including three inverse pairs asking the same question in opposite polarity.
- **Coverage.** Two articles yield no questions, so Method 2 covers 60 of 62.
- **No control arm.** Questions that are on-topic but *not* answerable from the article were never built. The on-topic gate keeps questions whose vocabulary matches the article, and the study then observes that Wikipedia gets used.

None of these change the finding this paper states. The claim is a ratio, $\text{lift} = Y/X$, and X and Y are drawn from the same corpus and the same authorship table under both designs; an instrument that raises or lowers how often models quote Wikipedia moves the numerator and denominator of the paste-rate statistic while leaving the authorship composition of what gets pasted unchanged. The limitations bound the *absolute* paste rate, how much these models quote, and leave the tier comparison, which is what §5.4 asserts, intact. What follows from the two designs disagreeing on volume while agreeing on every tier-level conclusion is taken up in §6.1.

#	Article	Author tier	Position	Question
1	1982 Lebanon War	Tier 1	late	What was the total number of PLO artillery pieces, including cannons and rocket launchers, present in Lebanon just before the 1982 Lebanon War began?
2	1982 Lebanon War	Tier 1	mid	What was the name of the military operation that marked the beginning of the 1982 Lebanon War?
3	1982 Lebanon War	untiered	lead	How many Lebanese airspace violations were recorded during the period spanning August 1981 through May 1982, prior to the 1982 Lebanon War?
5	2014 Gaza War	Tier 2	late	What method did the organization barred from entering Gaza since 2012 use to carry out its research during the 2014 Gaza War?
23	Battle of Jenin (2002)	Tier 3	lead	What type of buildings did Palestinian fighters target for booby-trapping during the Battle of Jenin (2002)?

#	Article	Section	Question
1	1982 Lebanon War	people also ask	How many people died in the Lebanon war in 1982?
3	2014 Gaza War	people also ask	What is true about both the 2008-2009 and 2014 Gaza wars?
6	Al-Aqsa Mosque	questions	can i go to al aqsa mosque
16	Balfour Declaration	letters	did palestine agree to the balfour declaration
46	Edward Said	in the past	did edward said have children

B Selector parameters

The two Tier 1 selectors are defined in §4.3. This appendix gives the frozen constants that section states in words.

B.1 BURST-95 cutoffs by activity band

An account is on BURST-95 if its editing span is at least 30 days and its `burst_30day_share` reaches the 95th percentile for its activity band, measured on span ≥ 30 editors in the Random editor sample. The high bands are supplemented with Prolific-editor pool probe rows for calibration stability only; those rows never enter a baseline denominator.

The cutoff falls with activity because burst share declines mechanically as an account accumulates edits: concentrating half of a 200-edit history into one month is common, and concentrating half of a 50,000-edit history into one month is not. Banding is what makes the comparison fair, and every threshold and every peer comparison in §5.1 is computed within band.

B.2 ECP crossing and pivot parameters

Over edits to ordinary articles, article talk pages, or user talk pages, counted in time order, the crossing index $n\backslash$ is the later of the 500th event and the first event at least 30 days after registration. An account is scorable if it has at least 500 total edits, a valid registration, and an untruncated extract reaching $n\backslash$, and is neither a bot nor an IP.

An I/P edit is a revision to any of the 62 frozen corpus titles. Pivot latency is the time from crossing to the account’s first I/P edit; ECP fast pivot requires that edit to fall at or after crossing and within 14 days of it. Accounts whose first I/P edit predates crossing are topic veterans and are excluded from the pivot-speed comparison.

Frozen parameters: 500 edits, 30-day tenure, 14-day pivot window.

C Upstream validation detail

Secondary analyses supporting §5.1 and §5.2. Primary results and the within-I/P band breakdown for BURST-95 are in the body.

C.1 Tier membership and sanction status

The frozen tier lists hold 995 accounts. Of these, 269 (Tier 1), 111 (Tier 2) and 558 (Tier 3) hold surviving prose in the authorship table and enter downstream measurement.

Sanction status was re-verified against the live en-wiki API on 2026-08-08 for all 938 prose-holding members: no Tier 1 or Tier 2 account was blocked, and 556 of 558 Tier 3 accounts were. The two exceptions are bans enforced without a local block, one WMF/global and one ArbCom. Block-episode counts come from the cohort block-log extract of 2026-08-06, covering events through 2026-07-07: every Tier 2 member shows at least three episodes, and 84 of the 269 prose-holding Tier 1 members show expired block history. The frozen lists do not

Band (lifetime mainspace edits)	50–200	200–1k	1k–5k	5k–10k	10k–50k	50k+
p95 cutoff on <code>burst_30day_share</code>	0.9474	0.8733	0.6244	0.4112	0.3655	0.1879

track subsequent block-log changes.

C.2 BURST-95 cross-cohort comparison (not used as a primary estimand)

Comparing I/P BURST-95 accounts to *all* Random editor sample accounts in the same edit-count band, flagged or not, gives 36.6% against 12.5% ever-blocked, a lift of $2.92\times$ ($p \approx 1.9 \times 10^{-15}$). By band: $3.47\times$ (50–200), $3.14\times$ (200–1k), $1.64\times$ (1k–5k), $1.59\times$ (5k–10k), then attenuating. The contrast mixes burst flagging with topic selection, so §5.1 reports only the within-cohort peer comparison. Machine-readable detail ships with the evidence bundles.

C.3 ECP fast pivot replication

On the Random editor sample the same construction finds only 32 fast pivoters among 3,581 scorable accounts, and the risk-arm lift (currently blocked $\rightarrow$ fast pivot) of $1.51\times$ is not significant ($p = 0.39$). This is expected by design, because the pivot target is the I/P corpus itself, and it marks ECP fast pivot as a **policy-specific** selector: meaningful where an extended-confirmed gate and topic concentration co-occur, and limited to those conditions.

D Method 2 replication results

600 answers: 200 questions $\times$ three models, search forced on, one completion each. **X** and the author-ship table are identical to Method 1 by construction, so the two designs differ only in their questions.

Verbatim paste appears in **39.0%** of Method 2 answers (234 / 600) against Method 1’s 26.5%. Every tier-level conclusion is unchanged: the same direction on every pool, the same ordering, and the same driving selector (ECP fast pivot, with BURST-95 contributing almost nothing). BURST-95 rests on a 1.27% corpus share and nine span-bearing answers, so its $0.22\times$ should be read as no evidence of over-representation, and it does not estimate suppression.

Figures: the figure in Appendix E (`paper/figures/fig1_lift_forest.svg`) supersedes the bundle-local charts for reporting purposes, because it carries the same lifts with their intervals, which the earlier charts predate. The originals remain in the evidence bundles as frozen

artefacts: a corpus-share-against-pasted-share chart for each design, and a point-estimate lift-by-pool chart for the combined set.

E Lift by tier and by selector, both designs

Figure 1. Lift by editor tier and by Tier 1 selector, for both question designs and pooled. The x-axis is logarithmic, so equal proportional departures from the null are equally far from the line whichever way they point. Read across a row’s three arms to see whether the designs agree; read the width of an interval to see what the estimate is worth.

Pool	X (corpus)	Y (span)	Lift	Answers ≥ 1 span
Ever blocked	52.16%	58.52%	1.12$\times$	209 / 600
Tier 1 union (burst $\cup$ ECP)	29.67%	43.08%	1.45$\times$	166 / 600
ECP fast pivot	28.56%	42.80%	1.50 $\times$	164 / 600
BURST-95	1.27%	0.28%	0.22 $\times$	9 / 600
Tier 3: standing sanction	11.46%	9.49%	0.83 $\times$	n/a

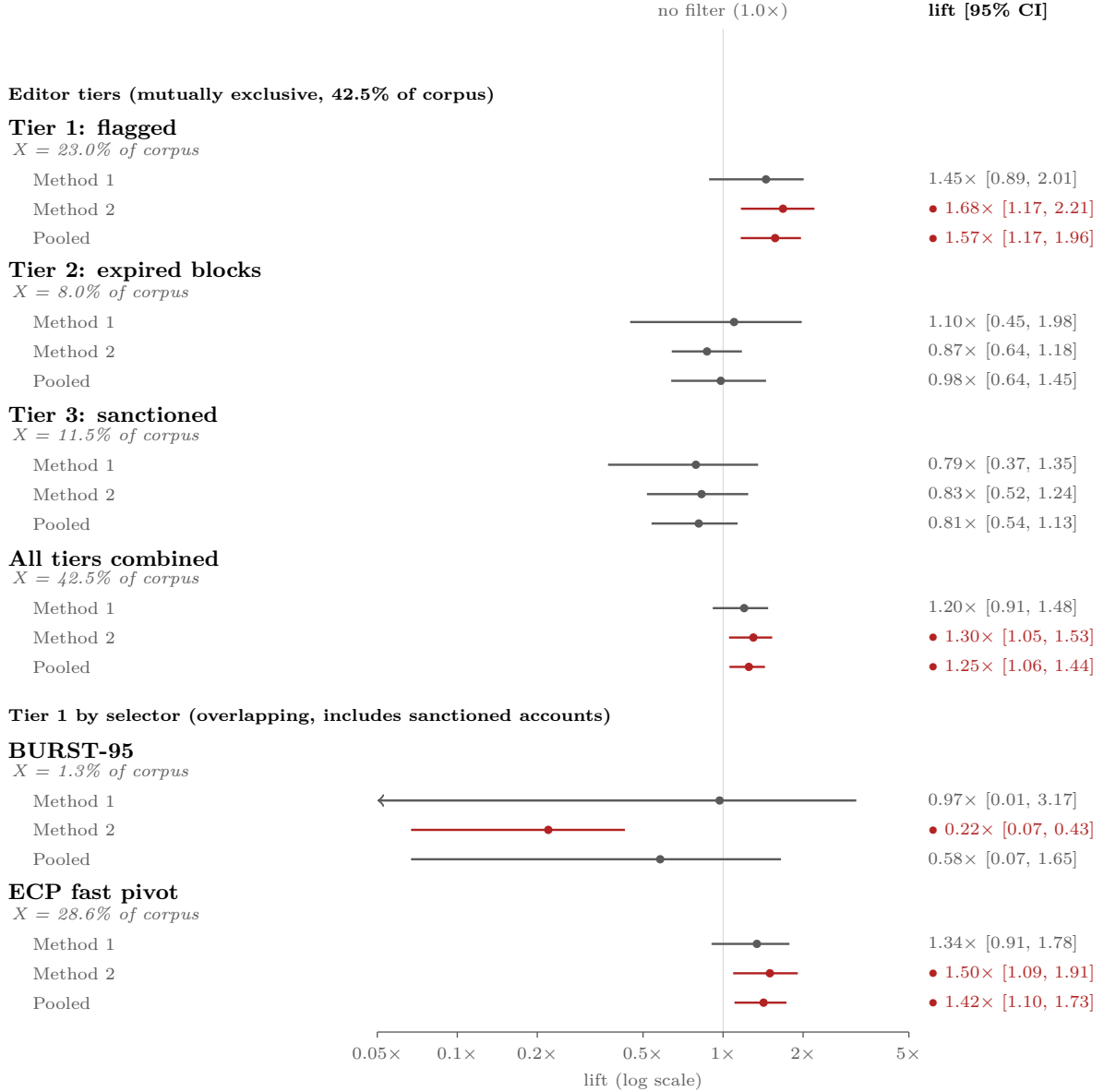


Figure 1: Lift by editor tier and by Tier 1 selector, for both question designs and pooled. Lift is a pool’s share of pasted span tokens (Y) over its share of corpus prose tokens (X); 1.0 is reproduction exactly in proportion to authorship. Intervals are 95% percentile CIs from a clustered bootstrap resampling answers (10,000 resamples, seed 42); ● marks an interval excluding 1.0. An arrowhead marks a bound outside the plotted range (BURST-95, Method 1: 0.005). The two blocks are different labelling systems: a selector row overlaps the tier rows and cannot be added to them.