Visibility Friction in the Hong–Kangxi Discourse on Xiaohongshu

Yidan Xia, Political Science

Project: Xiaohongshu (XHS) Hong–Kangxi posts visibility study.
Prepared for: Imagined stakeholder is the proposal review committee, including researchers familiar with computational text analysis. They will decide whether the project's measurement architecture and inferential framing are sound enough to be developed into fuller thesis analysis. The reader is assumed to be comfortable with regression and intercoder reliability but not with XHS-specific slang or the Hong–Kangxi rumor.

Executive summary

This memo asks whether public-search visibility friction inside one Xiaohongshu (XHS) Hong–Kangxi discourse event is associated with narrative direction, rhetorical mode, carrier material, or platform legibility. The practical problem for the thesis is not whether the platform "deleted" a topic in general, but whether the measurement architecture can support a narrower claim about differential visibility inside a single historical-nationalist discourse. The analytic corpus contains 1,594 unique event posts from a November 2025 keyword crawl. The outcome is `search_nonretrievable`: whether a collected post could still be recovered through public XHS search during the April 2026 verification check. Because the audit shows that this label mixes hard disappearance and search invisibility, the memo treats it as *visibility friction*, not confirmed deletion or censor intent.

The main result is descriptive. After excluding verification errors and mode/content-uncodable posts, the adjusted logit sample is $N = 1{,}463$. The clearest associations are feminist-lineage direction (OR = 2.06, 95% CI [1.22, 3.46], AME = +16.5 percentage points), Red Chamber suoyin carrier (OR = 2.30, 95% CI [1.64, 3.21], AME = +19.2 points), and medium OCR legibility relative to high-caption posts (OR = 1.37, 95% CI [1.05, 1.79], AME = +7.3 points). The OCR result is best interpreted as a format/searchability mechanism, not as evidence that OCR posts were deleted. Engagement is also heavily right-skewed, and very low engagement is associated with higher search non-retrievability; I therefore treat engagement as a visibility mechanism and sensitivity concern rather than a clean pre-treatment control. Anti-Manchu/Qing direction is positive but not statistically stable in the main model; anti-West is too rare to interpret beyond low-power evidence. New engagement and OCR/legibility sensitivity checks show that Red Chamber suoyin remains robust after excluding low-engagement or OCR-dependent posts, while feminist-lineage remains positive but becomes less precise in the strict high-caption-only sample. The thesis should therefore frame the quantitative finding as a within-event visibility association and reserve stronger claims about mechanisms for future verification, audit, and qualitative follow-up.

Decisions for the reader

- Use `search_nonretrievable`, but name it "visibility friction." Do not call it confirmed deletion unless later verification separates search invisibility, author deletion, and platform removal.

- Keep direction as the central theoretical construct, but report the current anti-Manchu/Qing result as unresolved rather than confirmed. The supported direction result is feminist-lineage visibility friction.
- Treat carrier variables as material controls and sensitivity checks. Red Chamber suoyin is empirically strong, but the coefficient should not be converted into a claim that "literary content" is targeted.
- Treat medium OCR legibility as a searchability/observability mechanism. Report OCR-restricted sensitivity checks so that format-driven non-retrievability is not confused with deletion or content moderation.
- Treat engagement as a visibility mechanism and sensitivity issue, not as a clean pre-treatment control. Report models with and without engagement, categorical engagement-bin controls, and low-engagement exclusions.
- For thesis-stage causal or mechanism claims, add repeated verification, stable account identifiers, and a random audit of non-adjudicated posts.

1 Problem, scope, and research question

The Hong–Kangxi rumor claimed that the Kangxi Emperor was the secret son of Empress Dowager Xiaozhuang and Hong Chengchou, a Ming general who surrendered to the Qing. The historical claim is not the object of this memo. The research problem is that the rumor reactivated several politically meaningful ways of narrating Chinese nationhood: anti-Manchu/Qing rupture, anti-Western pseudo-history, and feminist-lineage critique. These directions touch different premises in the state-led national story: the multi-ethnic nation framework (Mullaney, 2011; Brubaker, 2004; Wimmer, 2013), the national-humiliation chronology that centers nineteenth-century foreign aggression (Wang, 2012), Han-centered Ming/Qing boundary narratives (Fitzgerald, 1996; Carrico, 2017), and gendered or patrilineal assumptions about lineage continuity (Wang, 2017).

Chinese information control is known to be strategic, delegated, and frictional rather than reducible to universal deletion (King, Pan, and Roberts, 2013; Lorentzen, 2014; Roberts, 2018; Sun and Zhao, 2022). The missing piece here is intra-domain variation: **within the same Hong–Kangxi event corpus, which post features are associated with reduced public-search visibility?** I answer three operational questions. RQ1 describes the corpus by rhetorical mode, narrative direction, carrier material, and platform legibility. RQ2 tests whether anti-Manchu/Qing, anti-West, and feminist-lineage directions differ in visibility. RQ3 tests whether mode, carrier, or legibility help explain visibility beyond direction.

2 Data, outcome, and measurement design

The raw November 2025 crawl contains 2,797 keyword hits from 康熙瓜, 洪承畴, and 伪史. After removing 107 unrelated keyword-noise posts, the cleaned crawl contains 2,690 rows: 593 standalone Western pseudo-history posts, 503 pure historical-analysis posts, and 1,594 Hong–Kangxi event posts. The 1,594 event posts are the analytic corpus. Each record includes title, caption/body, OCR text where available, post type, keyword source, engagement metadata at collection, posting date, and April 2026 search-verification status.

The dependent variable is `search_nonretrievable`: 1 if verification search does not return the original post and 0 if it does. A random audit of 100 posts flagged likely deleted found 32 gone from

an identifiable profile, 27 still present on an author profile but invisible through public search, and 41 author-unverifiable cases. This is why the outcome is interpreted as public-search visibility friction rather than hard deletion. Appendix D reports the audit and reference-category rates.

Measurement uses a two-pass codebook. Coders read title, description/body, OCR text, and content-bearing hashtags together; they do not infer from unobserved video, comments, author identity, or external knowledge. Pass 1 codes rhetorical mode and evidence fields; Pass 2 codes direction and carrier. Platform legibility is script-derived only after the final reviewed `message_location` field is fixed; it collapses message location, post type, visible text, OCR text, and video-observability flags into a text-observability category. Production labels are source-tracked: 623 high-risk posts use human-adjudicated final fields, while 971 non-adjudicated posts use the LLM layer only where the pre-specified rule allows it. All 623 adjudicated rows merge into the final workbook; the non-adjudicated rows have no mode consensus violations and no direction-positive leakage outside the adjudication queue.

3 Findings

RQ1: Most event participation is not a theory-bearing direction

The event corpus is not dominated by overt political claims. Public claim/correction is the largest mode (609 posts, 38.2%), but playful spectatorship is also large (441, 27.7%). Information/evidence work and inquiry/reflection account for 238 (14.9%) and 225 (14.1%) posts. Direction is rarer than event participation: the final direction dummies identify 91 anti-Manchu/Qing posts (5.7%), 14 anti-West posts (0.9%), and 98 feminist-lineage posts (6.1%). Most posts therefore participate in the Hong–Kangxi event without advancing one of the three theory-bearing directions.

Carrier and legibility explain why a direction-only design would be too coarse. Under the default carrier rule, 878 posts use Hong–Kangxi-specific material, 235 use Red Chamber suoyin material, 218 use popular-culture intertext, and 293 use broader history/politics material; these categories overlap. Legibility is also uneven: 728 posts have high caption legibility, 570 rely on medium OCR legibility, 234 have low visual/video legibility, and 54 are video-centered but still caption-readable. Appendix C reports the distributions, bivariate rates, carrier-source comparisons, and sensitivity plots.

RQ2–RQ3: The adjusted signals are feminist-lineage direction, Red Chamber suoyin carrier, and medium OCR legibility

The main model is a descriptive logit for `search_nonretrievable`. It uses the 1,463 posts with non-error outcome and codable mode/content. It includes three direction dummies, categorical mode, four carrier controls, binary `has_evidence_work`, platform legibility, post type, posting date, keyword fixed effects, log visible-text length, and log engagement at collection. Standard errors are robust. Engagement is shown both with and without controls because engagement may partly reflect earlier visibility or ranking rather than a pre-treatment covariate; Appendix 28 adds bin controls and low-engagement exclusions for this reason.

The model answers RQ2 cautiously. Feminist-lineage direction is the only theory-bearing direction that remains clearly associated with higher public-search non-retrievability in the main model and in the with/without-engagement comparison. Anti-Manchu/Qing direction is positive but not statistically distinguished from zero in the main model. Its estimate also moves across LLM-only direction variants, so it should be treated as theoretically important but empirically unresolved. Anti-West direction has only 14 positives in the full corpus and 13 with non-missing outcomes; its wide interval is low power,

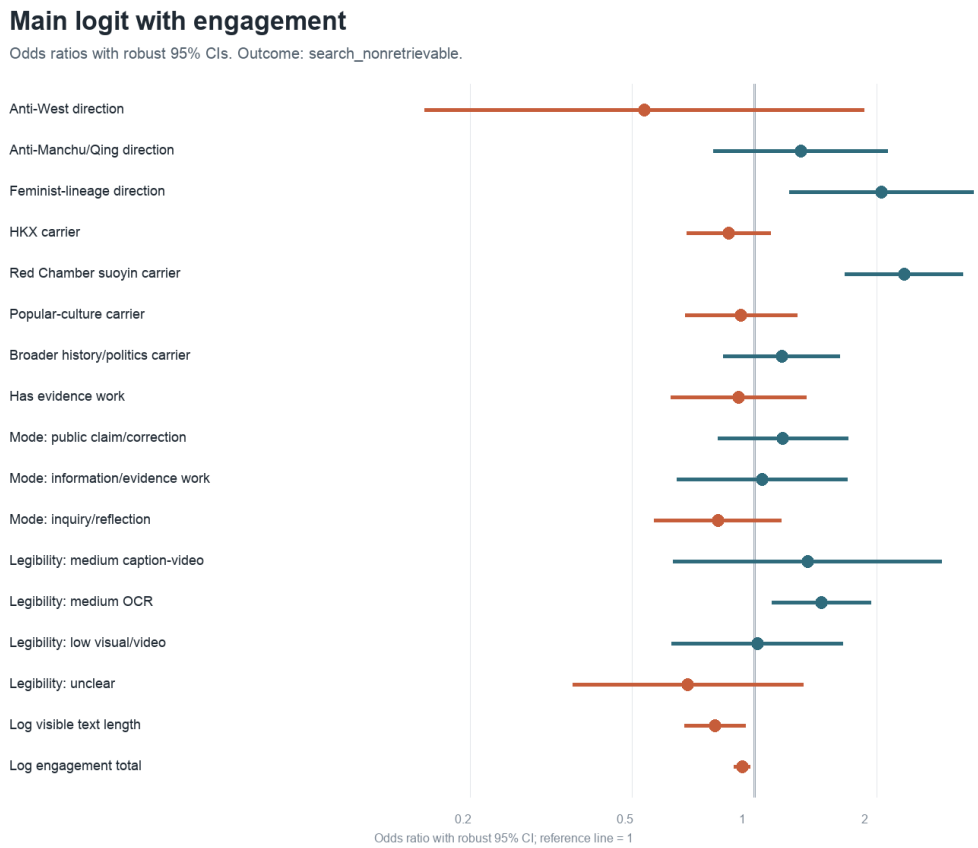

Figure 1: Key adjusted associations from the main logit with engagement controls. Points are odds ratios; bars are robust 95% confidence intervals; the vertical reference line is OR = 1. Values above 1 indicate higher odds of search non-retrievability. The full coefficient plot appears in Appendix C.

not evidence of safety or absence.

For RQ3, Red Chamber suoyin carrier is the strongest non-direction association. It remains large under the GPT-default, DeepSeek, carrier-union, and carrier-intersection variants. This does not prove that Red Chamber content is moderated as literature; it shows that posts using Red Chamber as interpretive material are more often search-nonretrievable after adjusting for direction, mode, legibility, post type, keyword, date, text length, and engagement. Medium OCR legibility is also positive, though weaker after deriving legibility from the final reviewed message-location field. This points to an observability pathway: image-text or OCR-mediated posts may be indexed, verified, or searched differently from caption-primary posts. Appendix 28 shows that the core feminist-lineage and Red Chamber estimates remain positive when low-engagement posts are excluded, while the medium-OCR estimate weakens. Appendix 29 adds OCR-restricted checks. Dropping medium-OCR posts preserves the Red Chamber estimate and keeps feminist-lineage positive; the stricter high-caption-only model preserves Red Chamber but makes the feminist-lineage estimate less precise.

Bottom line for thesis design

The quantitative evidence supports a narrow claim: within this event corpus, public-search visibility friction is associated with feminist-lineage direction, Red Chamber suoyin carrier, and medium OCR legibility. The three signals likely reflect different mechanisms. Medium OCR may partly be

Table 1: Key regression results, with and without engagement controls

Term	Main model with engagement: OR (95% CI)	AME	Without engagement: OR (95% CI)
Anti-West direction	0.53 [0.15, 1.81]	−0.146	0.53 [0.16, 1.73]
Anti-Manchu/Qing direction	1.27 [0.77, 2.09]	+0.057	1.29 [0.80, 2.11]
Feminist-lineage direction	2.06 [1.22, 3.46]	+0.165	1.90 [1.14, 3.16]
Hong–Kangxi carrier	0.86 [0.68, 1.09]	−0.036	0.85 [0.67, 1.08]
Red Chamber suoyin carrier	2.30 [1.64, 3.21]	+0.192	2.24 [1.61, 3.12]
Medium OCR legibility	1.37 [1.05, 1.79]	+0.073	1.33 [1.02, 1.74]
Log engagement total	0.93 [0.89, 0.97]	–	not included

Note: AME is the average discrete change reported by the notebook for binary predictors in the with-engagement model. The rare `tag_only_low_legibility` category is not interpreted because only two non-error posts fall in that cell.

search-procedure friction and low-exposure searchability; Red Chamber suoyin is robust even when low-engagement or OCR-dependent posts are removed; feminist-lineage remains positive but is more format-sensitive. The analysis does not yet support a causal claim about platform intent, nor a clean claim that anti-Manchu/Qing direction is selectively suppressed. The next stage should therefore combine repeated verification and account-level audit with qualitative close reading of the feminist-lineage and Red Chamber cases.

A Codebook development and fixed-instrument logic

The initial proposal asked whether anti-Manchu, anti-Western, and feminist posts had different deletion or survival rates on XHS. That design rested on three assumptions that data exploration falsified: that the three categories were common enough for a single-axis analysis, that they were cleanly separable from gossip and meme content, and that public-search non-retrievability approximated deletion. An earlier codebook lineage (v1 through v2-Alpha.2) attempted to operationalize the original design through a single mode-plus-topic architecture and was tested in smoke and reliability exercises; it ultimately failed because the topic field was being asked to carry direction, carrier, and context simultaneously. Therefore that lineage was abandoned. The replacement design changed (i) the outcome (deletion → visibility friction), (ii) the variable structure (single mode-plus-topic → direction + mode + carrier + legibility, two-pass), and (iii) the measurement strategy (classifier-first → reliability-first human/LLM-assisted coding with empty-state-valid labels). The theoretical core that the Hong–Kangxi event reopens load-bearing premises of state-led nationhood is preserved; the design now matches what the data can sustain.

The current codebook (v3.3-final, augmented) is the second of two distinct codebook lineages developed for this project. The first lineage (**v1 through v2-Alpha.2**) was abandoned after a reliability test showed that its construct definitions did not survive contact with the data. The second lineage (**v3 through v3.3-final, augmented**) was developed from scratch after that abandonment and is the instrument used in production. This appendix documents both lineages because the abandonment is itself part of the project's methodological commitment to validation-first measurement.

A.1 Abandoned first lineage (v1–v2-Alpha.2)

The first lineage used a *mode* (M1–M7) plus *topic* (T1–T7) plus *scope* plus *stance* architecture. Mode and topic were initially single-pass coded together, and topic was a single multilabel field that mixed narrative direction (T1 ming-Qing rupture, T2 anti-Western pseudo-history, T4 patriarchy-feminism) with carrier material (T3 Hong-Kangxi genealogy, T5 Red Chamber suoyin) and residual context (T6 general history-politics, T7 textbook historiography). The lineage began with a 112-post exploratory sample and side-by-side LLM description under v1/v1+, moved to v2 after adding `scope`, and then pivoted to v2-Alpha when clarifying that tone/mode and topic should be classified separately. The first 40-post smoke-test attempt was discarded because the codebook changed and API/rate-limit problems prevented a clean LLM run. A second 40-post smoke test under the revised prompt passed the pre-committed smoke criterion; a subsequent two-pass architecture split mode from topics/stance, and v2-Alpha.2 sharpened T3 to require explicit textual reference rather than tag-only or vague evocation.

The decisive old-lineage test was a 96-post reliability sample, with primary reported content analyses on $n = 89$ after excluding unrelated/noisy rows and recording a small number of technical issues. Mode reached human–LLM $\kappa = 0.51$, T1 reached 0.49, T5 reached 0.92, but **T3 reached only** $\kappa = 0.38$. Diagnosis traced the divergence to human coder drift rather than a simple prompt failure: the primary coder had implicitly treated event relevance as requiring a topic label and placed vague event-evocation posts into T3, while the LLMs read v2-Alpha.2 literally and left topics empty when no Hong–Kangxi paternity material was textually present. The diagnosis surfaced a deeper architectural problem: the topic field was being asked to do three different jobs at once (direction, carrier, context). The lineage was abandoned rather than patched again.

A.2 Replacement second lineage (v3–v3.3-final, augmented)

The replacement lineage was built around four design constraints derived from the v2-Alpha.2 reliability post-mortem: (i) separate rhetorical action from theory-bearing content from cultural carrier (the two-pass architecture), (ii) treat empty-state for both direction and carrier as a valid label rather than a missing label, (iii) reduce the number of substantive mode categories from six to four to lower category-boundary load on coders, and (iv) derive style-flag fields (`playful_shell_lexical`, Thin event reference) by deterministic script rather than by subjective coding. These constraints produced v3 notebook.

The replacement lineage used three *calibration* rounds (henceforth "round 1", "round 2", "round 3") to refine boundary rules on disjoint development samples drawn from the analytic corpus and not used elsewhere in the project. Round 1 ($n = 57$) tested v3 with two human coders and three LLMs; H1–H2 mode reached $\kappa \approx 0.56$, but H1–H2 multilabel direction reached macro $\kappa \approx 0.33$ and `direction_unclassifiable` versus `direction=[]` confusion collapsed multilabel direction's effective n. v3 was simplified to v3.1 (six raw modes compressed to four substantive modes plus `unable_to_classify`; `direction_unclassifiable` removed; `direction=[]` formalized as a valid state) and to v3.2 (added `evidence_work` and `claim_target` as four-way secondary qualifiers; reframed Thin event reference as script-derived). Round 2 ($n = 76$) tested v3.2 with two human coders and three LLMs; H1–H2 mode reached $\kappa \approx 0.67$, direction_status $\kappa \approx 0.64$, and carrier macro $\kappa \approx 0.65$, but `playful_shell` as a human-coded field reached $\kappa \approx 0.06$, claim_target reached $\kappa \approx 0.37$, and the LLMs systematically over-called direction and carrier (H1: 6 direction positives vs LLMs: 12–24; H1: 32 carrier positives vs LLMs: 47–68). v3.2 was patched to v3.2.2 (added explicit positive examples for each direction; tightened the rule that Hong–Kangxi paternity material alone is not anti-Manchu; added `broader_history_politics_material` as a fourth carrier and removed the residual-context bucket; required that `playful_shell` be derived lexically rather than coded). Round 3 ($n = 49$ targeted hard cases) tested v3.3; H1–H2 carrier macro κ rose from 0.65 to 0.72 and claim_target rose from 0.37 to 0.62. v3.3 was given micro-clarifications (all judgments combine title, body, OCR, and content-bearing hashtags; conditionals that finalize a claim are public claim, conditionals that remain open hypotheticals are inquiry; finalizing cues such as 怪不得/果然/事实证明 must be context-dependent; Red Chamber as referenced object does not trigger Red Chamber suoyin carrier) and frozen as v3.3-final *before* the held-out reliability sample was coded.

The v3 series did not use a robust LLM set during calibration; only the held-out reliability test added Claude Opus 4.7 and DeepSeek V4 Pro Thinking. The v3-series total development sample size is therefore $n = 57 + 76 + 49 = 182$ posts, none of which appears in the 240-post reliability sample.

A.3 Post-reliability augmentation

Following the held-out reliability test, 119 of the 232 analyzable posts were selected for adjudication because they contained human–human or human–LLM conflicts on substantive fields. Adjudication produced (i) final field-by-field labels for the disputed cases, used as gold labels for these posts in the production dataset; and (ii) a set of worked boundary-case examples appended to the codebook as illustrations under existing rules, "v3.3-final, augmented." Examples include cases such as: when 陈道明 (an actor's name) appears as a pun crossing from popular-culture carrier to playful-only mode; when 玉牒/族谱 cues serve only Hong–Kangxi paternity rather than broader historiographical work; when 怪不得 X crosses from playful affect to public-claim correction; when an open hypothetical conditional remains inquiry rather than crossing to public claim. The codebook constructs (mode,

direction, carrier, legibility) and the legal label sets are unchanged.

The post-reliability step follows the general best-practice logic that LLM annotation requires a stable codebook, explicit validation, and caution against changing the task after validation (Törnberg, 2024; Fang et al., n.d.). I therefore distinguish two project-specific interventions. *Definitional refinement* would change constructs or legal label sets and would require a new held-out validation. *Worked-example augmentation* clarifies the application of unchanged rules without changing constructs. The current step is the latter; the held-out reliability metrics in Appendix B remain valid estimates of the locked v3.3-final instrument, while the augmented examples clarify production application without changing constructs.

A small number of these worked examples (6–10, selected for representative coverage of the most common boundary types) are included in production prompts as few-shot examples; the remainder are kept in the codebook as reference but not in prompts, to keep prompt length tractable.

B Held-out reliability test for the v3.3-final codebook

This appendix is separate from Appendix A because it reports the performance of a locked measurement instrument, not another round of codebook revision. Appendix A explains how the instrument was developed; this appendix estimates whether trained humans and LLMs can apply the fixed instrument reproducibly.

B.1 Purpose and design

This appendix reports the held-out reliability test for the fixed v3.3-final codebook. The codebook was fixed before this test. The test estimates the quality of the measurement instrument, and would not be used to revise the construct definitions. The empirical question is whether, within one Hong–Kangxi / Mourning Ming discourse event, visibility friction tracks narrative direction, rhetorical mode, carrier material, or platform legibility. The three theory-bearing directions are `anti_west_direction`, `anti_manchu_or_anti_qing_direction`, and `feminist_lineage_direction`.

The coding architecture separates rhetorical action from theory-bearing content. Pass 1 classifies `mode`, `evidence_work`, and `claim_target`. Pass 2 classifies `direction` and `carrier`. Coders read all visible material together: title, description/body, OCR text, and content-bearing hashtags. They do not infer from unobserved video content, comments, author identity, or external knowledge. `direction` and `carrier` are separate: carrier material does not by itself imply direction. This separation matters because many posts use Hong–Kangxi, Red Chamber, popular-culture, or broader historical materials without advancing one of the three theory-bearing narrative directions.

The primary human coder (H1) is treated as the reference for human–LLM comparison because H1 is the production coder. The second human coder (H2) provides the local human–human benchmark. The H1–H2 benchmark is not a metaphysical truth label; it is an estimate of how reproducible the codebook is when applied by trained human coders. LLM performance is therefore interpreted against this human benchmark, not against perfect accuracy.

Table 2: Input sets in the updated reliability run

LLM set	Mode rows	Content rows	Combined rows
current	720	720	720
robust	480	480	480

Note: The updated run reads 232 primary human rows, two LLM sets, and 1,200 combined LLM rows. The current set contains three models: Claude Sonnet 4.6, DeepSeek V3.2, and GPT-5.4. The robust set contains Claude Opus 4.7 and DeepSeek V4 Pro Thinking.

B.2 Metrics and interpretation

For single-label and binary fields, this appendix reports accuracy, Cohen's κ, macro F1, and weighted F1. For multilabel fields, it reports the average binary κ across labels, macro positive F1, macro precision, and macro recall. For all-rater comparisons, it reports Krippendorff's α. Because several labels are rare, accuracy can be misleading: a model can obtain high accuracy by predicting the majority class. The analysis therefore emphasizes κ, macro F1, precision, recall, and per-label error structure. Rigid universal thresholds are avoided. The H1–H2 benchmark is the local ceiling for this task: a model is strong if it approaches the human–human benchmark on the same field and does not show a systematic error pattern that threatens the substantive analysis. This is especially important

for `anti_manchu_or_anti_qing_direction`, because false positives or false negatives in that label can directly change the main visibility-friction results.

B.3 Human–human benchmark

B.3.1 Single-label and binary fields

Table 3: H1–H2 agreement for single-label and binary fields

Field	n	Accuracy	Cohen's κ	Macro F1	Weighted F1
`mode`	228	0.776	0.677	0.712	0.782
`evidence_work`	218	0.812	0.685	0.561	0.886
`claim_target`	218	0.706	0.477	0.475	0.701
`direction_status`	228	0.890	0.737	0.735	0.895
`mode_codable`	228	0.969	0.448	0.723	0.975
`content_codable`	228	1	–	1	1
`has_evidence_work`	228	0.855	0.713	0.854	0.854
`playful_shell_lexical`	228	0.961	0.903	0.951	0.960
Thin event reference	228	0.969	0.616	0.808	0.972

The human–human benchmark is strongest for `direction_status` and for the script-derived `playful_shell_lexical` variable. `mode` is usable and close to the expected range for latent rhetorical categories in short social-media text. The detailed `evidence_work` field has a reasonable κ, but its macro F1 is low because rare categories, especially `mixed_or_unclear`, are difficult. The binary `has_evidence_work` field is much more stable. `claim_target` is the weakest substantive single-label field and should be treated as secondary.

The main H1–H2 mode confusion is between public claim/correction and information/evidence work: 14 H1 public-claim posts were coded by H2 as evidence work, and 11 H1 evidence-work posts were coded by H2 as public claims. This is a boundary between finalizing a public-facing claim and arranging evidence or explanation. The main direction-status confusion is the boundary between `no_theory_direction` and `anti_manchu_or_anti_qing_direction`: H2 coded 9 H1 no-theory posts as anti-Manchu/Qing and 6 H1 anti-Manchu/Qing posts as no-theory.

B.3.2 Multilabel fields

Table 4: H1–H2 agreement for multilabel fields

Field	n	Macro κ	Macro positive F1	Precision	Recall	H1 positives	H2 positives
`direction`	228	0.762	0.783	0.741	0.830	63	70
`carrier`	228	0.625	0.736	0.778	0.711	248	212

The direction field is the strongest main content field. Carrier is usable but less stable. The lower carrier agreement reflects a real conceptual burden: coders must decide whether a post merely references the event or actually uses one of the specified carrier materials.

Table 5: H1–H2 per-label agreement for direction and carrier

Label	H1+	H2+	TP	FP	FN	κ	F1	P/R
AW direction	7	8	5	3	2	0.655	0.667	0.625/0.714
AM/Q direction	31	34	24	10	7	0.695	0.739	0.706/0.774
Feminist direction	25	28	25	3	0	0.936	0.943	0.893/1.000
HKX carrier	143	102	90	12	53	0.445	0.735	0.882/0.629
Red Chamber carrier	47	41	37	4	10	0.803	0.841	0.902/0.787
Pop culture carrier	13	12	9	3	4	0.704	0.720	0.750/0.692
History/politics carrier	45	57	33	24	12	0.547	0.647	0.579/0.733

Note: AW = anti-West; AM/Q = anti-Manchu/Qing; HKX = Hong–Kangxi-specific material. P/R reports precision/recall treating H1 as reference. These statistics use the reliability sample; the 80-post boundary supplement is not a prevalence estimate.

Three patterns matter. First, `feminist_lineage_direction` is highly reproducible: F1 = 0.943 and $\kappa = 0.936$. Second, `anti_manchu_or_anti_qing_direction` is usable but sensitive: F1 = 0.739 and $\kappa = 0.695$. Third, `hongkangxi_specific_material` has a large threshold disagreement: H1 coded 143 positives and H2 coded 102. H2 is stricter relative to H1. This does not make the field unusable, but it means HKX carrier must be interpreted with care, especially when a post contains only a generic Hong–Kangxi or Kangxi-rumor reference.

B.4 Human coder versus current LLMs

The current LLM set contains Claude Sonnet 4.6, DeepSeek V3.2, and GPT-5.4. After completion fixes, all three models have full mode rows and full content rows.

B.4.1 Single-label and binary fields

Table 6: H1 vs current LLMs for single-label and binary fields: Cohen's κ with accuracy in parentheses

Field	H1–H2	Claude Sonnet 4.6	DeepSeek V3.2	GPT-5.4
`mode`	0.677	0.649 (0.763)	0.593 (0.720)	0.664 (0.776)
`evidence_work`	0.685	0.438 (0.807)	0.660 (0.809)	0.543 (0.820)
`claim_target`	0.477	0.529 (0.731)	0.568 (0.756)	0.577 (0.757)
`direction_status`	0.737	0.538 (0.776)	0.552 (0.784)	0.585 (0.815)
`has_evidence_work`	0.713	0.767 (0.884)	0.716 (0.858)	0.785 (0.892)
`playful_shell_lexical`	0.903	0.905 (0.961)	0.874 (0.948)	0.927 (0.970)
Thin event reference	0.616	0.000 (0.966)	0.351 (0.970)	0.656 (0.978)

Note: The best current model is clear by row: GPT-5.4 for most fields and DeepSeek V3.2 for detailed `evidence_work`. The detailed field is not used as a main endpoint despite its kappa because rare subcategories remain unstable.

For `mode`, GPT-5.4 is the best current model ($\kappa = 0.664$), closely followed by Claude Sonnet 4.6 ($\kappa = 0.649$). Both are close to the H1–H2 benchmark ($\kappa = 0.677$). DeepSeek V3.2 is lower ($\kappa = 0.593$). Current LLMs can be used for mode pre-coding, but mode disagreement cases still need human review if the field enters a main model.

For `direction_status`, all current LLMs remain below the human benchmark. GPT-5.4 has the highest current κ (0.585), followed by DeepSeek V3.2 (0.552) and Claude Sonnet 4.6 (0.538). This field should not be accepted directly from current LLMs for final substantive analysis.

For `has_evidence_work`, the current LLMs perform very well: GPT-5.4 at $\kappa = 0.785$, Claude Sonnet 4.6 at 0.767, and DeepSeek V3.2 at 0.716. The detailed `evidence_work` categories remain less useful as main endpoints, but the binary indicator is strong. For `claim_target`, the LLMs agree with H1 more than H2 does, but H1–H2 agreement is weak and the macro F1 values for the LLMs remain around 0.50; the field is better retained as a secondary or audit variable.

B.4.2 Multilabel direction and carrier fields

Table 7: H1 vs LLMs for multilabel direction and carrier fields

Set	Field	Model	n	κ	F1	P	R	H1+	M+
current	direction	Claude Sonnet 4.6	222	0.695	0.732	0.675	0.810	65	85
current	direction	DeepSeek V3.2	221	0.668	0.704	0.678	0.741	64	76
current	direction	GPT-5.4	229	0.671	0.708	0.689	0.737	65	77
current	carrier	Claude Sonnet 4.6	222	0.492	0.664	0.548	0.933	253	390
current	carrier	DeepSeek V3.2	221	0.458	0.650	0.582	0.792	251	319
current	carrier	GPT-5.4	229	0.562	0.690	0.649	0.803	254	269
robust	direction	Claude Opus 4.7	228	0.712	0.746	0.685	0.830	65	86
robust	direction	DeepSeek V4 Pro	229	0.714	0.740	0.803	0.690	64	56
robust	carrier	Claude Opus 4.7	228	0.568	0.728	0.641	0.880	254	332
robust	carrier	DeepSeek V4 Pro	229	0.538	0.663	0.652	0.713	252	254

Note: H1+ and M+ are total positive labels across multilabel slots, not post counts. The table is portrait-oriented to keep surrounding interpretive text on the same page.

For the direction multilabel field, current Claude is the best current model ($\kappa = 0.695$, macro positive F1 = 0.732). DeepSeek V3.2 and GPT-5.4 are close but lower. All three are below the H1–H2 direction benchmark ($\kappa = 0.762$, macro positive F1 = 0.783). Direction is therefore suitable for LLM screening, but not for unadjudicated final labels.

For carrier, GPT-5.4 is the best current model ($\kappa = 0.562$, macro positive F1 = 0.690), but it is still below the H1–H2 carrier benchmark ($\kappa = 0.625$, macro positive F1 = 0.736). Claude Sonnet 4.6 has very high recall but low precision for carrier: it predicts 390 carrier positives against 253 H1 positives in the comparable rows, which means it over-calls carrier materials. DeepSeek V3.2 also over-calls some carrier labels and underperforms GPT-5.4 on macro κ.

B.5 Robust LLM set and current–robust comparison

Table 8: Best current vs best robust LLM, with H1–H2 ceiling

Field	Type	H1–H2	Best current / robust	Robust-current
`mode`	single	0.677	GPT-5.4 0.664 / DS V4 0.682	0.018
`direction_status`	single	0.737	GPT-5.4 0.585 / DS V4 0.647	0.062
`has_evidence_work`	single	0.713	GPT-5.4 0.785 / Opus 0.766	−0.019
`evidence_work`	single	0.685	DS V3.2 0.660 / DS V4 0.452	−0.208
`claim_target`	single	0.477	GPT-5.4 0.577 / DS V4 0.586	0.009
`playful_shell_lexical`	single	0.903	GPT-5.4 0.927 / Opus 0.885	−0.042
Thin event reference	single	0.616	GPT-5.4 0.656 / DS V4 0.741	0.086
`direction`	multilabel	0.762	Sonnet 0.695 / DS V4 0.714	0.019
`carrier`	multilabel	0.625	GPT-5.4 0.562 / Opus 0.568	0.006

Note: DS = DeepSeek. Values are the reliability statistic reported for that field; rows compare the best model within each generation rather than a single universal winner.

The robust set is better for some central fields. It improves `mode` through DeepSeek V4 Pro, improves `direction_status`, improves multilabel `direction`, and slightly improves multilabel `carrier` through Claude Opus 4.7. But the robust set is not uniformly better. It is worse for detailed `evidence_work`, worse for `has_evidence_work` relative to current GPT-5.4, and worse for `playful_shell_lexical` relative to current GPT-5.4. It also does not solve the anti-Manchu/Qing boundary problem: Claude Opus improves recall but still over-calls; DeepSeek V4 Pro is more precise but misses many H1 positives.

The conclusion is field-specific. The second LLM set is better for direction-level robustness and for some mode/codability tasks, but it is not a general replacement for the current set. Stronger models do not automatically produce better social-science measurements. Model choice should be field-specific and benchmarked against human agreement.

B.6 All-rater agreement

Table 9: Krippendorff's α across human and LLM raters

Field	H1–H2 only	H1 + current LLMs	H1 + robust LLMs
`mode`	0.677	0.679	0.611
`evidence_work`	0.685	0.579	0.474
`has_evidence_work`	0.709	0.766	0.707
`claim_target`	0.479	0.640	0.528
`direction_status`	0.737	0.621	0.636
`playful_shell_lexical`	0.903	0.902	0.872
Thin event reference	0.616	0.321	0.430
`mode_codable`	0.447	0.278	0.332
`content_codable`	–	0.259	0.135

Note: The current alpha uses H1 plus three current LLMs ($n = 232$, four raters). The robust alpha uses H1 plus two robust LLMs ($n = 232$, three raters). Low alpha for codability fields reflects extreme class imbalance and should not be interpreted as a substantive failure of the construct.

The all-rater α table supports the same conclusion. Current LLMs produce stronger all-rater agreement for `mode`, `has_evidence_work`, `claim_target`, and `playful_shell_lexical`. Robust LLMs produce stronger all-rater agreement for `direction_status` and Thin event reference. Neither set dominates.

B.7 Main error patterns and implications

Table 10: Qualitative interpretation of the main reliability patterns

Pattern	Evidence from the reliability test	Implication for production coding
Anti-Manchu/Qing boundary is the main risk	H1–H2 F1 is 0.739, but current LLM F1 ranges only from 0.533 to 0.568. Current models over-call anti-Manchu/Qing; robust DeepSeek V4 Pro becomes more conservative but misses more positives.	Use LLMs as screeners only. All positive anti-Manchu/Qing labels and all model-disagreement cases should receive human QA or adjudication.
Feminist lineage is stable	H1–H2 F1 is 0.943. Current LLM F1 is around 0.920–0.923, and robust DeepSeek V4 Pro reaches 0.980.	This label can be more heavily automated, with spot checks on positives and short playful posts.
Red Chamber carrier is stable	H1–H2 F1 is 0.841. Current DeepSeek/GPT and robust Opus are close to or above this level.	This is the safest carrier label for LLM-assisted production.
HKX carrier has a human threshold problem	H1 codes 143 positives while H2 codes 102. LLM support varies widely: current GPT predicts 118 positives; robust Opus predicts 189.	Do not interpret HKX carrier prevalence naively. Audit generic Kangxi/Hong–Kangxi references, father-line-only claims, and thin event references.
Popular-culture carrier is over-called by LLMs	H1–H2 F1 is 0.720. LLM F1 ranges from 0.433 to 0.553. Most errors are false positives.	Use human review for positive popular-culture labels, especially actor-name jokes, CP puns, entertainment references, and meme-like titles.
Broader history/politics carrier is a boundary bucket	H1–H2 F1 is 0.647. Robust Opus approaches this level; other models are weaker or over-call.	Retain as a main carrier with QA. Avoid letting generic "history," "experts," "records," or "official history/unofficial history" cues automatically trigger the label when they only serve HKX paternity.
Detailed evidence-work categories are not main endpoints	H1–H2 macro F1 is 0.561, and robust LLMs do not improve the field. The binary `has_evidence_work` is much stronger.	Use `has_evidence_work` in quantitative models. Keep detailed `evidence_work` for audit or qualitative interpretation.
`claim_target` remains secondary	H1–H2 κ is 0.477. LLMs can agree with H1 more closely, but macro F1 remains modest.	Use `claim_target` as a secondary robustness or descriptive field, not as a main theory-bearing endpoint.

B.8 Use of labels in the full corpus

The reliability test supports a mixed human–LLM strategy. It does not support fully automated final coding for all fields.

Table 11: Production strategy by field

Field	Use	Reason
mode	LLM-assisted coding with disagreement QA.	Best LLMs reach the H1–H2 benchmark.
direction	High-recall LLM screening plus human adjudication for positives and disagreements.	Direction is theoretically central and still below human ceiling.
direction status	Use adjudicated or QAed labels for main analysis.	Best robust LLM improves performance but remains below H1–H2.
anti-Manchu/Qing	Human QA for all model-positive cases; audit likely negatives when feasible.	Systematic false positives and false negatives remain.
feminist lineage	Human adjudication for any model-positive case; otherwise mostly automated with spot checks.	High agreement, but it remains a theory-bearing direction.
anti-West	Human QA for positives and candidate retrieval.	Low support makes metrics unstable.
carrier	LLM pre-code plus human QA for positives and model disagreements.	Carrier is usable but below direction in reliability.
Red Chamber carrier	Semi-automated with spot checks.	Strong human and LLM performance.
HKX carrier	QA threshold cases.	Human coders differ in strictness.
pop-culture carrier	Human review of positives.	LLMs over-call.
history/politics carrier	Positive labels require QA.	Boundary with HKX-service evidence is difficult.
has evidence work	Use as binary quantitative field.	Strong H1–H2 and LLM performance.
evidence-work detail	Audit and qualitative use only.	Macro F1 and category imbalance make it weak as an endpoint.
claim target	Secondary or robustness field.	Human benchmark is weak.
playful shell	Script-derived and spot-checked.	High agreement; deterministic derivation is preferable.
thin event reference	Script-derived or spot-checked.	Rare label; prevalence-sensitive metrics.

B.9 Methodological rationale for staged calibration and a single locked reliability test

The staged calibration tests and the final reliability test follow standard measurement logic from content analysis and survey methodology. The project begins with theoretical constructs—direction,

mode, carrier, and legibility—and operationalizes them as coding rules. In survey-methodology terms, this is a move from abstract constructs to observable indicators, and the main risk is measurement error: the coded answer may deviate from the construct the project intends to measure (Groves et al., 2009). In content analysis, this risk is handled through a codebook, coder training, pilot coding, intercoder reliability tests, and final reliability reporting (Krippendorff, 2018).

The calibration tests are development instruments. Their purpose is not to produce final reliability estimates. Their purpose is to reveal whether the codebook can be applied to real data, whether categories are mutually intelligible, whether coders confuse topic with rhetorical mode, and whether LLM prompts reproduce the intended distinctions. This is why the early tests could be used to revise the codebook, move from single-pass coding to a two-pass architecture, remove unstable fields, and convert `playful_shell` into a script-derived variable. In social-science measurement, this is analogous to questionnaire pretesting or pilot coding: the point is to reduce preventable ambiguity before the final measurement instrument is locked.

The final reliability test has a different status. Once the v3.3-final codebook is fixed, the held-out test estimates reproducibility and model performance. It is not used to change the codebook unless the project is willing to treat this sample as a new development sample and draw a new held-out validation sample. This separation between development data and validation data is especially important for LLM annotation. Recent work on LLMs as annotators stresses that LLM performance is task-specific, prompt-sensitive, and potentially biased; LLM outputs must therefore be validated against a human benchmark, with prompt/codebook details documented and with error analysis by label rather than by one overall accuracy number (Törnberg, 2024; Ziems et al., 2024). Automated text-as-data methods more generally require problem-specific validation and cannot replace careful construct definition and close reading (Grimmer and Stewart, 2013).

The design also treats LLM labels as predicted measurements rather than truth. This matters because even high annotation accuracy can bias downstream statistical estimates if prediction errors are systematic. Recent work on LLM annotations for social science emphasizes combining machine labels with expert annotations and accounting for prediction error when the labels enter downstream inference (Egami et al., 2024; Fang et al., n.d.). The use of two LLM sets in this project therefore serves two purposes: it tests whether a stronger or different model family improves measurement, and it reveals whether errors are stable across models. The robust set improves some fields, especially `direction` and `direction_status`, but it does not remove the need for human QA on theory-bearing direction labels.

The substantive coding architecture is also theory-driven. The project studies visibility friction within a single discourse event rather than broad topic deletion. This design draws on work showing that authoritarian information control can operate through selective censorship, delegated platform enforcement, and friction rather than only through direct deletion (King, Pan, and Roberts, 2013; Roberts, 2018; Sun and Zhao, 2022). The direction variables reflect theories of nationalism, ethnic boundary making, historical memory, and identity politics (Boym, 2001; Brubaker, 2004; Wimmer, 2013). The distinction between direction, carrier, and mode prevents the analysis from treating every Hong–Kangxi reference as the same kind of political content. This is essential because the discourse includes serious claims, evidence work, playful spectatorship, inquiry, Red Chamber suoyin, popular-culture intertext, and broader historical-political frames.

B.10 Bottom-line conclusion

The reliability test supports the fixed v3.3-final codebook for the main constructs, with field-specific qualifications. Human–human agreement is strong for `direction` and `direction_status`, acceptable for `mode`, and moderate for `carrier`. The binary evidence-work indicator is reliable, but detailed evidence-work categories should remain audit fields. LLMs are useful as assisted coders and high-recall screeners. They are not validated substitutes for human adjudication on all theory-bearing labels.

The robust LLMs are partly better but not uniformly better. DeepSeek V4 Pro is strongest for `mode`, `direction_status`, and Thin event reference. Claude Opus 4.7 is strongest for robust carrier coding and performs well as a high-recall direction screener. Current GPT-5.4 remains strong for `mode`, `has_evidence_work`, `playful_shell_lexical`, and current carrier coding. The main unresolved problem is anti-Manchu/Qing direction, while anti-West is too rare for stable automated performance. All three theory-bearing direction positives therefore receive human adjudication before entering the main visibility-friction analysis.

C Production variables, final corpus checks, and regression robustness

This appendix reports the production workbook and notebook outputs behind the within-event model. It keeps the main memo focused on the result while showing the source-tracked labels, corpus checks, carrier comparisons, and sensitivity models.

C.1 Corpus checks and final-label provenance

The final analysis workbook contains 1,594 unique Hong–Kangxi event posts. All 623 human-adjudicated rows are merged into the final analysis data; the remaining 971 non-adjudicated rows use LLM-layer fields when allowed by the pre-specified production rule. Non-adjudicated posts have no mode consensus violations and no direction-positive leakage outside the adjudication queue.

Table 12: Final analysis workbook checks

Input rows	Eligible rows	Analysis rows	Unique IDs	Human adj.	Non-adj.	Outcome yes	Outcome error
2,690	1,594	1,594	1,594	623	971	767	23

Additional checks: non-adjudicated mode consensus violations = 0; non-adjudicated direction-positive violations = 0; missing `analysis_mode` = 0; missing `platform_legibility` = 0; auto-pre-review platform-legibility mismatches after final message-location review = 247; final-rule platform-legibility mismatches = 0. Across the full corpus, raw carrier labels have any-label disagreement on 507 posts and exact-set agreement on 1,087 posts.

The direction finalization rule is deterministic: adjudicated posts use human final labels; non-adjudicated posts are set to direction-negative only because both LLMs were negative on all three direction labels and the post did not trigger the high-risk queue. Carrier is handled differently. The default carrier variables use human adjudication for queued posts and GPT-5.4 carrier labels for non-adjudicated posts because GPT-5.4 is the stronger production carrier source in the adjudicated comparison. The workbook also preserves GPT-5.4, DeepSeek, union, intersection, and per-label disagreement carrier variants for sensitivity checks.

C.2 Final descriptive distributions

Legibility preprocessing rule. `platform_legibility` is not a content label and is not assigned by LLMs. The pipeline first uses OCR-enriched text to propose `message_location` and flags uncertain rows for review; human review then overwrites uncertain `message_location` values; only after that final field is fixed does the analysis layer derive `platform_legibility`. The derivation collapses final message location, post type, title/description length, OCR text, hashtag-cleaned visible text, and video-observability flags into an analysis-level observability category. The resulting variable is used as a platform/text-observability control, not as a measure of political sensitivity.

Figures 2–4 clarify why the regression should not be a direction-only model. Mode is split between public claim/correction and non-claim participation: public claims are the largest category, but playful spectatorship alone accounts for more than one-quarter of the corpus. Direction is much rarer than event participation, so the three direction dummies identify a narrow theory-bearing subset rather than the event as a whole. Legibility is also heterogeneous: high-caption posts are the largest group, medium-OCR posts make up more than one-third of the corpus, and low visual/video posts remain a nontrivial smaller group. This makes legibility a necessary observability control, especially because the outcome itself is produced by a public-search procedure.

Table 13: Script-derived message-location and platform-legibility labels

Label	Meaning
Platform legibility	
`high_caption_legibility`	Title/body text is the main information source and is long enough for text analysis.
`medium_ocr_legibility`	The post mainly depends on OCR-extracted image text; quality depends on OCR accuracy.
`medium_caption_video_legibility`	The post is video-centered, but the caption/body is long enough to transmit substantive information.
`low_visual_or_video_legibility`	The main information is likely in video or visual material, while available text is weak.
`tag_only_low_legibility`	Visible text is almost entirely hashtags or too short for reliable text classification.
`unclear_legibility`	The script cannot confidently determine the legibility level.
Message location	
`title_primary`	The title carries the main information.
`caption_primary`	Title plus description/body carry the main information.
`image_text_primary`	OCR-extracted image text carries the main information.
`mixed_caption_image`	Caption/body and image text both carry substantive information.
`mixed_caption_video`	Video is central, but the caption/body also carries enough information for analysis.
`video_primary_unobserved`	The post is video-centered and the main content is not directly observed by the text pipeline.
`low_text_or_tag_only`	Visible text is too short or mostly hashtags.
`unclear_location`	The script cannot confidently locate the main information source.

The main mapping is: `title_primary` and `caption_primary` with sufficient text map to high caption legibility; `image_text_primary` and `mixed_caption_image` map to medium OCR legibility; `mixed_caption_video` maps to medium caption-video legibility; `video_primary_unobserved` and weak-text image/video cases map to low visual/video legibility; `low_text_or_tag_only` maps to tag-only low legibility; unresolved cases map to unclear legibility.

Table 14: Mode, direction, carrier, and legibility distributions in the 1,594-post corpus

Construct	Main distribution	Interpretation
Mode	Public claim/correction 609 (38.2%); playful spectatorship 441 (27.7%); information/evidence work 238 (14.9%); inquiry/reflection 225 (14.1%); unable 81 (5.1%)	The event is not only political assertion; playful and inquiry/evidence modes are large.
Direction status	No theory direction 1,343 (84.3%); anti-Manchu/Qing 82 (5.1%); anti-West 10 (0.6%); feminist lineage 93 (5.8%); mixed 9 (0.6%); uncodable 57 (3.6%)	Theory-bearing direction is rare; dummy counts are larger because mixed posts can count in multiple directions.
Direction dummies	Anti-West 14 (0.9%); anti-Manchu/Qing 91 (5.7%); feminist lineage 98 (6.1%)	Direction is multilabel, so dummy totals do not equal direction-status totals.
Default carrier dummies	Hong–Kangxi 878 (55.1%); Red Chamber 235 (14.7%); popular culture 218 (13.7%); broader history/politics 293 (18.4%)	Carriers overlap and are controls/sensitivity variables, not mutually exclusive topics.
Legibility	High caption 728 (45.7%); medium OCR 570 (35.8%); low visual/video 234 (14.7%); medium caption-video 54 (3.4%); unclear 5 (0.3%); tag-only low 3 (0.2%)	Legibility is derived from final reviewed message location and should not be treated as a political-content label.

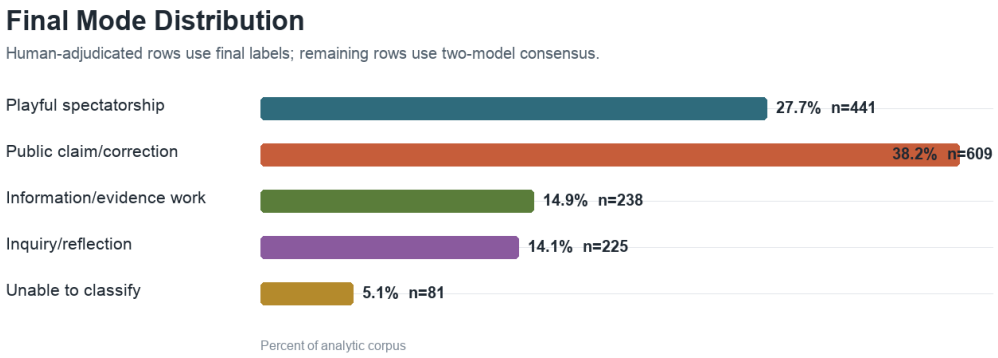

Figure 2: Final mode distribution.

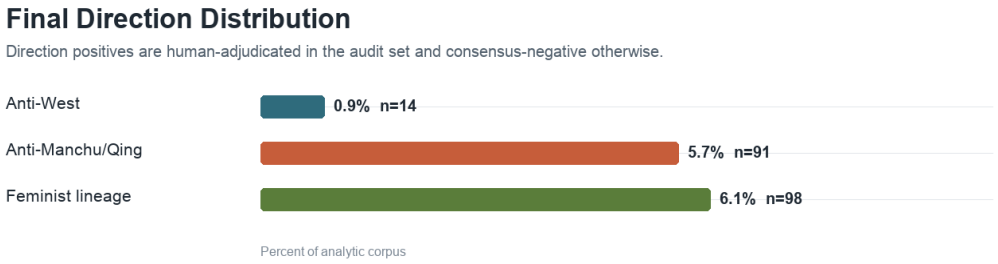

Figure 3: Final direction-dummy distribution. Direction is multilabel; posts can be positive on more than one direction.

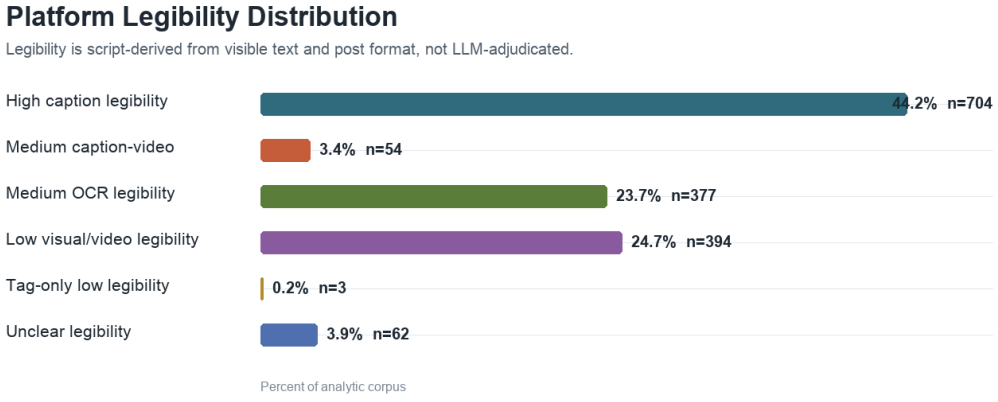

Figure 4: Platform-legibility distribution. Legibility is derived from final reviewed message location and OCR/video observability rather than from LLM content judgment.

C.3 Bivariate visibility by construct

Bivariate visibility rates are descriptive. They do not adjust for other fields and should not be used as the main answer to RQ2 or RQ3.

Table 15: Selected bivariate search-nonretrievable rates

Construct level	Non-error N	Search-nonretrievable	Rate	Wilson 95% CI
Playful spectatorship	437	212	48.5%	43.9–53.2
Public claim/correction	603	318	52.7%	48.7–56.7
Information/evidence work	235	112	47.7%	41.4–54.0
Inquiry/reflection	216	108	50.0%	43.4–56.6
Anti-Manchu/Qing direction	88	54	61.4%	50.9–70.9
Anti-West direction	13	6	46.2%	23.2–70.9
Feminist-lineage direction	97	63	64.9%	55.0–73.7
High caption legibility	728	346	47.5%	43.9–51.2
Medium OCR legibility	549	298	54.3%	50.1–58.4
Low visual/video legibility	233	126	54.1%	47.7–60.4
Medium caption-video legibility	54	29	53.7%	40.6–66.3

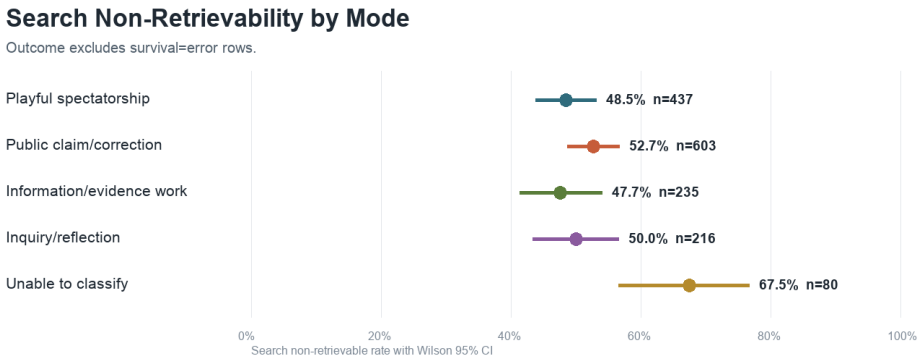

Figure 5: Bivariate search-nonretrievability by mode.

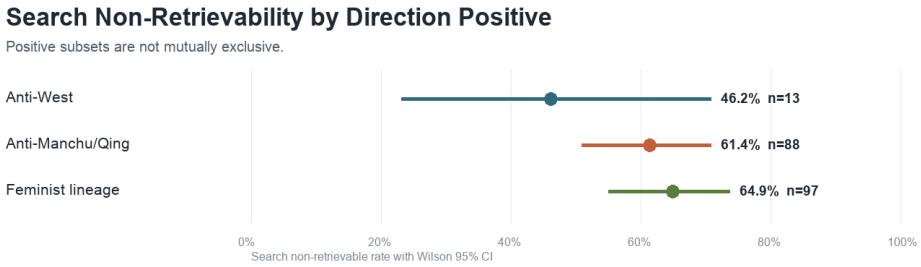

Figure 6: Bivariate search-nonretrievability by direction-positive dummy.

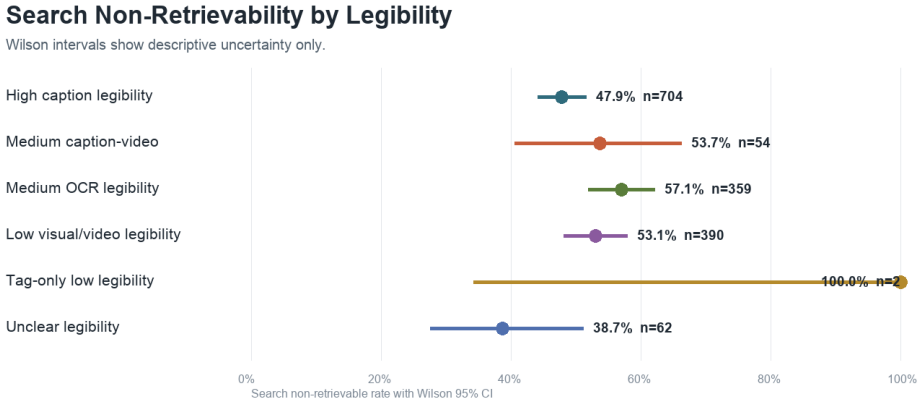

Figure 7: Bivariate search-nonretrievability by platform legibility.

The bivariate mode plot shows only modest separation among codable modes: public claim/correction is somewhat higher (52.7%) than playful spectatorship (48.5%), information/evidence work (47.7%), and inquiry/reflection (50.0%). The high rate for `unable_to_classify` is not substantively interpreted because those rows are filtered out of the regression sample. Direction is more separated descriptively: anti-Manchu/Qing and feminist-lineage posts are above 60%, while anti-West has a very wide interval because only 13 non-error posts are positive. Legibility shows an observability pattern: medium OCR, low visual/video, and medium caption-video posts are all above high-caption posts, but their rates are close enough that the regression should treat legibility as a format/searchability control rather than as a standalone moderation result. These plots are therefore useful diagnostics, but they motivate the adjusted models rather than replacing them.

C.4 Carrier handling and production disagreement

Carrier is retained because it helps separate material pathways from theory-bearing direction, but it is not treated as a core theoretical IV. The default rule is human-adjudicated carrier for the 623 queued posts and GPT-5.4 carrier for the 971 non-queued posts. GPT-5.4 is preferred over DeepSeek for default carrier because, against human-adjudicated labels, its exact-set agreement is 83.6%, compared with 71.3% for DeepSeek. Per-label GPT-5.4 carrier kappas in the adjudicated set are 0.798 for Hong–Kangxi, 0.950 for Red Chamber, 0.915 for popular culture, and 0.859 for broader history/politics.

Table 16: Carrier raw prevalence under GPT-5.4 and DeepSeek variants

Variant	Hong–Kangxi	Red Chamber	Popular culture	Broader history/politics
GPT-5.4 raw	826 (51.8%)	236 (14.8%)	222 (13.9%)	316 (19.8%)
DeepSeek raw	928 (58.2%)	187 (11.7%)	209 (13.1%)	196 (12.3%)
Default analysis rule	878 (55.1%)	235 (14.7%)	218 (13.7%)	293 (18.4%)

Note: The default rule combines human adjudication for queued posts with GPT-5.4 for non-queued posts. The default counts are derived from raw GPT counts and adjudicated human-vs-GPT positive counts.

Table 17: Bivariate search-nonretrievability by raw carrier model variant

Variant	Carrier	Non-error N	Search-nonretrievable	Rate	Wilson 95% CI
GPT-5.4	Hong–Kangxi	820	393	47.9%	44.5–51.3
GPT-5.4	Red Chamber	229	151	65.9%	59.6–71.8
GPT-5.4	Popular culture	219	101	46.1%	39.6–52.7
GPT-5.4	Broader history/politics	311	162	52.1%	46.5–57.6
DeepSeek	Hong–Kangxi	919	454	49.4%	46.2–52.7
DeepSeek	Red Chamber	182	120	65.9%	58.8–72.4
DeepSeek	Popular culture	206	96	46.6%	39.9–53.4
DeepSeek	Broader history/politics	193	103	53.4%	46.3–60.3

Interpretation: Red Chamber suoyin shows the clearest descriptive carrier contrast under both models. Hong–Kangxi and popular-culture carriers do not, by themselves, indicate higher search non-retrievability. Because carriers overlap with direction, legibility, and mode, this table is descriptive; the adjusted model treats carriers as controls and sensitivity variants.

Table 18: Carrier agreement against human-adjudicated labels in the 623-post queue

Model	Label	Human pos.	Model pos.	Accuracy	κ	Positive F1
GPT-5.4	Hong–Kangxi	411	359	0.904	0.798	0.922
GPT-5.4	Red Chamber	109	110	0.986	0.950	0.959
GPT-5.4	Popular culture	79	83	0.981	0.915	0.926
GPT-5.4	Broader history/politics	159	182	0.944	0.859	0.897
DeepSeek	Hong–Kangxi	411	392	0.854	0.682	0.887
DeepSeek	Red Chamber	109	91	0.958	0.845	0.870
DeepSeek	Popular culture	79	84	0.963	0.838	0.859
DeepSeek	Broader history/politics	159	124	0.892	0.695	0.763

Table 19: Exact carrier-set agreement against human adjudication

Model	Exact-set matches	Human-adjudicated N	Exact-set agreement
GPT-5.4	521	623	83.6%
DeepSeek	444	623	71.3%

Exact-set agreement requires the full four-label carrier set to match, not merely a per-label majority. This is why GPT-5.4 is used as the non-adjudicated default carrier source while DeepSeek remains in sensitivity analysis.

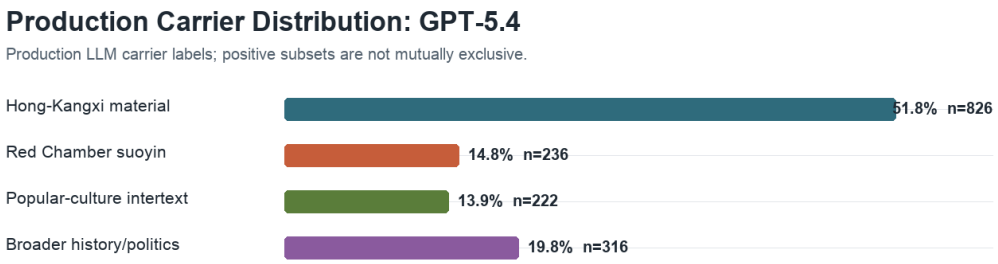

Figure 8: Raw GPT-5.4 carrier distribution.

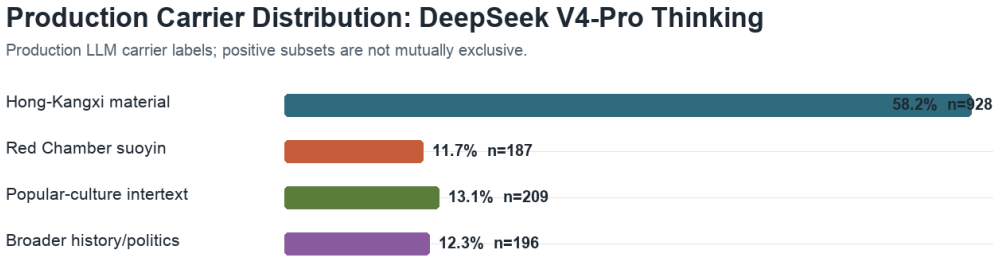

Figure 9: Raw DeepSeek carrier distribution.

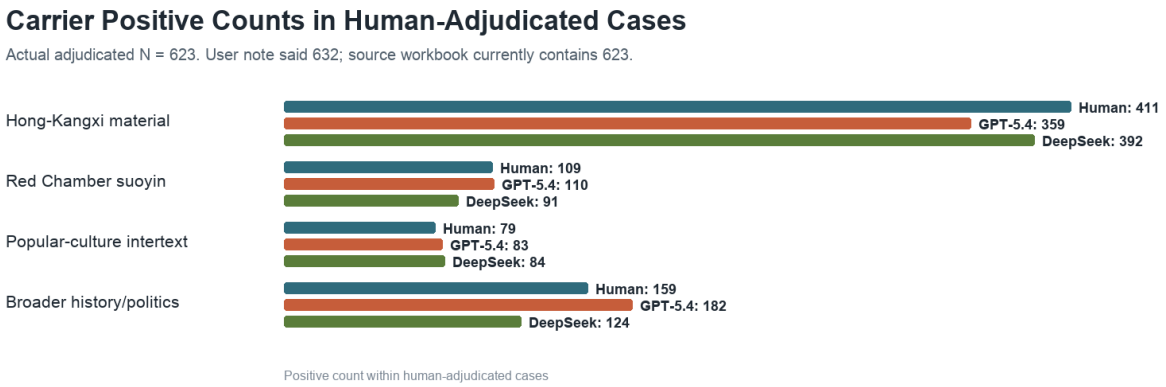

Figure 10: Carrier positive counts in the human-adjudicated queue versus production LLM outputs.

C.5 Direction co-occurrence and cross-structure profiles

The three theory-bearing directions are not simply different names for the same posts. Pairwise overlap is rare. Anti-West overlaps with anti-Manchu/Qing in four cases, but otherwise the direction dummies mostly identify separate interpretive moves. This matters for the regression because the feminist-lineage result is not being driven by systematic co-occurrence with anti-Manchu/Qing or anti-West labels.

Table 20: Pairwise direction co-occurrence in the full 1,594-post corpus

Direction pair	Joint positives	First direction N	Second direction N	Joint / first	Joint / second
Anti-West + Anti-Manchu/Qing	4	14	91	28.6%	4.4%
Anti-West + Feminist lineage	0	14	98	0.0%	0.0%
Anti-Manchu/Qing + Feminist lineage	5	91	98	5.5%	5.1%

The direction fields are multilabel, so overlap is allowed. The table reports full-corpus final direction dummies before excluding non-error/codable regression rows.

The profile plots show that direction is also tied to different rhetorical and material pathways. All three direction-positive groups are overwhelmingly claim-oriented: public claim/correction accounts for 78.6% of anti-West posts, 82.4% of anti-Manchu/Qing posts, and 91.8% of feminist-lineage posts, compared with 32.3% among no-theory posts. Carrier profiles are more differentiated. Anti-West posts are mostly broader-history/politics material (85.7%), anti-Manchu/Qing posts combine broader-history/politics (72.5%), Hong–Kangxi material (61.5%), and Red Chamber suoyin (35.2%), while feminist-lineage posts rely heavily on Hong–Kangxi material (61.2%) but much less on broader-history/politics (9.2%). Legibility profiles add a format dimension: anti-Manchu/Qing is relatively caption-legible, whereas feminist-lineage posts have more low visual/video cases than anti-Manchu/Qing.

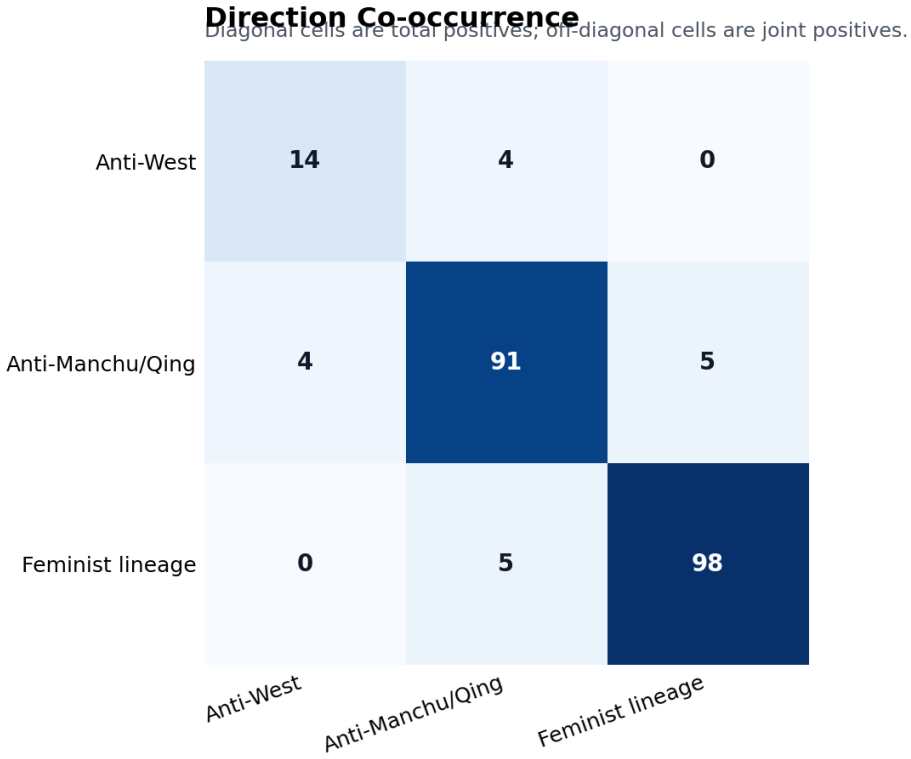

Figure 11: Direction co-occurrence heatmap. Diagonal cells are total positives; off-diagonal cells are joint positives.

Table 21: Direction profile summary across mode, carrier, legibility, visibility, and engagement

Group	Full N	Mode profile	Carrier profile	Legibility profile	Visibility / engagement profile
No theory	1,343	Public claim 32.3%; playful 31.5%	HKX 56.7%; other carriers each 14–16%	High caption 46.6%; OCR 36.8%; low visual/video 13.0%	Regression-sample N = 1,277; rate 48.6%; median engagement 57
Anti-West	14	Public claim 78.6%; inquiry 14.3%	Broader history/politics 85.7%; Red Chamber 21.4%	High caption 64.3%; OCR 28.6%; low visual/video 7.1%	Regression-sample N = 13; rate 46.2%; median engagement 183; low support
Anti-Manchu/Qing	91	Public claim 82.4%; evidence work 9.9%	Broader history/politics 72.5%; HKX 61.5%; Red Chamber 35.2%	High caption 56.0%; OCR 34.1%; medium caption-video 5.5%	Regression-sample N = 88; rate 61.4%; median engagement 56.5
Feminist lineage	98	Public claim 91.8%; playful 4.1%	HKX 61.2%; Red Chamber 11.2%; broader history/politics 9.2%	High caption 46.9%; OCR 27.6%; low visual/video 18.4%	Regression-sample N = 94; rate 64.9%; median engagement 221

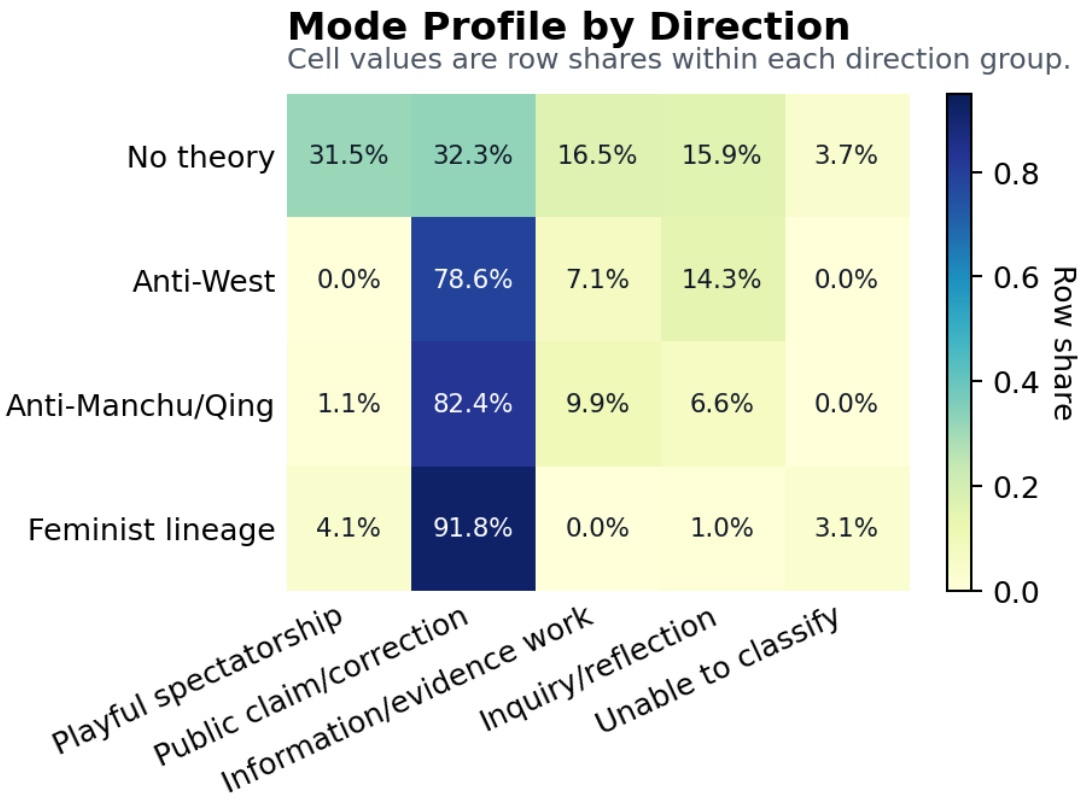

Figure 12: Mode profile by direction group. Cell values are row shares within each group.

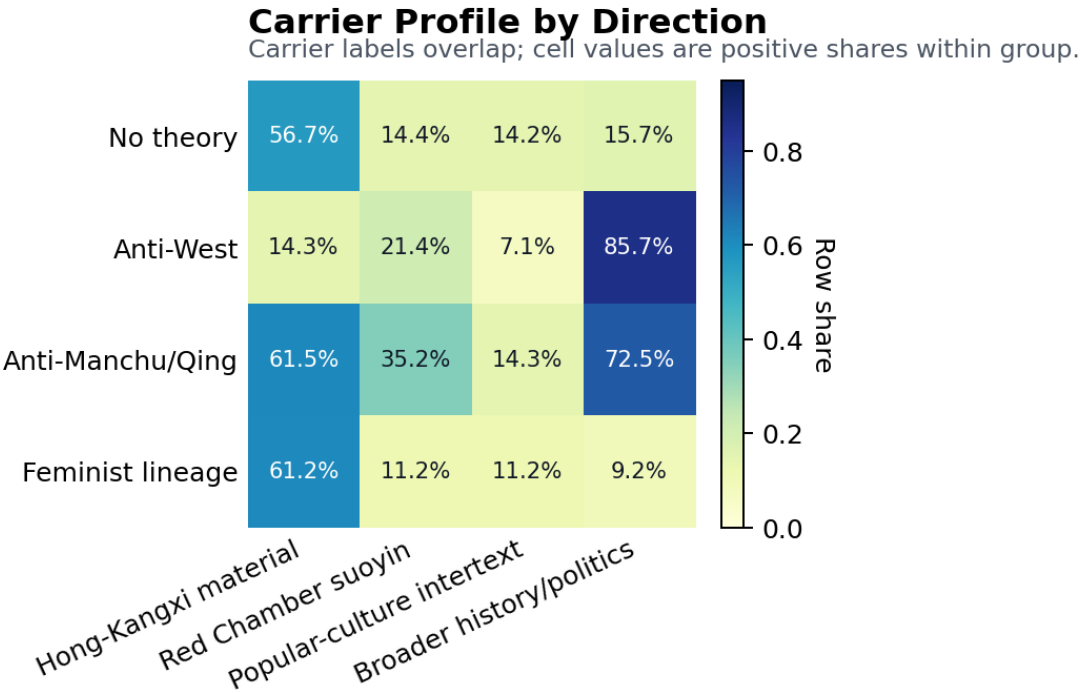

Figure 13: Carrier profile by direction group. Carrier labels overlap, so row percentages do not sum to 100%.

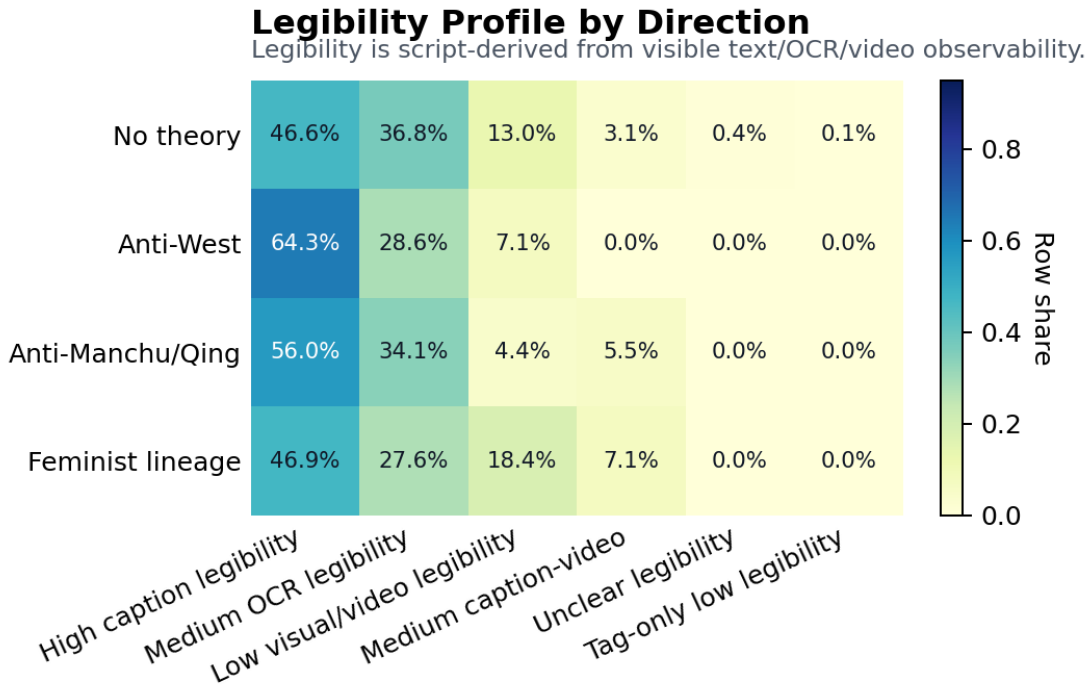

Figure 14: Legibility profile by direction group. Legibility is derived from final reviewed message location, visible text, OCR, and video observability.

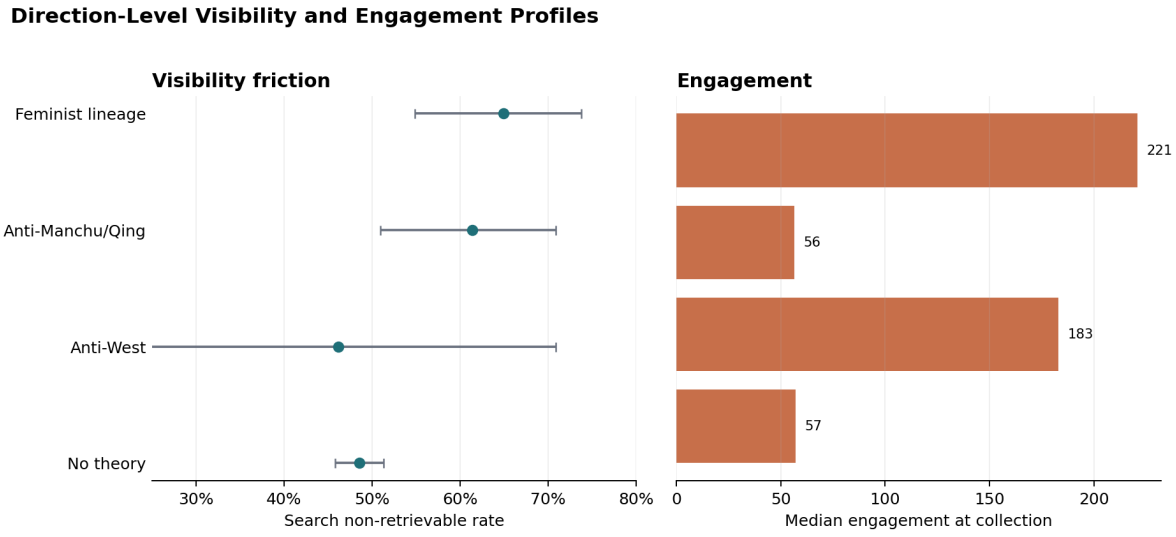

Figure 15: Direction-level bivariate visibility and engagement profile in the non-error, codable regression sample. Error bars are Wilson 95% intervals; engagement is the median count at collection.

C.6 Direction-focused and within-direction regression extensions

The supplemental regressions ask two narrower questions. First, if each direction is compared against no-theory posts under the same reduced comparison setup, does the direction coefficient reproduce the main model pattern? Second, within the direction-positive subset, do carrier, legibility, or engagement factors appear to explain which posts are search-nonretrievable? These models are descriptive and low-powered; they are included to diagnose structure, not to replace the main specification.

Table 22: Direction-focused logits against no-theory baseline

Comparison	Model N	Direction N	Direction events	OR	95% CI	p
Anti-West vs no theory	1,290	13	6	0.76	[0.23, 2.52]	0.653
Anti-Manchu/Qing vs no theory	1,365	88	54	1.29	[0.78, 2.12]	0.316
Feminist lineage vs no theory	1,371	94	61	2.17	[1.29, 3.64]	0.003

Each row estimates the target direction against no-theory posts, controlling for mode, carrier, evidence-work, legibility, post type, posting date, visible text length, and engagement. Anti-West remains a low-support estimate.

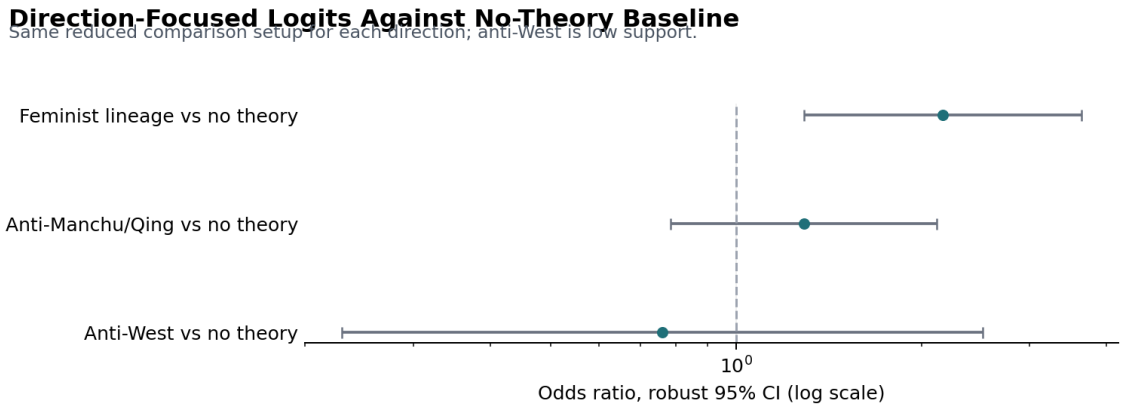

Figure 16: Direction-focused logits against no-theory baseline.

The focused models reproduce the main result: feminist-lineage direction remains the only direction with a clear adjusted association. Anti-Manchu/Qing is positive but not statistically stable; anti-West is too small to read as a substantive null. The within-direction models are even more cautious. Anti-West is not estimated because the subset has only 13 non-error, codable posts and six events. Within anti-Manchu/Qing posts, Red Chamber suoyin is the only clearly positive reduced-model factor (OR = 4.36, 95% CI [1.23, 15.51]), suggesting that the anti-Manchu/Qing visibility pattern is partly concentrated in posts that connect the rumor to Red Chamber suoyin material. Within feminist-lineage posts, the reduced model points instead to a negative HKX-carrier association, while other carrier, legibility, and engagement factors are not independently stable. In the engagement OLS checks, search-nonretrievability is not associated with clearly lower engagement within either estimable direction; if anything, the point estimates are positive but imprecise.

Table 23: Reduced within-direction models and engagement checks

Direction	Reduced visibility logit summary	Engagement OLS summary
Anti-West	Not estimated: $N = 13$, events = 6.	Not estimated because the subset is too small.
Anti-Manchu/Qing	Red Chamber carrier OR = 4.36 [1.23, 15.51], $p = .023$; HKX carrier OR = 0.29 [0.08, 1.03], $p = .055$; log engagement OR = 1.18 [0.93, 1.49], $p = .165$.	Search-nonretrievable coefficient = 0.72 log points [-0.33, 1.78], $p = .180$.
Feminist lineage	HKX carrier OR = 0.33 [0.12, 0.91], $p = .032$; public-claim mode OR = 3.33 [0.46, 24.35], $p = .236$; log engagement OR = 1.13 [0.93, 1.36], $p = .219$.	Search-nonretrievable coefficient = 0.69 log points [-0.44, 1.83], $p = .229$.

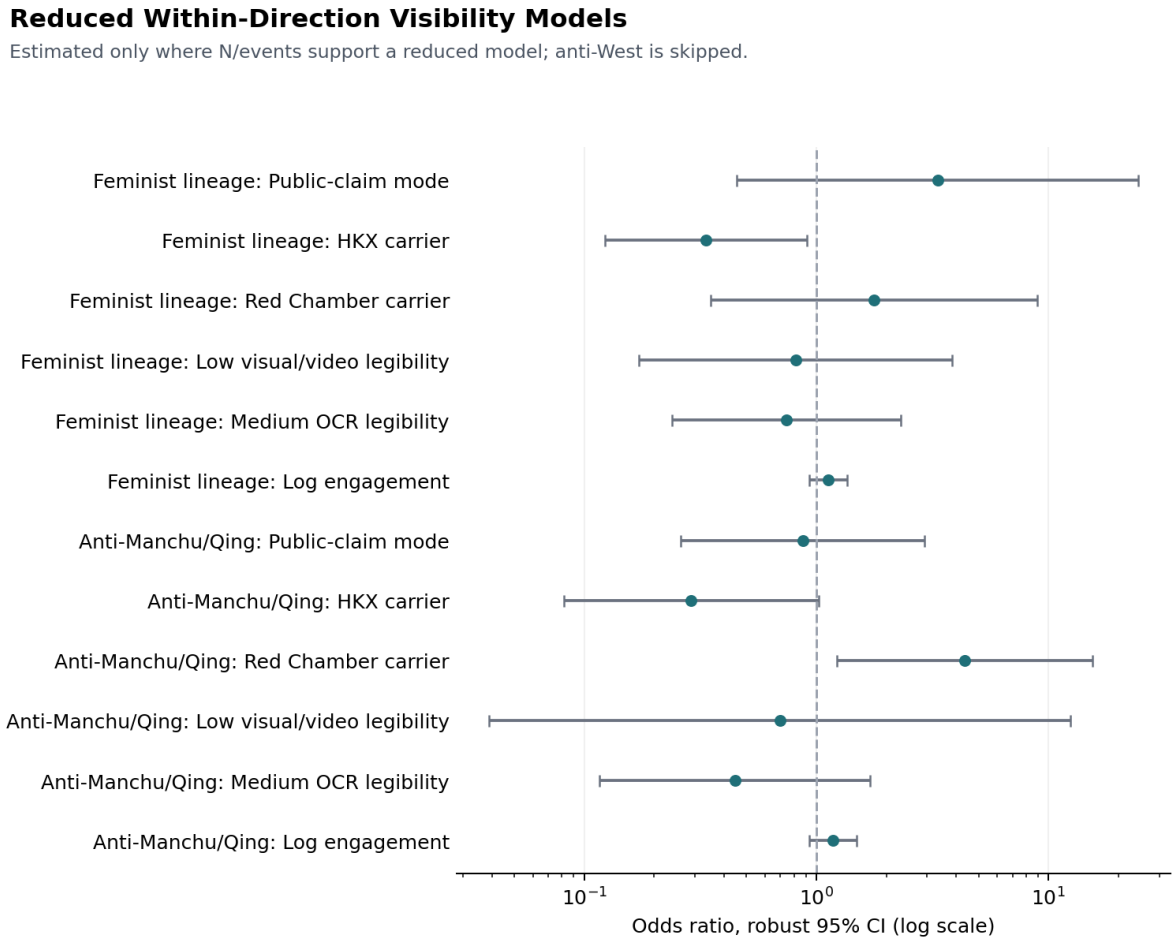

Figure 17: Reduced within-direction visibility models. Only anti-Manchu/Qing and feminist-lineage subsets have enough observations for the reduced model.

C.7 Regression specification

The main descriptive model is:

$$\begin{aligned}\text{logit } \Pr(Y_i = 1) = \alpha &+ \beta_1 AMQ_i + \beta_2 AW_i + \beta_3 FEM_i \\ &+ \boldsymbol{\beta}_4 Mode_i + \boldsymbol{\beta}_5 Carrier_i \\ &+ \beta_6 Evidence_i + \boldsymbol{\beta}_7 Legibility_i \\ &+ X_i\boldsymbol{\gamma} + \delta_{q(i)} + \varepsilon_i.\end{aligned}$$

Here Y_i is public-search non-retrievability. *Direction* contains the three direction dummies; *Mode* is categorical with playful spectatorship as baseline; *Carrier* contains the four carrier dummies; *Evidence* is the binary has-evidence-work indicator; *Legibility* is script-derived; X_i contains post type, posting date, log engagement, and log visible-text length; and $\delta_{q(i)}$ are keyword fixed effects. The sample excludes `survival=error` and mode/content-uncodable posts. Standard errors are robust; account-level clustering is not used because stable account IDs were not captured.

The coefficient plots below transform the fitted logit coefficients into odds ratios. The vertical reference line is OR = 1, which means no difference in the odds of search non-retrievability relative to the relevant baseline category or a one-unit increase for continuous predictors. Values to the right of 1 indicate higher odds; values to the left indicate lower odds. The x-axis is shown on a log scale so that distances above and below 1 are visually comparable.

Table 24: Regression model diagnostics and design variants

Model	N	Parameters	Event rate	Pseudo-R^2	Direction source	Carrier source
Main with engagement	1,463	24	50.2%	0.049	Final analysis	Final analysis
Main without engagement	1,463	23	50.2%	0.044	Final analysis	Final analysis
Without Hong–Kangxi carrier	1,463	23	50.2%	0.048	Final analysis	Final analysis minus HKX
Direction-status model	1,463	25	50.2%	0.049	Direction status	Final analysis
LLM direction union	1,463	24	50.2%	0.048	Raw LLM union	Final analysis
LLM direction intersection	1,463	24	50.2%	0.051	Raw LLM intersection	Final analysis
Carrier union	1,463	24	50.2%	0.048	Final analysis	Carrier union
Carrier intersection	1,463	24	50.2%	0.044	Final analysis	Carrier intersection
DeepSeek carrier	1,463	24	50.2%	0.044	Final analysis	DeepSeek

All models converged. The sample is constant across models after excluding `survival=error` and mode/content-uncodable posts. Pseudo-R^2 is used only as a diagnostic, not as a model-selection target.

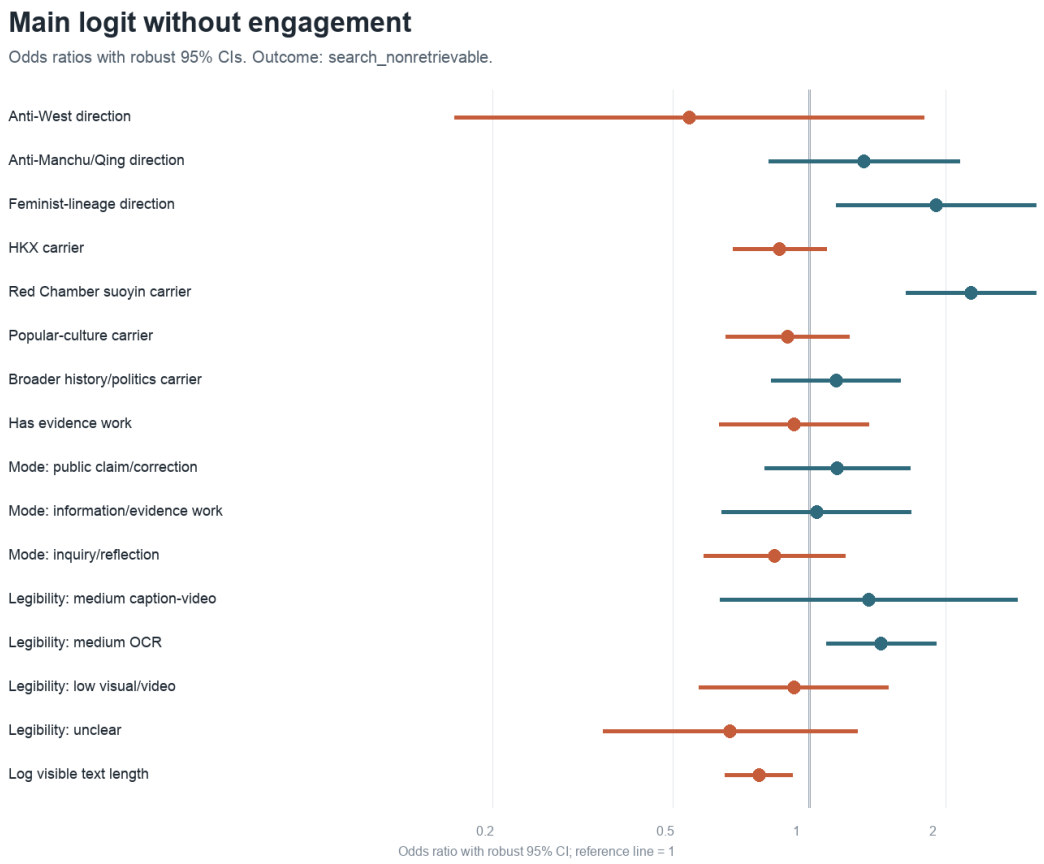

Figure 18: Main logit without engagement controls, plotted as odds ratios with robust 95% confidence intervals. The vertical reference line is OR = 1.

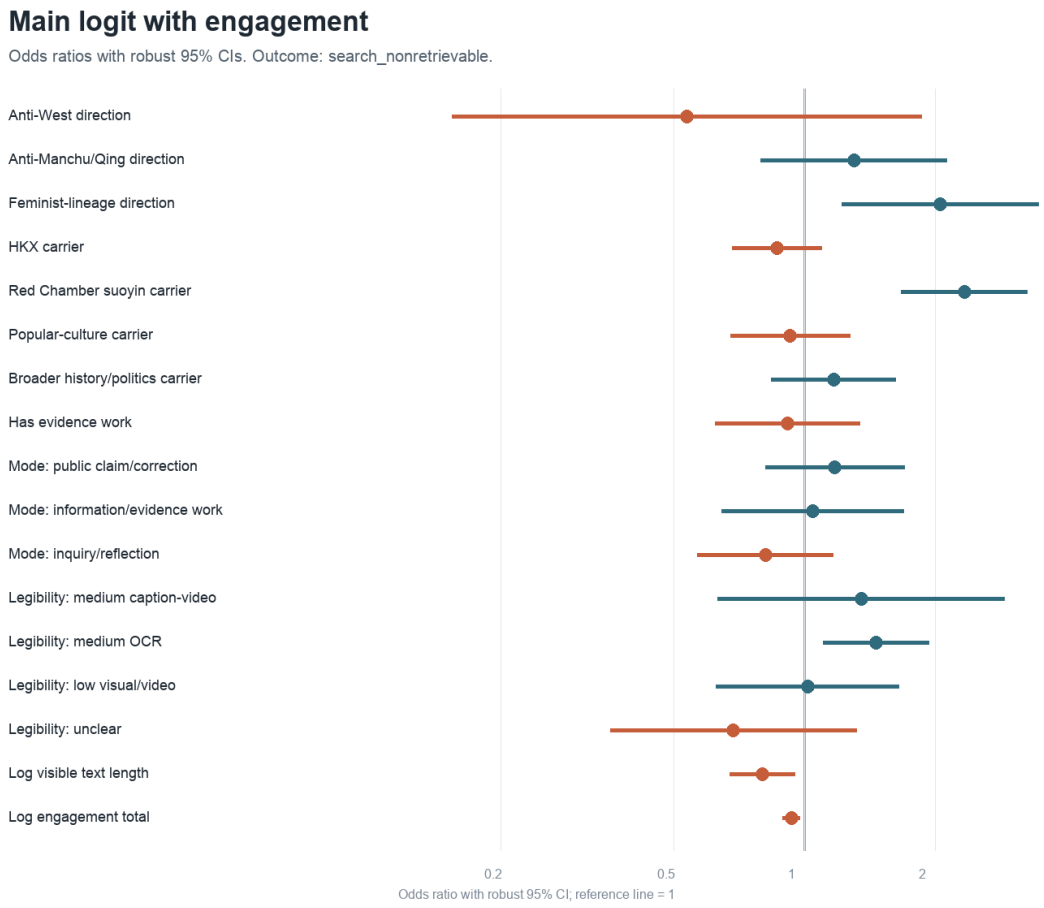

Figure 19: Main logit with engagement controls, plotted as odds ratios with robust 95% confidence intervals. This is the figure summarized in the body; the vertical reference line is OR = 1.

C.8 Sensitivity models

The sensitivity checks ask whether the main pattern depends on engagement, the Hong–Kangxi carrier, final human-adjudicated direction labels, the GPT-default carrier rule, or OCR/searchability mechanics. The stable signals are feminist-lineage direction, Red Chamber suoyin carrier, and medium OCR legibility, but they are not interpreted in the same way: medium OCR is a likely observability mechanism, while Red Chamber and feminist-lineage are substantive content associations. Anti-Manchu/Qing direction remains positive in most variants but does not stabilize into a clear result. Anti-West remains too rare.

Table 25: Sensitivity model summary for central terms

Model	Anti-West OR	Anti-Manchu/Qing OR	Feminist OR	Red Chamber OR	Medium OCR OR
Main with engagement	0.53	1.27	2.06	2.30	1.37
Main without engagement	0.53	1.29	1.90	2.24	1.33
Without HKX carrier	0.55	1.25	2.06	2.35	1.37
LLM direction union	0.73	1.17	1.94	2.31	1.36
LLM direction intersection	0.78	1.49	2.56	2.30	1.36
Carrier union	0.54	1.31	2.00	2.28	1.37
Carrier intersection	0.56	1.30	2.05	2.18	1.36
DeepSeek carrier	0.59	1.33	2.03	2.24	1.37
Direction-status model	0.41	1.29	2.14	2.29	1.37

Note: Direction-status model entries for the three directions are categorical direction-status coefficients rather than dummy coefficients; they are included only to show qualitative robustness. Full coefficient plots follow.

Engagement requires a separate diagnostic because it is substantively ambiguous. Low engagement could make a post less likely to attract moderation attention, but it could also mean that the post had little ranking, indexing, or search exposure and was therefore harder to recover through the public search procedure. Because likes and total engagement are extremely right-skewed, the log-engagement control in the main model should not be read as a clean causal adjustment. I therefore add three checks: the distribution of low engagement across corpus categories, the bivariate relationship between engagement bins and search non-retrievability inside the Hong–Kangxi regression sample, and regression sensitivity models that omit engagement, model it categorically, or exclude low-engagement posts.

Table 26: Engagement imbalance across the three cleaned corpus categories

Corpus category	N	Median likes	Likes ≤ 9	Median total engagement	Total engagement ≤ 9
Hong–Kangxi event	1,594	33	31.6%	60	22.6%
Pure Western pseudo-history	593	13	42.8%	38	25.6%
Pure historical analysis	503	22	34.0%	40	24.1%

Total engagement is likes + comments + collections + shares. The pseudo-history and pure-history rows come from the cleaned reference corpus using `pure_fake_his=yes` and `pure_his=yes`; the Hong–Kangxi row is the 1,594-post event corpus.

These checks answer the low-engagement concern directly. The imbalance matters: in the regression sample, posts with zero or 1–9 total engagements are search-nonretrievable more often than mid-engagement posts. This means that the negative log-engagement coefficient in the main model should not be interpreted as "less engagement causes censorship" or "more engagement protects posts." A more cautious reading is that low-exposure posts are less reliably recovered by public search, while high-exposure posts are easier to locate. At the same time, the substantive estimates do not collapse when engagement is omitted, modeled with bins, or restricted to posts above the very-low-engagement

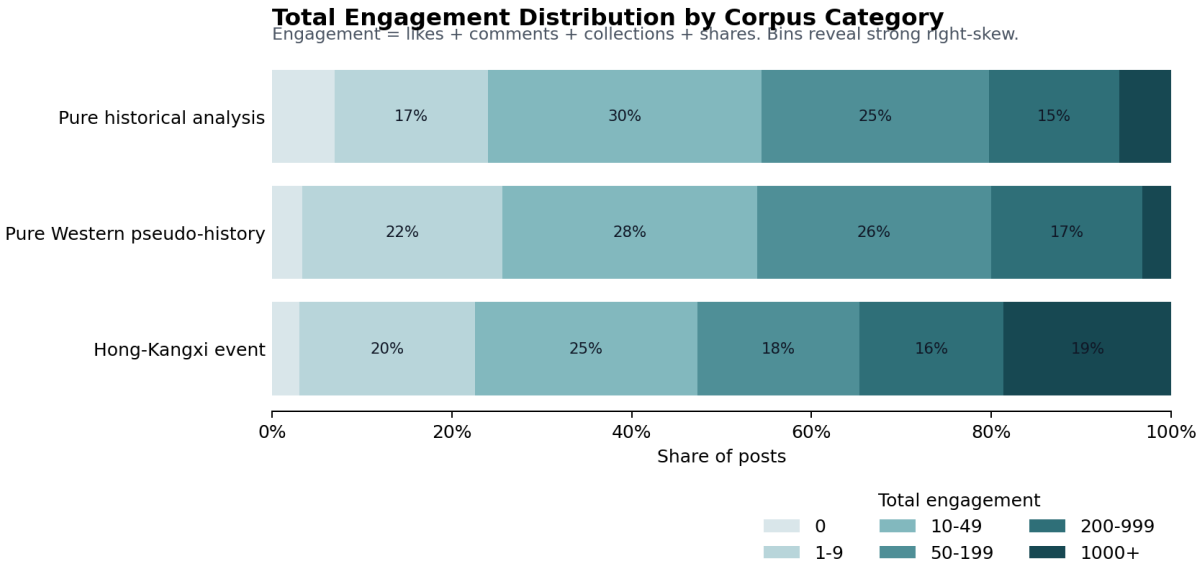

Figure 20: Total-engagement bin composition by corpus category. The stacked bars show that all three corpus categories contain many low-engagement posts, while the Hong–Kangxi event corpus has the largest high-engagement tail.

Table 27: Search non-retrievability by total-engagement bin in the Hong–Kangxi regression sample

Total engagement bin	N	Non-retrievable N	Rate	Wilson 95% CI
0	45	28	62.2%	47.6–74.9%
1–9	277	165	59.6%	53.7–65.2%
10–49	365	184	50.4%	45.3–55.5%
50–199	265	127	47.9%	42.0–53.9%
200–999	230	95	41.3%	35.1–47.8%
1000+	281	136	48.4%	42.6–54.2%

Table 28: Engagement sensitivity for central terms

Model	N	Event rate	Anti-Manchu/Qing OR	Feminist OR	Red Chamber OR	Medium OCR OR
Main with log engagement	1,463	50.2%	1.27 [0.77, 2.09]	2.06 [1.22, 3.46]	2.30 [1.64, 3.21]	1.37 [1.05, 1.79]
Without engagement	1,463	50.2%	1.29 [0.80, 2.11]	1.90 [1.14, 3.16]	2.24 [1.61, 3.12]	1.33 [1.02, 1.74]
Engagement-bin controls	1,463	50.2%	1.30 [0.79, 2.14]	2.04 [1.21, 3.44]	2.32 [1.65, 3.24]	1.36 [1.04, 1.78]
Exclude total engagement ≤ 9	1,141	47.5%	1.42 [0.83, 2.43]	2.18 [1.27, 3.75]	2.51 [1.74, 3.63]	1.26 [0.92, 1.73]
Exclude likes ≤ 9	1,003	47.0%	1.59 [0.91, 2.77]	2.27 [1.29, 4.02]	2.39 [1.62, 3.54]	1.28 [0.91, 1.80]

The engagement-bin model replaces linear log engagement with categorical total-engagement bins, using 10–49 as the baseline. In that model, the 1–9 bin itself has OR = 1.52 [1.09, 2.13] relative to 10–49, while the zero-engagement bin is positive but imprecise (OR = 1.77 [0.90, 3.49]).

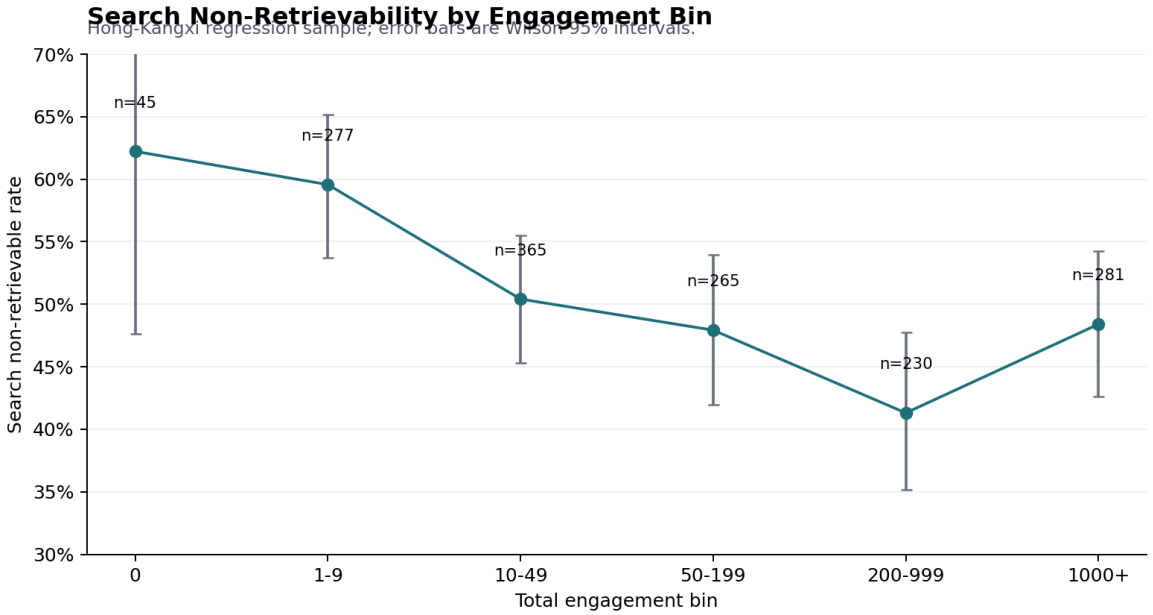

Figure 21: Bivariate search non-retrievability by total-engagement bin in the Hong–Kangxi regression sample. The lowest-engagement bins have the highest non-retrievability rates, so engagement is part of the visibility problem rather than a neutral control.

threshold. Feminist-lineage direction and Red Chamber suoyin remain positive in all five specifications. Medium OCR legibility is the term most affected by excluding low-engagement posts, which supports the broader interpretation that the OCR result partly captures format and searchability rather than a clean political-content effect.

One additional concern is mechanical rather than political: OCR-dependent or weak-text posts may be search-nonretrievable because the search procedure has less reliable text to match, not because the post was removed or downranked. The main model already partially addresses this by controlling for platform legibility, post type, and visible-text length, but that is not the same as removing the mechanism. I therefore add three restricted-sample checks. The first drops medium-OCR posts. The second keeps only caption-searchable posts: high-caption posts plus video posts whose captions carry enough substantive text. The third is the strictest bound and keeps only high-caption posts.

Table 29: OCR/legibility sensitivity for central terms

Model	N	Event rate	Anti-Manchu/Qing OR	Feminist OR	Red Chamber OR	Log engagement OR
Main with legibility controls	1,463	50.2%	1.27 [0.77, 2.09]	2.06 [1.22, 3.46]	2.30 [1.64, 3.21]	0.93 [0.89, 0.97]
Exclude medium OCR	941	48.0%	1.75 [0.95, 3.23]	1.90 [1.04, 3.48]	2.46 [1.63, 3.70]	0.97 [0.92, 1.03]
Caption-searchable only	780	47.8%	1.83 [0.96, 3.48]	1.81 [0.92, 3.57]	2.35 [1.53, 3.60]	0.94 [0.88, 1.00]
High-caption only	727	47.5%	1.78 [0.91, 3.47]	1.70 [0.83, 3.49]	2.40 [1.55, 3.70]	0.94 [0.88, 1.01]

The main model also estimates the medium-OCR coefficient itself: OR = 1.37 [1.05, 1.79]. Medium OCR is not estimable in the restricted samples because those rows are removed or because legibility no longer varies enough to identify the term.

This check changes the interpretation in a useful way. Medium OCR remains a plausible format/searchability mechanism in the main model, so I should not describe it as equivalent to removal or censorship. Once medium-OCR posts are dropped, the feminist-lineage and Red Chamber suoyin estimates remain positive; Red Chamber remains especially stable. In the stricter caption-searchable

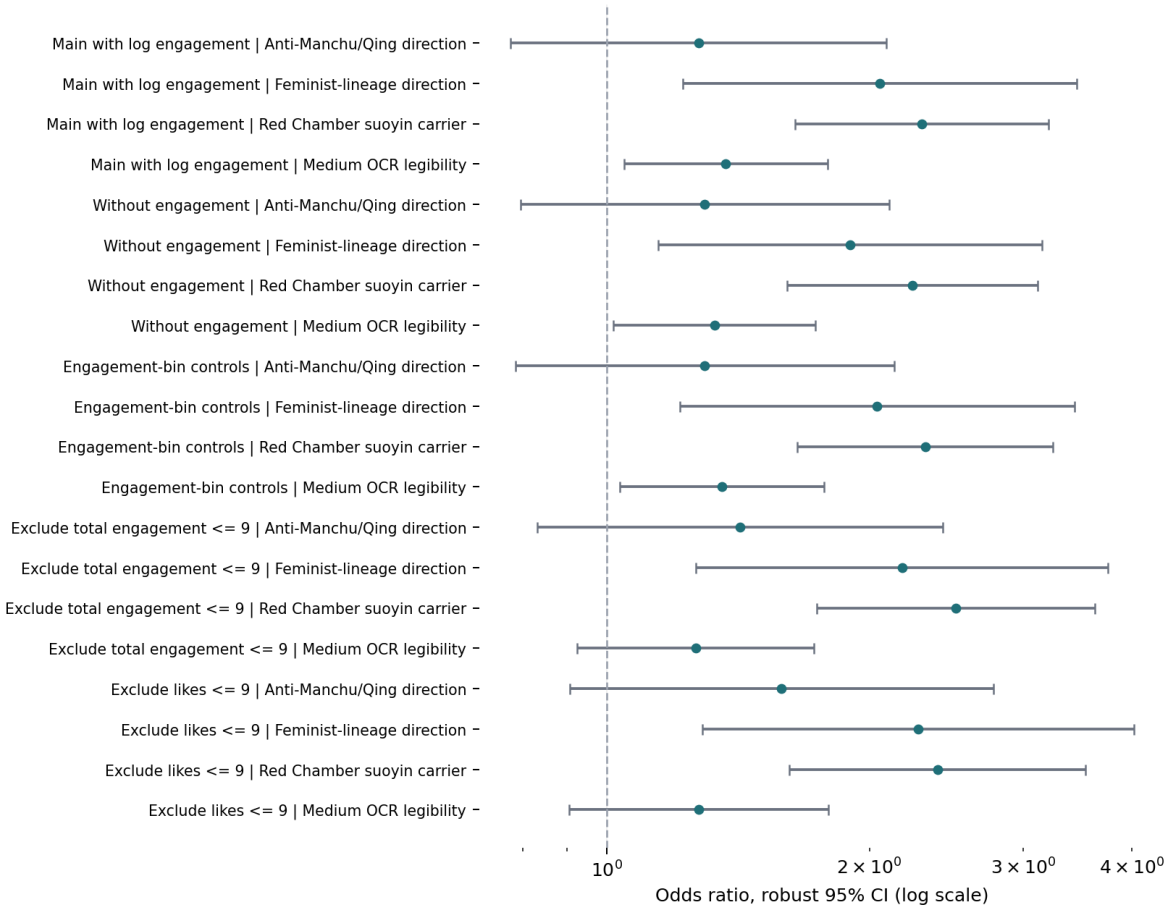

Figure 22: Engagement sensitivity for central terms. Points are odds ratios with robust 95% confidence intervals; the vertical reference line is OR = 1.

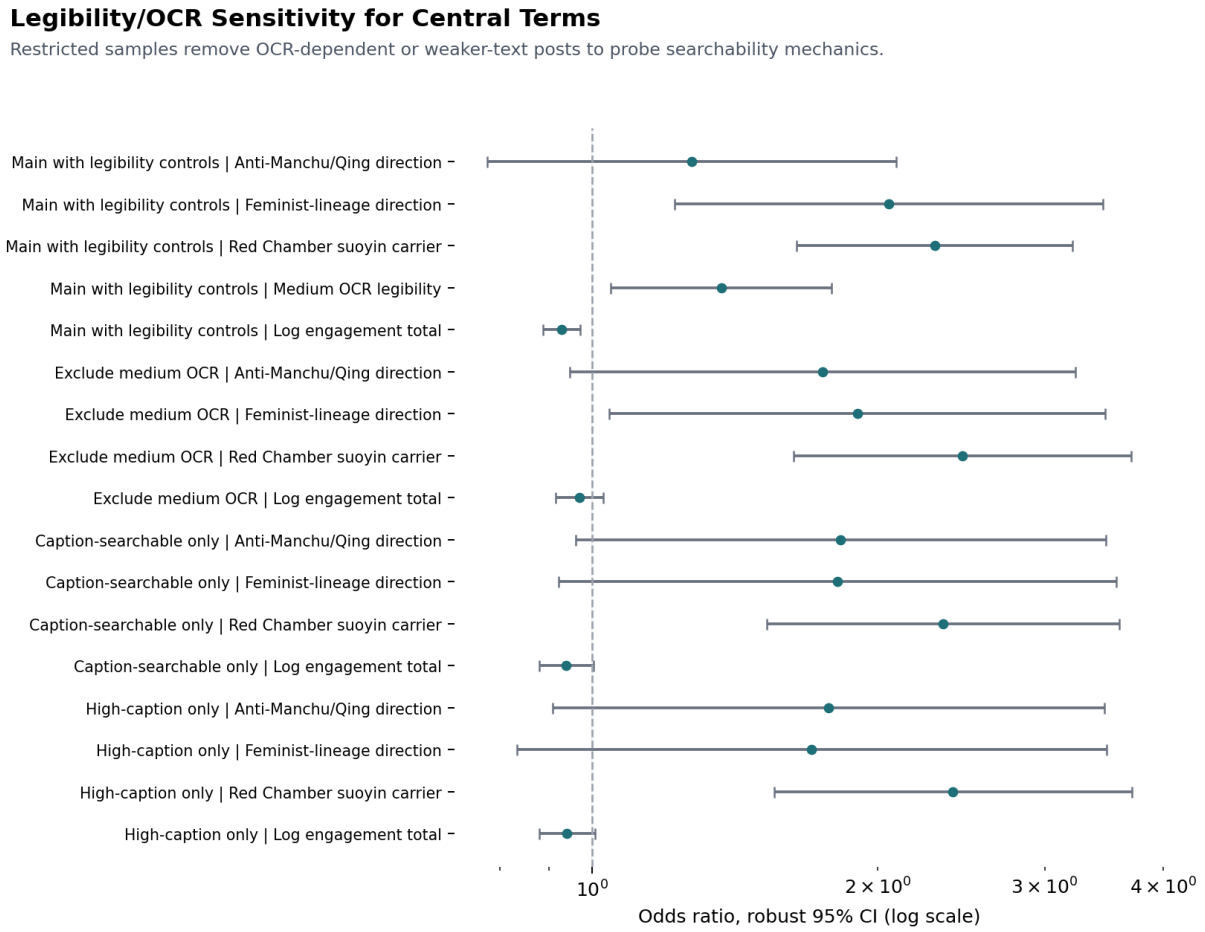

Figure 23: Legibility/OCR sensitivity for central terms. Restricted samples test whether direction, carrier, and engagement results survive after removing OCR-dependent search friction.

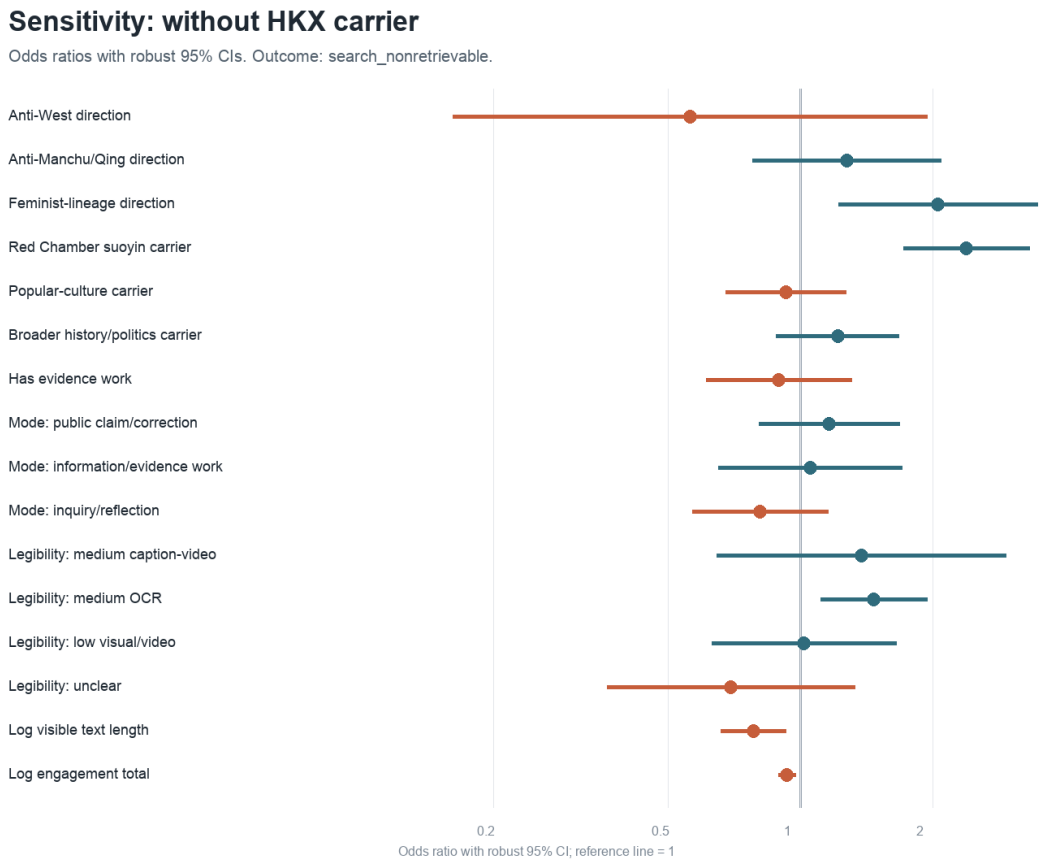

Figure 24: Sensitivity model excluding the Hong–Kangxi carrier control, plotted as odds ratios. The vertical reference line is OR = 1.

and high-caption-only samples, Red Chamber remains clearly to the right of OR = 1, while the feminist-lineage estimate remains positive but loses precision and crosses 1. The engagement coefficient also weakens once OCR-mediated posts are removed. The right conclusion is therefore not that OCR explains away the substantive findings, but that the feminist-lineage result should be reported as partly format-sensitive, whereas Red Chamber suoyin is robust to this searchability check.

The sensitivity plots should be read as a robustness map rather than as nine competing specifications. Because the plots are on an odds-ratio scale, intervals that cross the OR = 1 line are not statistically distinguishable from no adjusted association at the conventional 95% level. Engagement controls do not drive the main finding: feminist-lineage direction, Red Chamber suoyin carrier, and medium OCR legibility stay to the right of 1 when engagement is excluded, although the medium-OCR estimate is smaller after final message-location correction. Dropping the Hong–Kangxi carrier does not materially change the Red Chamber or feminist-lineage estimates, which lowers the concern that the HKX carrier is simply absorbing all event-material structure. Raw LLM direction variants are most informative for anti-Manchu/Qing: the estimate moves across union and intersection definitions and its interval crosses 1 in the main final-label specifications, consistent with the measurement concern that raw LLMs over-call this boundary-heavy label. Carrier variants are more reassuring. Red Chamber suoyin remains large and to the right of 1 under GPT-default, union, intersection, and DeepSeek carrier definitions, so the carrier result is not an artifact of choosing one carrier source.

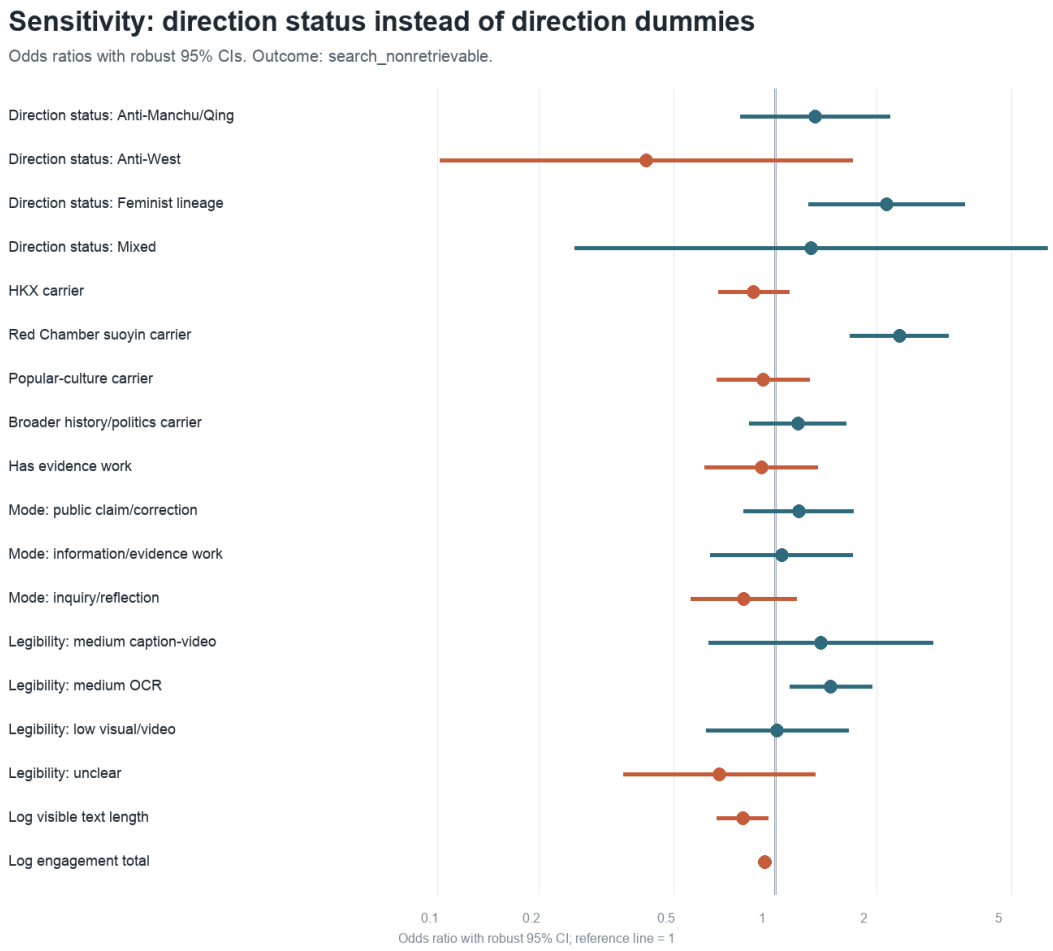

Figure 25: Sensitivity model using categorical direction status instead of direction dummies, plotted as odds ratios. The vertical reference line is OR = 1.

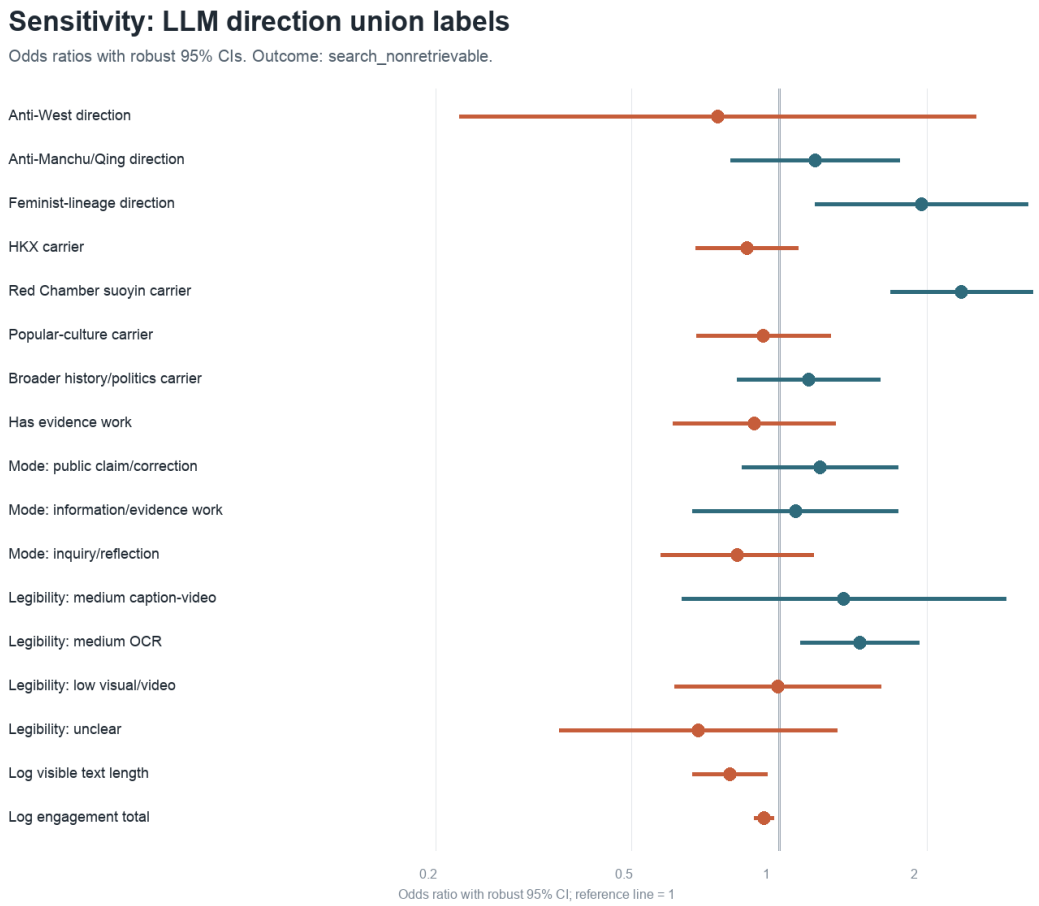

Figure 26: Sensitivity model using raw LLM direction union labels, plotted as odds ratios. The vertical reference line is OR = 1.

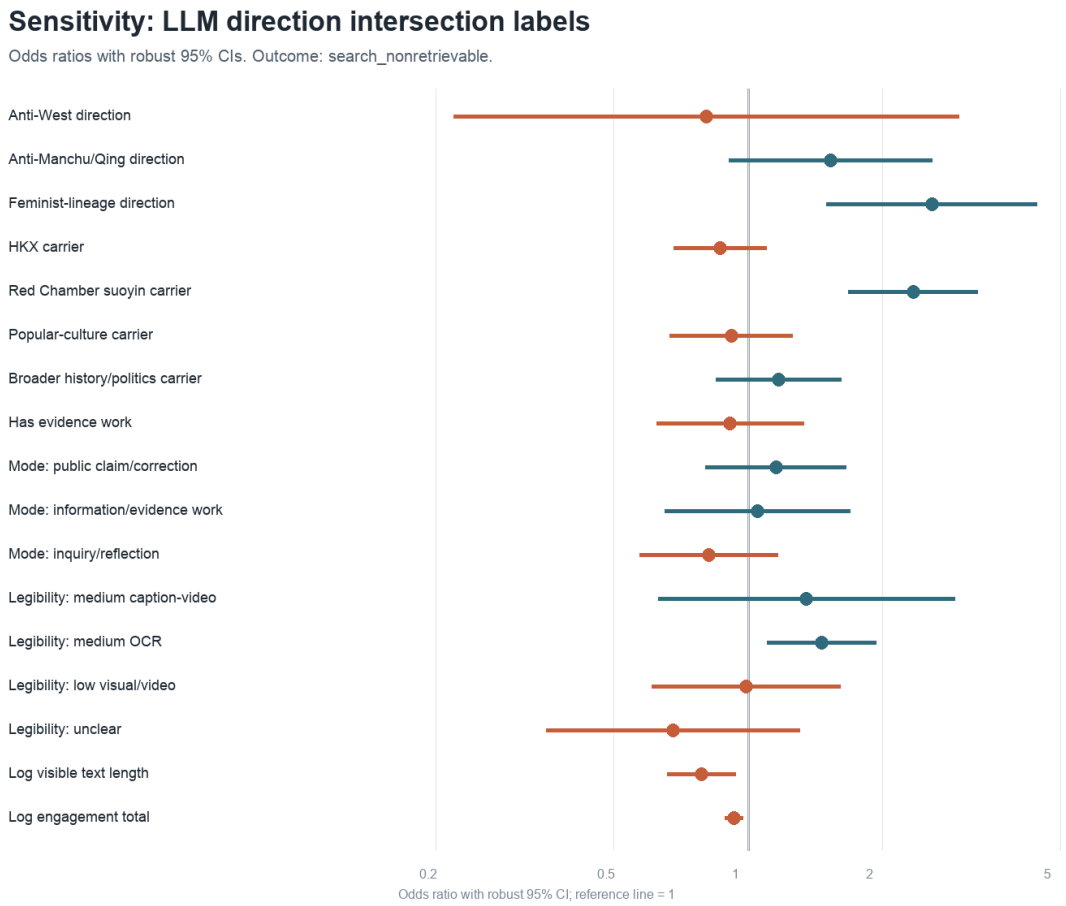

Figure 27: Sensitivity model using raw LLM direction intersection labels, plotted as odds ratios. The vertical reference line is OR = 1.

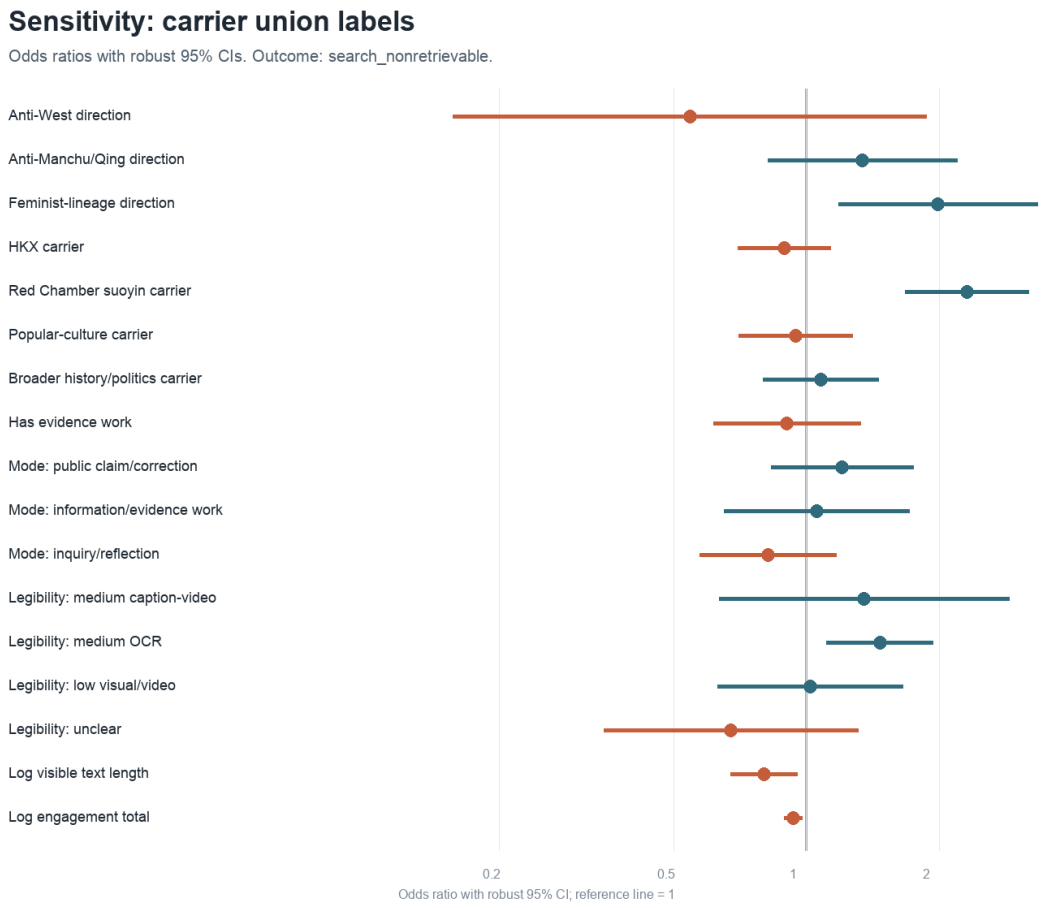

Figure 28: Sensitivity model using carrier union labels, plotted as odds ratios. The vertical reference line is OR = 1.

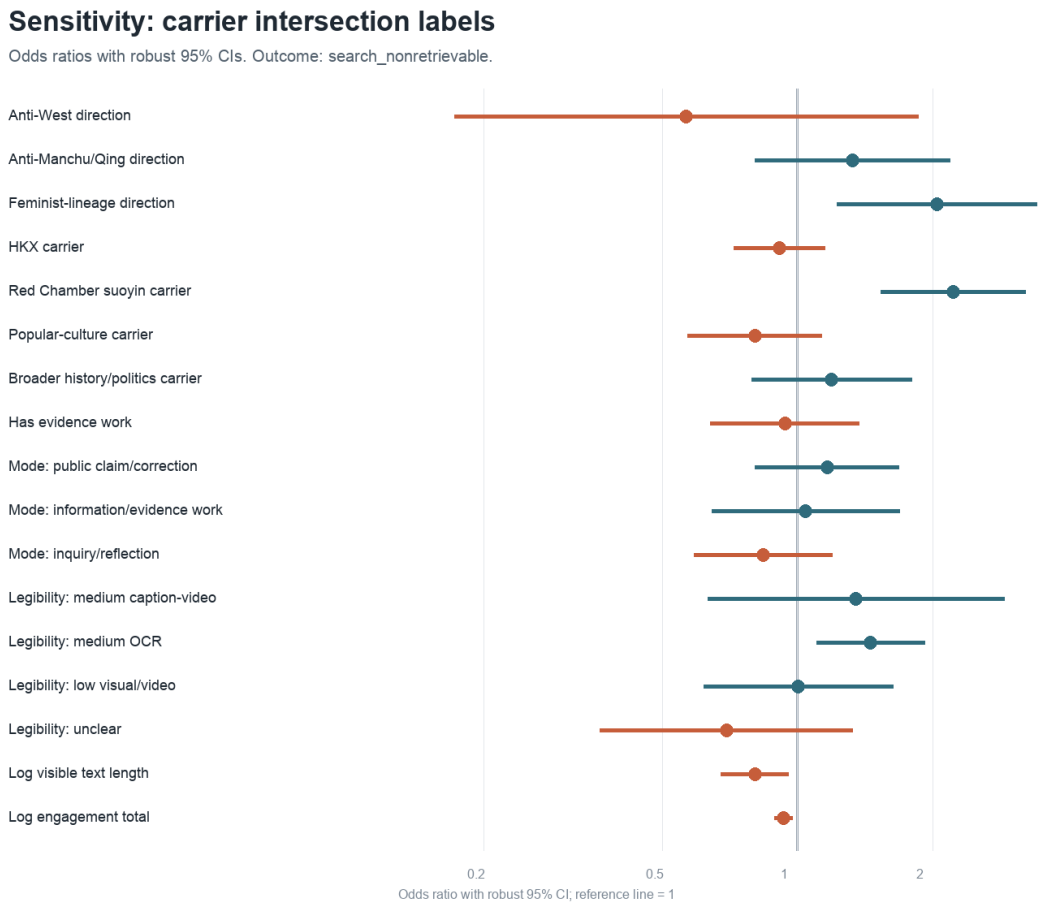

Figure 29: Sensitivity model using carrier intersection labels, plotted as odds ratios. The vertical reference line is OR = 1.

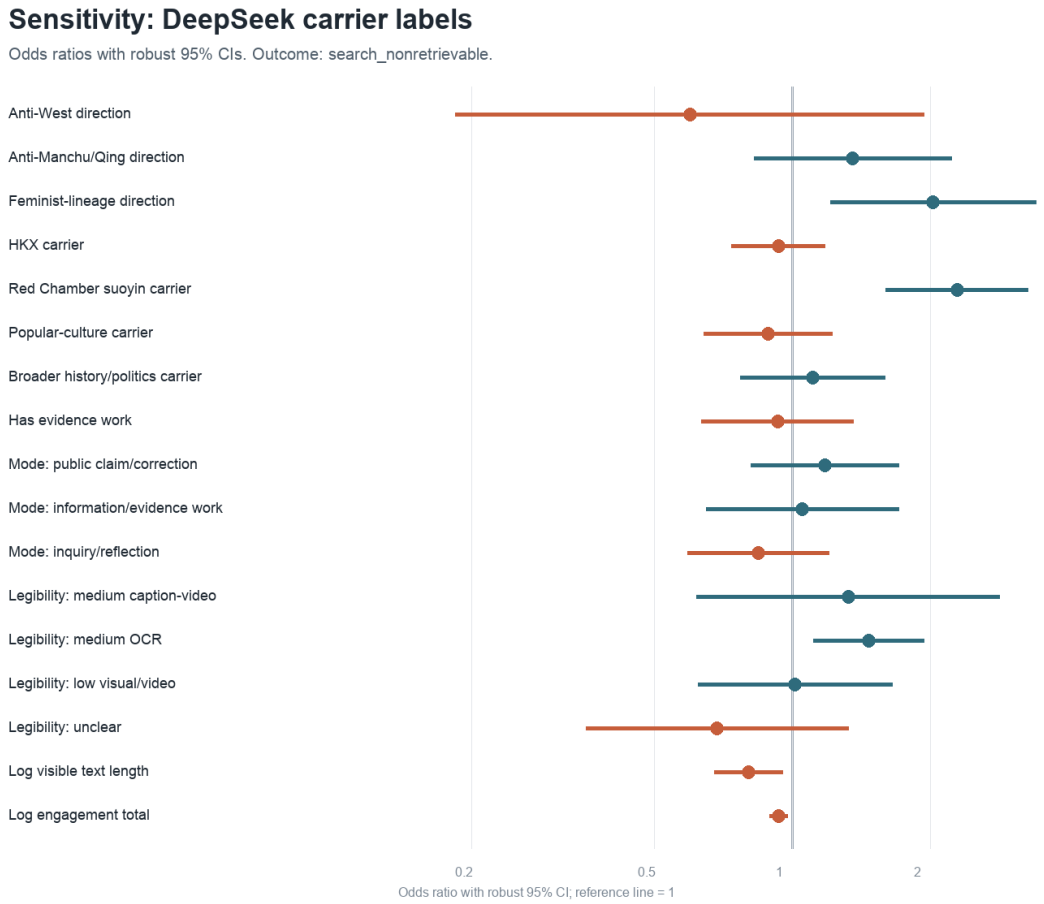

Figure 30: Sensitivity model using DeepSeek carrier labels, plotted as odds ratios. The vertical reference line is OR = 1.

C.9 Variable-selection rationale

Table 31: Fields downgraded to descriptive, qualitative, filter, or sensitivity use

Field	Reliability / support issue	Treatment
Detailed four-way `evidence_work`	Overall $\kappa = 0.685$ but macro F1 only 0.561; rare subtypes unstable	Exclude from main regression; use binary `has_evidence_work`; discuss detailed types qualitatively.
`claim_target`	H1–H2 $\kappa = 0.477$	Exclude from main regression; use only for descriptive audit or qualitative examples of first-order vs second-order claims.
Mode/content codable flags	Extreme class imbalance; near-all posts codable makes κ unstable or undefined	Use only as sample filters; disclose excluded counts.
Thin event reference	Moderate reliability but very low prevalence and potential quasi-separation	Use descriptively; not a main IV.
Raw LLM direction labels	Direction errors are non-random, especially anti-Manchu/Qing over-call	Do not use as final main labels; retain for LLM-only sensitivity and error analysis.

Table 30: Regression variables retained and measurement rationale

Field	Role in regression	Reliability evidence	Decision rule
Anti-Manchu/Qing direction	Direction IV	H1–H2 $\kappa = 0.695$, F1 = 0.739; current LLM F1 only 0.533–0.568	Include with source-tracked final labels: human labels for queued posts, consensus negative otherwise; report LLM-only sensitivity because raw LLMs over-call this label.
Anti-West direction	Direction IV	H1–H2 $\kappa = 0.655$, F1 = 0.667; low support	Include with source-tracked final labels; report counts and wide intervals as low-power evidence.
Feminist-lineage direction	Direction IV	H1–H2 $\kappa = 0.936$, F1 = 0.943; robust DeepSeek F1 ≈ 0.98	Include with source-tracked final labels; production positives and mode-disagreement hidden positives are human-adjudicated.
`mode`	Mode control	H1–H2 $\kappa = 0.677$, macro F1 = 0.712; best robust LLM $\kappa = 0.682$	Include as categorical field; human-adjudicate two-model mode disagreements.
Red Chamber suoyin carrier	Carrier control	H1–H2 $\kappa = 0.803$, F1 = 0.841	Include; strongest carrier label.
Hong–Kangxi specific carrier	Carrier control	H1–H2 $\kappa = 0.445$, F1 = 0.735; H1 143 positives vs H2 102	Include as event-material control; report with/without sensitivity.
Popular-culture intertext carrier	Carrier control	H1–H2 $\kappa = 0.704$, F1 = 0.720; LLMs over-call	Include as control; preserve disagreement/provenance flags.
Broader history/politics carrier	Carrier control	H1–H2 $\kappa = 0.547$, F1 = 0.647	Include as control with sensitivity; not a central theory IV.
Has evidence work (binary)	Binary control	H1–H2 $\kappa = 0.713$; GPT-5.4 $\kappa = 0.785$	Include binary version only; detailed subtypes are qualitative.
Legibility and text-location fields	Platform-legibility controls	Derived after final reviewed `message_location` is fixed, using post type, title/body/OCR text, hashtag-cleaned text length, OCR quality, and video-observability flags; see Table 13.	Include as observability/format controls; do not interpret as political-content labels.
Metadata controls	Controls	Not content-coded	Include post type, date, keyword FE, text length, and engagement; report with/without engagement, engagement-bin controls, and low-engagement exclusions.

D Search-verification audit and reference-category visibility benchmarks

This appendix documents the descriptive visibility benchmark summarized in the main text. It is intentionally separate from the within-event regression. The benchmark uses all three cleaned corpus categories: Hong–Kangxi event posts, pure historical-analysis posts, and standalone Western pseudo-history posts. The latter two categories are reference baselines, not part of the 1,594-post analytic corpus used for the main regression.

D.1 Verification procedure and audit

Direct URL checks are not reliable for this project because XHS URLs contain expiring security tokens. The verification script therefore approximates a manual public-search workflow: it enters the first 20 characters of a post's title into XHS search, or the beginning of the post text when no title is available; waits for search results to render; and checks whether the original post's unique ID appears in the rendered DOM. A DOM match confirms public-search retrievability. A non-match indicates search non-retrievability under the query procedure, not necessarily hard deletion.

To calibrate the raw `likely_deleted` label, I randomly sampled 100 posts from the `likely_deleted` set and manually checked whether each post could be recovered through an identifiable author profile. The audit did not overwrite the production outcome; it characterizes what the search-absence label means.

Table 32: Random audit of 100 posts flagged `likely_deleted`

Audit outcome	Count	Share	Interpretation
Confirmed gone from identifiable author profile	32	32%	Search absence includes hard disappearance from the identifiable profile, although the audit cannot distinguish platform deletion from author deletion.
Profile-alive but search-invisible	27	27%	At least one-quarter of search-absent cases are still present on author profiles and should be interpreted as visibility friction rather than hard deletion.
Unverifiable author/post status	41	41%	The post was not retrievable through search, but the author profile could not be confidently identified, mainly because XHS display names are non-unique.

Note: The audit sample is restricted to posts already flagged `likely_deleted`; it is therefore not an estimate of the overall corpus deletion rate. All 100 audit cases were absent from direct title/text search by construction. The audit checked whether they could be recovered through identifiable author profiles.

D.2 November-only and all-period rates

The preferred descriptive benchmark compares categories within the November 2025 posting window. This reduces, but does not eliminate, temporal confounding. The denominator includes rows with `survival=error`, matching the corpus-level rate used in the audit summary; excluding error rows changes the November rates to approximately 51.5%, 45.3%, and 45.6%, respectively.

Table 33: November 2025 search-nonretrievable rates

Category	N	Vis.	Invis.	Err.	Rate, Wilson 95% CI
Hong–Kangxi	1,559	745	792	22	50.8% (48.3, 53.3)
Pure history	286	156	129	1	45.1% (39.4, 50.9)
Pseudo-history	83	43	36	4	43.4% (33.2, 54.1)

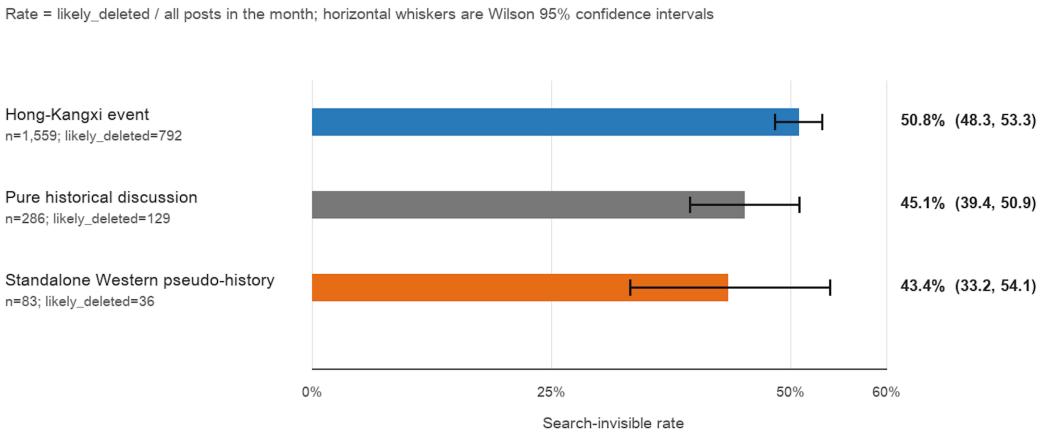

Figure 31: Search-nonretrievable rates by corpus category, November 2025. Error bars show Wilson 95% intervals.

Table 34: All-period rates through 2025-11-19

Category	N	Rate, Wilson 95% CI
Hong–Kangxi	1,594	50.4% (48.0, 52.9)
Pure history	503	42.7% (38.5, 47.1)
Pseudo-history	593	50.1% (46.1, 54.1)

E Limitations and future extensions

This project's main limitation is outcome ambiguity. `search_nonretrievable` can reflect hard platform deletion, author deletion, account suspension, search downranking, index decay, query drift, or account-identification noise. The audit shows that more than one mechanism is present. The correct current interpretation is therefore visibility friction under a public-search procedure, not censorship intent or confirmed deletion.

The second limitation is collection selection. The corpus contains posts that were searchable through 康熙瓜, 洪承畴, or 伪史 at collection time. It misses posts removed before collection, posts never indexed, posts that circulated through other terms, private or follower-only circulation, and comments. This means the analysis describes the retrievable event archive, not the complete social experience of the rumor.

The third limitation is measurement error in rare and boundary-heavy labels. Anti-West direction has only 14 final positives in the full corpus, so a wide interval cannot distinguish no association from low power. Anti-Manchu/Qing direction is theoretically central but empirically difficult because Hong–Kangxi paternity material alone is not anti-Manchu/Qing under the codebook. The current human-adjudication design protects against unverified model-positive direction calls, but it does not fully rule out shared two-model false negatives outside the queue. A thesis-stage random audit of non-queued posts should be added.

Carrier variables require especially cautious interpretation. Red Chamber suoyin is reliable and substantively interpretable, but Hong–Kangxi carrier and broader-history carrier are threshold-sensitive. They are best used as material controls and sensitivity checks. A strong carrier coefficient should be interpreted as a visibility association with a material pathway, not as evidence that the platform targets that carrier as a political category.

Legibility is also a limitation and a finding. OCR-mediated and video/image-primary posts do not have the same observable text as caption-primary posts. Medium OCR legibility is associated with higher search non-retrievability, but the corrected estimate is modest and could reflect platform indexing, verification difficulty, post format, image-text content, or moderation. The OCR-restricted sensitivity check helps bound this problem: Red Chamber remains robust after OCR-dependent posts are removed, while feminist-lineage remains positive but is less precise under the strict high-caption-only restriction. Future work should improve multimodal capture by saving screenshots, extracting all image panels, preserving video frames where feasible, and rerunning verification with multiple query strings.

Engagement controls are useful but risky. Engagement at collection may reflect earlier platform ranking and visibility; controlling for it can remove part of the outcome process. This memo therefore reports models with and without engagement. Future work should record engagement at repeated time points and model visibility as a longitudinal process rather than a single end-state.

Future extensions should prioritize five upgrades. First, repeat public-search verification over time and estimate survival or retrievability curves rather than one binary April 2026 status. Second, add stable account identifiers or profile snapshots so that search invisibility, profile disappearance, and author deletion can be separated more cleanly. Third, draw an independent random validation sample from non-queued posts and apply predicted-variable correction if unadjudicated LLM labels remain right-hand-side variables. Fourth, compare the Hong–Kangxi event with other historical-nationalist events to distinguish event-specific patterns from generic pseudo-history search friction. Fifth, combine the quantitative model with close qualitative readings of feminist-lineage and Red Chamber suoyin

cases, because those are the two strongest adjusted signals and likely contain different mechanisms.

References

Boym, Svetlana. 2001. *The Future of Nostalgia*. New York: Basic Books.

Brubaker, Rogers. 2004. *Ethnicity without Groups*. Cambridge, MA: Harvard University Press.

Carrico, Kevin. 2017. *The Great Han: Race, Nationalism, and Tradition in China Today*. Berkeley: University of California Press.

Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2024. "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses." Working paper.

Fang, Qixiang, Javier Garcia Bernardo, and Erik-Jan van Kesteren. n.d. "A Methodological Guide on Using Large Language Models for Text Annotation in the Social Sciences and Humanities with Python and R." Manuscript.

Fitzgerald, John. 1996. "The Nationless State: The Search for a Nation in Modern Chinese Nationalism." In *Chinese Nationalism*, edited by Jonathan Unger, 56–85. Armonk, NY: M.E. Sharpe.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3): 267–297.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2nd ed. Hoboken, NJ: Wiley.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2): 326–343.

Krebs, Ronald R., and Patrick Thaddeus Jackson. 2007. "Twisting Tongues and Twisting Arms: The Power of Political Rhetoric." *European Journal of International Relations* 13(1): 35–66.

Krippendorff, Klaus. 2018. *Content Analysis: An Introduction to Its Methodology*. 4th ed. Thousand Oaks, CA: SAGE.

Lorentzen, Peter. 2014. "China's Strategic Censorship." *American Journal of Political Science* 58(2): 402–414.

Mullaney, Thomas S. 2011. *Coming to Terms with the Nation: Ethnic Classification in Modern China*. Berkeley: University of California Press.

Roberts, Margaret E. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton, NJ: Princeton University Press.

Sun, Taiyi, and Quansheng Zhao. 2022. "Delegated Censorship: The Dynamic, Layered, and Multistage Information Control Regime in China." *Politics & Society* 50(2): 191–221.

Törnberg, Petter. 2024. "Best Practices for Text Annotation with Large Language Models." *Sociologica* 18(2): 67–85. doi:10.6092/issn.1971-8853/19461.

Wang, Zheng. 2012. *Never Forget National Humiliation: Historical Memory in Chinese Politics and Foreign Relations*. New York: Columbia University Press.

Wang, Zheng. 2017. *Finding Women in the State: A Socialist Feminist Revolution in the People's Republic of China, 1949–1964*. Berkeley: University of California Press.

Wimmer, Andreas. 2013. *Ethnic Boundary Making: Institutions, Power, Networks*. Oxford: Oxford University Press.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50(1): 237–291.