The Language Has Changed, But the Signal Remains

Register Mutation and Machine Blindness in Manosphere Reddit Discourse, 2012-2025

M6001 Capstone Academic Output

Student ID: 23441701514

Word count: 6,000 words

June 2026

Abstract

Between 2012 and 2025, Reddit communities tied to the manosphere changed in ways that are easy to miss. The blunt subcultural vocabulary of the early years has given way to the language of wellness, self-improvement, and personal accountability, but whether this reflects a genuine shift in ideology, or simply a more sophisticated way of concealing it, is far from clear. This study investigates that question and asks what it means for automated content detection.

Using an explanatory sequential mixed-methods design, this study analyses 17,950 post titles from four subreddits across three periods: Legacy (up to 2016), Pill (2017-2021), and Sigma (2022-2025). Jaccard distance and term frequency-inverse document frequency (TF-IDF) analysis track vocabulary changes, revealing a 63% turnover in surface terms. Ideological alignment with manosphere definitions, however, increased across both anchor sets (+0.037 on Ging anchors; +0.055 on Tanner & Gillardin anchors). Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analysis indicates that the emotional tone remained largely unchanged ($\varepsilon^2 = 0.0003$) despite the evolving vocabulary.

A temporal generalisation test shows that a Legacy-trained classifier performs less strongly on Sigma-era content than on its training data: F1 falls from 0.976 to 0.733, a 24.4-percentage-point train-to-test drop. However, cross-validation within the Legacy training set gives a lower baseline (CV F1 = 0.744), meaning the more conservative CV-to-test drop is only 1.2 percentage points. This qualifies rather than removes the detection-degradation claim: the model's strongest apparent failure is between Legacy calibration and Sigma-era transfer, while some weakness is already visible under within-era validation. The findings are brought together in a diagnostic framework for secondary school educators, developed in response to recognition gaps foregrounded by the Keeping Children Safe in Education (KCSiE) 2026 proposed revisions and Relationships, Sex and Health Education (RSHE) statutory guidance.

1. Introduction

Between 2012 and 2025, the online manosphere underwent a substantial and consequential linguistic transformation. A legible subcultural lexicon characterised the "Legacy era" (2012-2016), redpill, hypergamy, SMV, looksmaxxing, serving as explicit ideological markers legible to researchers, moderators, and practitioners alike precisely because the vocabulary was distinctive. By the "Sigma era" (2022–2025), this lexicon had been substantially displaced by the language of mainstream wellness and self-improvement: discipline, standards, mindset, accountability, monk mode. The register has migrated. Whether the underlying ideology has migrated with it, or whether linguistic change has made the same ideological content harder to detect, is the central question this study addresses.

Migration creates a three-domain recognition gap. In practice, this matters because schools, platforms, and safeguarding systems are increasingly asked to recognise misogynistic online influence at the same time as that influence becomes less visibly hostile at the lexical surface. Automated classifiers trained on Legacy-era data are calibrated in a language that no longer characterises what they are meant to detect. Teachers and safeguarding professionals trained to spot explicit misogynistic language are poorly positioned to recognise the same ideology dressed as self-improvement advice. Studies that document register shift rarely test whether detection performance degrades as a result; those that test detection rarely examine why classifications fail. This study spans all three gaps.

This study examines the subject through a longitudinal corpus analysis of 17,950 post titles from four Reddit communities, r/TheRedPill, r/PurplePillDebate, r/Seduction, and r/AllPillDebate, covering 2012 to 2025 across three analytical periods. The analysis is structured around three research questions:

RQ1. To what extent has the manosphere's core vocabulary shifted toward mainstream wellness and self-improvement registers, and is that shift manosphere-specific or a general property of Reddit discourse over time?

RQ2. Does lexical mutation accompany ideological change, or does ideological content persist beneath surface register shifts, and can computational methods distinguish between the two?

RQ3. How does register change affect automated detection performance, and what rhetorical mechanisms explain detection failure at the level of individual posts?

An explanatory sequential mixed-methods design (Creswell & Plano Clark, 2018) integrates four quantitative methods, Jaccard distance, TF-IDF discriminating vocabulary analysis, VADER sentiment analysis, and Sentence-BERT (SBERT) semantic similarity scoring, culminating in a temporal generalisation test, with qualitative content analysis of twelve purposively selected cases drawn directly from quantitative failure points. The quantitative strand establishes the scale and character of linguistic mutation; the qualitative strand identifies the rhetorical mechanisms behind it.

Three contributions follow. Empirically, the study finds that 63% vocabulary turnover coexists with increasing semantic proximity to academic definitions of manosphere ideology: surface and ideological change come apart. Conceptually, it contributes the term machine blindness to describe how automated systems can fail when ideological content persists beneath changing registers. Methodologically, it adapts temporal generalisation testing (King & Dehaene, 2014) to measure detection degradation under register shift, yielding a 24.4pp F1 drop and two mechanistically distinct failure modes. A practitioner-facing diagnostic framework addresses the recognition gap foregrounded by the KCSiE 2026 proposed revisions and RSHE statutory guidance.

2. Background

Computational social scientists rarely draw on speech-act theory; gender studies scholars rarely evaluate classifier performance; platform governance researchers rarely inspect the linguistic mechanisms behind detection failures. These are not arbitrary omissions; they reflect different disciplinary epistemologies and publication cultures. This study's contribution sits precisely in the space between them. For computational social science, that means taking seriously the pragmatics of language use rather than treating texts as unordered token sequences. For gender and masculinity studies, it means engaging more directly with the affordances and limits of automated detection when theorising how misogyny circulates online. For platform governance research, it means grounding regulatory debates in empirical evidence about how linguistic and ideological change interact in real-world moderation systems.

This paper extends three strands of prior work: large-scale longitudinal manosphere research (Ribeiro et al., 2021), computational-sociolinguistic studies of jargon and embeddings (Farrell et al., 2020), and lexicon-first analyses of coded misogynistic cryptolects (Gothard et al., 2021). Rather than measuring community migration or extracting stable jargon, it asks what happens when ideologically salient language becomes less jargon-like and more compatible with ordinary self-improvement and advice registers.

2.1 The Manosphere: Taxonomy and Sigma-Era Evolution

Scholarly accounts of the manosphere generally begin with Ging (2019), whose typology remains the standard reference point. She identifies four broad strands: men's rights activists (MRAs), Men Going Their Own Way (MGTOW), pick-up artists (PUAs), and incels. The communities are loosely affiliated and organised around shared anti-feminist ideology, expressed through a subcultural lexicon that was, at least in this early period, distinctive. Gynocentrism, hypergamy, looksmaxxing, these terms did not circulate outside the communities generating them, which is part of what made legacy-era discourse legible to researchers and moderators.

That has changed. Tanner and Gillardin (2025) document the emergence of what they call the 'neo-manosphere,' in which overt adversarial anti-feminism is displaced by something that looks, on the surface, considerably more like ordinary self-improvement content. The sigma male archetype sits at the centre of this. Positioned outside the conventional alpha/beta hierarchy and defined by self-sufficiency and emotional detachment, it draws heavily on wellness culture, Stoic philosophy, and financial independence discourse.

The theoretical tools for thinking about this come from Bridges and Pascoe (2014), whose account of hybrid masculinities describes how gender formations can strategically incorporate progressive or subordinated masculine identities to reproduce hegemonic gender structures while insulating them from critique. Connell and Messerschmidt (2005) provide the underlying framework. Sigma-era wellness discourse does exactly what hybrid masculinity theory predicts: borrowing the vocabulary of self-determination and personal responsibility to repackage hierarchical gender logics as aspiration rather than ideology. Krendel's (2020) corpus-assisted discourse analysis of manosphere lexis provides a methodological precedent for the TF-IDF approach used here.

Haslop et al. (2024) push this further into the attention economy, showing through their analysis of the 'Andrew Tate effect' that influencer masculinity is built on the gap between emotional register and ideological payload. Content that reads motivational encodes misogynistic premises, and that gap is not a flaw in the content; it is the point. What Evans and Ringrose (2025) add is that this does not stay online: alpha/beta hierarchies and gendered performance norms are showing up in physical peer cultures, carried there by short-form social media, which is part of why the register shift this study documents matters beyond the platforms where it originates.

Survey evidence links exposure to Tate-affiliated content with controlling relationship attitudes (Women's Aid, 2023) and positive views of his misogynistic positions among boys aged 6–15 (YouGov, 2023). Pinkett (2026) documents how this operates practically: the ideology arrives not as ideology, but as advice about discipline, standards, and self-improvement, making recognition difficult for adults in educational settings.

Austin’s distinction between locutionary and illocutionary acts provides the analytical mechanism for this, grounding the claim that issuing certain utterances is itself a kind of action rather than merely describing one (Austin, 1962, pp. 6–7, 100–101). Searle developed the framework systematically as a theory of rule-governed linguistic behaviour, and Skinner extended speech-act analysis to the contextual interpretation of ideological discourse (Searle, 1969; Skinner, 2002, pp. 103–127). Surface register and illocutionary force can diverge substantially: a phrase may be propositionally neutral while illocutionarily performing a rejection of women as legitimate relationship partners. The ideology is in the doing, not the saying. This is what this study calls structural detection failure, and it runs through the qualitative analysis in Section 4.

2.2 Computational Approaches to Detection Failure

The computational literature on hate speech detection has documented cross-register failure, though the temporal dimension is less well studied. Davidson et al. (2017) found that classifiers misclassify sexist content as merely offensive rather than ideological, anticipating the degradation examined here. Aggarwal and Stevenson (2024) find that manospheric authors maintain a recognisable linguistic style across subreddits even as terminology shifts, suggesting detectable continuity persists through register migration. de Kock (2024a, 2024b) uses dynamic words and user embeddings to show that sub-community surface trajectories

diverge while underlying gendered orientation stays consistent, supporting the treatment of Legacy, Pill, and Sigma as distinct eras within one ideological field.

Ushio and Camacho-Collados (2024) provide the most directly relevant empirical baseline. They demonstrate that hate speech detection degrades consistently with temporal distance from training data, while sentiment classification stays stable. For trope-level classification, the case for large language models is made by Törnberg (2024) and Ziems et al. (2024), both of whom find LLMs outperform fine-tuned classifiers on low-resource annotation tasks with limited historical training data. That is the methodological situation here, and it is why this study uses a zero-shot Llama approach.

The literature does not distinguish temporal failure, a calibration problem arising when classifier training language has evolved, from structural failure, which is more fundamental: ideology expressed illocutionarily evades detection regardless of training recency. Treating these as variations of the same problem produces solutions that fix one while leaving the other.

Machine blindness, as used here, is therefore narrower than general concept drift and broader than adversarial language. It names a failure mode specific to ideological discourse, where the central difficulty lies in recovering illocutionary meaning rather than tracking distributional change or intentional evasion. This focus on the pragmatic level of ideology is what distinguishes machine blindness from existing taxonomies of temporal degradation in Natural Language Processing (NLP).

2.3 Platform Governance and the Detection Gap

Gillespie (2018) frames platform moderation as a form of private governance, which helps explain why detection failures follow systematic patterns rather than appearing randomly. Automated systems are built at a moment in time; they encode assumptions about what harmful content looks like at that moment. As language evolves, those assumptions drift out of alignment with what the systems are supposed to catch, a governance failure rooted in the mismatch between how platforms invest in moderation and how quickly linguistic landscapes move.

Marwick and Lewis (2017) document the deliberate side of this: ironic, coded, and deniable language is used strategically to stay below moderation thresholds. Sigma wellness discourse is an advanced instance of this, achieving a kind of strategic invisibility that reduces friction with automated systems without requiring the underlying ideological content to change.

Ofcom's (2024) illegal content risk assessment framework requires platforms to update risk assessments as language evolves. This study provides evidence that the obligation is structurally difficult to meet when linguistic mutation outpaces retraining cycles.

3. Methods

3.1 Research Design

This study employs an explanatory sequential mixed-methods design (Creswell & Plano Clark, 2018) situated within a critical realist paradigm, which distinguishes relatively stable ideological structures (the real) from the events and experiences they generate (the actual and the empirical) (Bhaskar, 1978, pp. 46-48). Manosphere ideology, the real, is operationalised through a 66-statement anchor taxonomy (Ging, 2019; Tanner & Gillardin, 2025) remapped onto four tropes: Rejection, Crisis, Hate, and Prescription; post titles across three eras constitute the actual. The framework treats surface vocabulary mutation and ideological persistence as the central claim the empirical analysis tests rather than assumes. The key conceptual problem is therefore not only that language changes, but that the relationship between lexical surface and ideological function becomes unstable. The design is integrative rather than additive: each strand addresses a different level of explanation required by the research questions.

3.2 Corpus

The corpus comprises 17,950 post titles from four manosphere-adjacent subreddits. Inclusion required two or more of: (1) explicit self-identification with manosphere discourse; (2) documented community overlap; (3) ideological focus on gendered grievance as a primary organisational theme; (4) prior administrative action (quarantine or ban) by Reddit.

Table 1. Manosphere corpus composition by subreddit and era

Subreddit	Legacy (≤2016)	Pill (2017–21)	Sigma (2022–25)	Total
r/TheRedPill	1,973	2,349	1,388	5,710
r/PurplePillDebate	1,752	1,982	1,519	5,253
r/Seduction	2,326	2,230	1,490	6,046
r/AllPillDebate	N/A	N/A	942	942
Total	6,051	6,561	5,339	17,950
Control corpora	300	N/A	300	600

Note. AllPillDebate is absent from the Legacy and Pill eras (founded 2022). Control corpora: r/AskMen and r/selfimprovement. N = 17,950 manosphere posts total.

r/PurplePillDebate is included as a site of linguistic contact between manosphere and mainstream registers, a high-volume location where manosphere vocabulary intersects with counter-discourse, rather than as a pure advocacy community. r/AllPillDebate was founded in approximately 2021, contributing posts only to the Sigma era (n = 942; 17.6% of Sigma titles), introducing a potential era-specific confound addressed in the limitations section.

3.3 Quantitative Pipeline

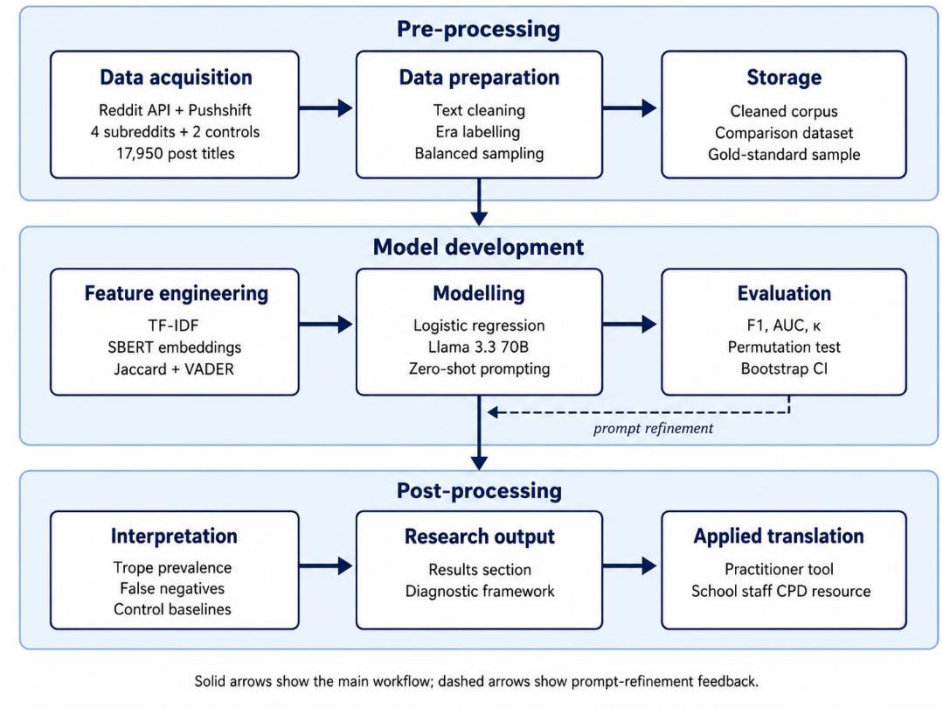

Figure 1. Mixed-methods pipeline for computational discourse analysis. The study proceeds through three linked stages: quantitative description of register change, quantitative detection testing, and qualitative explanation of classification failure. The design is sequential and integrative: quantitative outputs identify patterns and failure cases, while qualitative analysis explains the mechanisms behind those failures.

Titles were lowercased, tokenised, and stripped of standard English stopwords. Lemmatisation was deliberately avoided to preserve subcultural morphological forms (looksmaxxing, alphafuxx) that carry ideological signal and would be normalised away by a standard lemmatiser; this choice is applied consistently across all token-based analyses. SBERT operates at the sentence level and is unaffected by this decision. Avoiding lemmatisation may inflate lexical turnover by preserving morphological variants, but this is retained because such variants are themselves part of the subcultural register under investigation.

Lexical analysis. Jaccard distance (1 minus the proportion of shared vocabulary) measures vocabulary turnover between era pairs. TF-IDF then identifies which specific terms characterise each era's register, establishing the destination of change that Jaccard distance, as a scalar measure, cannot capture.

Sentiment analysis. VADER compound scores test affective-tone stability across eras. Because sentiment scores violated normality assumptions, Kruskal-Wallis is used in place of ANOVA. VADER's reliance on a fixed lexicon may produce systematic underperformance on short or ironic titles; however, the finding of interest is cross-era stability rather than absolute level.

Semantic similarity. Sentence-BERT (all-MiniLM-L6-v2) maps post titles to the 66-statement anchor taxonomy: 33 statements from Ging (2019) and 33 from Tanner & Gillardin (2025). Cosine similarity scores measure semantic proximity to academic definitions of manosphere tropes. Endorsement cannot be inferred from surface similarity alone.

Temporal generalisation. Following King and Dehaene (2014), a logistic regression classifier trained on TF-IDF features from Legacy-era titles is evaluated on Sigma-era titles. Two figures are reported: the train-to-test drop, which captures how far performance falls from the fitted Legacy classifier to Sigma-era transfer, and the more conservative CV-to-test comparison, which checks whether the apparent transfer gap remains after accounting for within-era generalisation.

3.4 Trope Classification

Llama 3.3 70B Versatile (Groq API, temperature = 0) operates as a zero-shot classifier. Zero-shot is preferred in low-resource settings with limited historical training data (Törnberg, 2024). Prompts follow the MinZero/MaxZero framework (Bunt et al., 2025), varying definitional constraints on a 35-post development set without leaking test data.

Evaluation uses a 65-post held-out test set drawn from a 100-post classifier gold standard, following a 35-post development set. A separate 12-case qualitative sample was used for mechanism analysis, giving 112 coded cases across the validation and qualitative phases. Following Song et al. (2020), single-coder annotation introduces systematic bias into performance estimates: κ values measure classifier-coder agreement rather than intercoder reliability, and F1 scores indicate relative classifier difficulty across tropes rather than absolute accuracy benchmarks.

Table 2. LLM trope classifier performance (Llama 3.3 70B; Groq API; temperature = 0; n = 65 test posts)

Trope	F1	Cohen's κ	Reliability	Interpretation
Prescription	0.842	0.777	Substantial	Validated; retained for prevalence analysis
Rejection	0.769	0.617	Moderate	Retained cautiously; interpreted as indicative
Crisis	0.400	0.377	Weak	Excluded from prevalence analysis
Hate	0.471	0.399	Weak	Excluded from prevalence analysis

Note. κ values report classifier-coder agreement on a single-coder gold standard, not intercoder reliability (Song et al., 2020). Prescription is treated as validated. Rejection is retained cautiously because its F1 is acceptable but κ remains below the stronger 0.65 threshold used for robust prevalence claims. Crisis and Hate are excluded from prevalence interpretation.

3.5 Qualitative Phase

The qualitative phase is not intended to generate thematic generalisations about manosphere discourse; it is designed to identify mechanism-level explanations for quantitatively observed classification failures. Twelve cases were purposively selected from three quantitatively identified regions: (1) false negatives from the temporal generalisation test; (2) high-SBERT/neutral-VADER outliers; and (3) collocate-shift pairs, instances where a key term's surrounding context migrates substantially across eras. The objective was analytical variation across failure types rather than representativeness of the broader corpus.

Each case is analysed in two sequential passes: Pass 1 applies the locutionary/illocutionary distinction (Austin, 1962; Searle, 1969; Skinner, 2002) to identify why the classifier misses a post whose ideological function is evident to a human reader; Pass 2 applies reflexive thematic analysis (Braun & Clarke, 2006) as a separate subsequent pass, preventing deductive speech-act categories from collapsing the inductive themes.

Table 3. Qualitative case corpus (n = 12): metadata and coding summary

ID	Era	Subreddit	Trope	Selection	Theme	SBERT	VADER
1	Legacy	r/TheRedPill	Rejection	False Negative	Theme 2: Fairness inversion	0.561	−0.026
2	Pill	r/PurplePillDebate	Crisis	False Negative	Theme 1: Epistemic authority	0.759	0.000
3	Legacy	r/PurplePillDebate	Hate	False Negative	Theme 1: Epistemic authority	0.736	−0.026
4	Pill	r/TheRedPill	Prescription	False Negative	Theme 3: Softened prescription	0.671	0.000
5	Legacy	r/PurplePillDebate	Rejection	Affective Camouflage	Theme 2: Fairness inversion	0.768	0.000
6	Pill	r/PurplePillDebate	Prescription	Affective Camouflage	Theme 3: Softened prescription	0.825	0.000
7	Pill	r/PurplePillDebate	Crisis	Affective Camouflage	Theme 1: Epistemic authority	0.762	0.000
8	Sigma	r/PurplePillDebate	Hate	Affective Camouflage	Theme 2: Fairness inversion	0.812	0.000
9	Legacy	r/TheRedPill	Prescription	Collocate Shift	Theme 3: Softened prescription	0.590	0.000

10	Sigma	r/TheRedPill	Prescription	Collocate Shift	Theme 3: Softened prescription	0.550	0.000
11	Legacy	r/PurplePillDebate	Crisis	Collocate Shift	Theme 1: Epistemic authority	0.601	0.052
12	Pill	r/PurplePillDebate	Rejection	Collocate Shift	Theme 3: Softened prescription	0.601	0.052

Note. FN = false negative; AC = affective camouflage; CS = collocate-shift pair. SBERT = cosine similarity to Rejection anchor set.

3.6 Integration Strategy

The quantitative and qualitative strands are integrated by design, not by juxtaposition. In mixed-methods terms, this is a connecting strategy: quantitative outputs directly determine which posts enter qualitative analysis, and the three selection criteria map precisely onto the two failure modes the study seeks to explain: temporal failure (the classifier cannot generalise across registers) and structural failure (the ideology is carried illocutionarily and therefore evades surface-based detection in any era). The integration point is therefore procedural as well as interpretive: classifier failures generate the qualitative case sample, and qualitative analysis explains the mechanisms that produced those failures. The qualitative strand does not illustrate quantitative findings; it provides the mechanistic account of why those findings arise.

Ethics. The study uses publicly available Reddit data with all identifiers removed prior to analysis. Example phrases are included in practitioner-facing materials only where analytically necessary and are lightly paraphrased to prevent direct attribution or searchability.

4. Findings

The findings are organised as a staged argument rather than by numerical research-question order. Section 4.1 establishes whether register change is directional rather than merely large; Section 4.2 tests whether that change produces detection degradation; and Section 4.3 uses qualitative analysis to explain the rhetorical mechanisms behind that degradation.

4.1 Register Destination, Not Volume (RQ1)

Jaccard distance analysis shows substantial vocabulary turnover across eras, with the Legacy to Sigma distance of 0.6276, approximately 63% turnover, representing the largest inter-era gap in the manosphere corpus. However, control communities also show high Legacy to Sigma turnover (Jaccard distance = 0.7751), exceeding the manosphere value by 0.1475. The scale of lexical change over thirteen years is therefore partly a general property of Reddit discourse rather than a manosphere-specific phenomenon. The mutation argument cannot rest on volume of change alone; it must rest on direction.

TF-IDF indicates the direction of lexical change. Legacy-era titles are dominated by explicit subcultural markers, trp (Rank 2), pill (Rank 4), red (Rank 5), fr (Rank 6), game (Rank 8), which are entirely absent from the Pill and Sigma top 15. By the Sigma era, the highest-ranked terms are generic: women, men, dating, girls, sex, want, vocabulary indistinguishable at the surface level from mainstream relationships discourse. Control communities acquire none of

these markers in any era, consistent with their disappearance from Sigma manosphere titles reflecting manosphere-specific register change rather than general Reddit drift.

Table 4. TF-IDF discriminating vocabulary by era, top 10 terms

Rank	Legacy (≤2016)	Pill (2017–21)	Sigma (2022–25)
1	trp	women	women
2	women	men	men
3	men	girl	dating
4	pill	sex	get
5	red	get	girls
6	fr	girls	sex
7	girl	want	want
8	game	dating	man
9	like	woman	woman
10	get	man	girl

Note. Legacy terms in ranks 1, 4–6, 8 (trp, pill, red, fr, game) are explicit subcultural markers absent from Pill and Sigma top-10, evidencing vocabulary shedding. N = 17,950 posts.

SBERT similarity to both anchor sets increases monotonically across all three eras. For Ging (2019) anchors: Legacy = 0.3483, Pill = 0.3581, Sigma = 0.3843. For Tanner & Gillardin (2025) anchors: Legacy = 0.3013, Pill = 0.3284, Sigma = 0.3559. If register mutation were accompanied by ideological dilution, similarity to manosphere-relevant anchor statements would be expected to decrease; instead, it increases across both taxonomies.

The absolute SBERT shifts are small in cosine terms and should not be read as large semantic transformations at the individual-post level. Their evidential value lies in their consistency across both independent anchor sets and their convergence with TF-IDF, VADER, and qualitative failure-case evidence.

A construct validity comparison with r/ExRedPill is critical to interpreting this finding. That community, which exists to critique and deconstruct manosphere ideas, scores comparably to the manosphere corpus on Ging anchors in Pill and Sigma eras. High SBERT similarity on the Ging instrument therefore reflects articulation of the ideological categories, engaging with them, referencing them, debating them, regardless of whether that engagement is affirmative or critical. The T&G instrument, built on Sigma-era TikTok language less familiar to ExRedPill members, provides better discrimination: the manosphere corpus outscores ExRedPill across all eras.

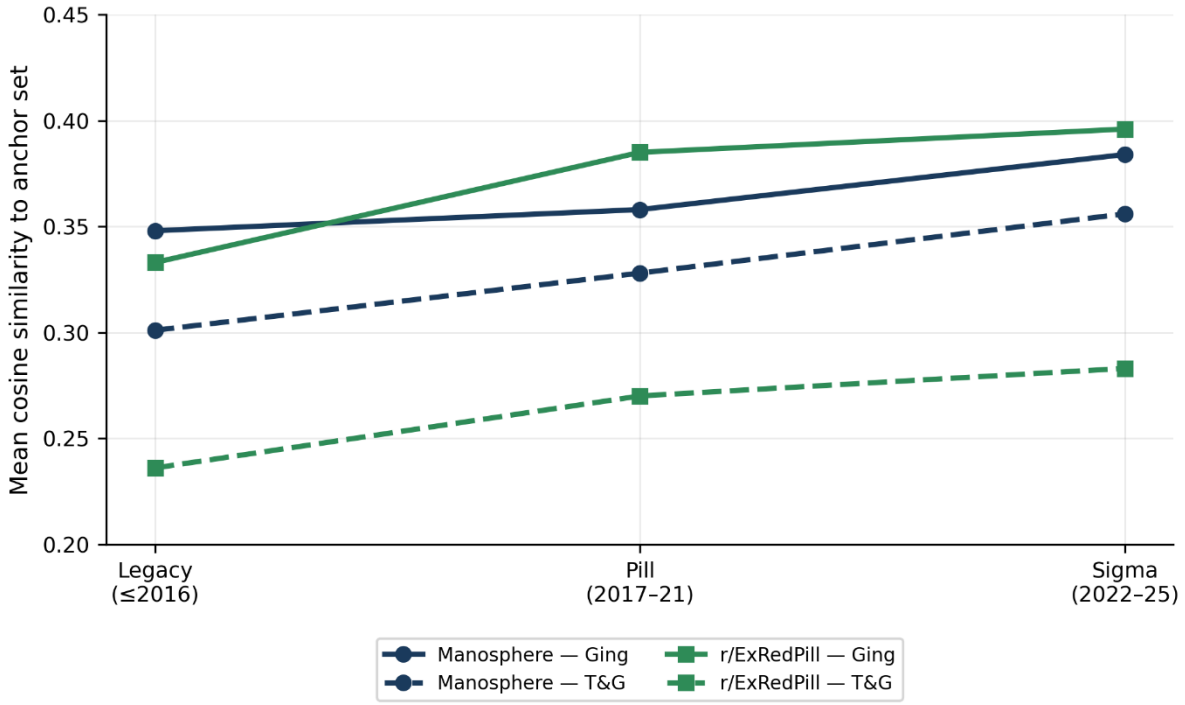

Figure 2. Construct validity: SBERT cosine similarity by corpus and era. The construct-validity comparison shows that SBERT similarity captures engagement with manosphere ideological categories, but not endorsement on its own. Tanner and Gillardin anchors discriminate more consistently between manosphere and r/ExRedPill content than Ging anchors, supporting their use for analysing Sigma-era discourse.

Note. Solid lines represent Ging anchors; dashed lines represent Tanner and Gillardin anchors. r/ExRedPill is used as a critical comparison corpus because it discusses manosphere concepts without necessarily endorsing them.

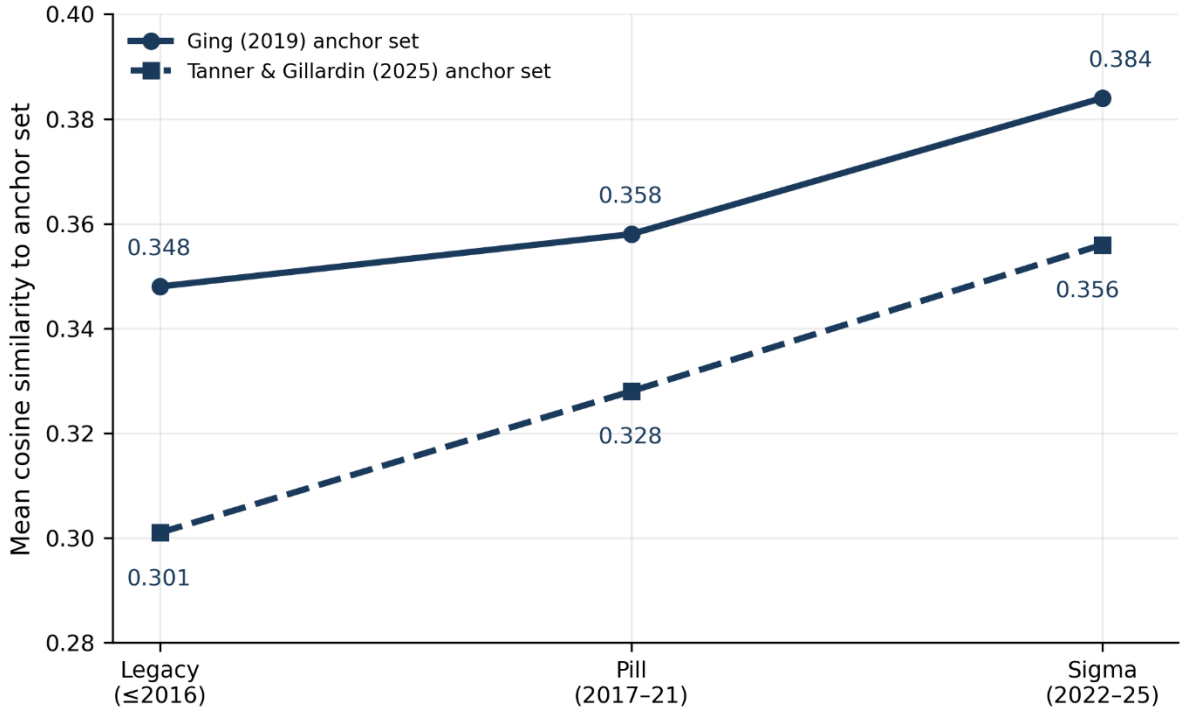

Figure 3. Mean SBERT cosine similarity across three eras. Mean semantic similarity to both anchor sets increases from Legacy to Sigma eras, despite substantial vocabulary turnover. This pattern suggests that surface lexical change does not correspond to ideological dilution; instead, Sigma-era content remains semantically aligned with manosphere ideological categories.

Note. SBERT similarity indicates proximity to the anchor taxonomy, not endorsement. The result should therefore be interpreted as evidence of ideological articulation rather than direct attitudinal commitment.

VADER sentiment remains negligibly stable across eras ($H = 6.92$, $p = 0.0315$, $\varepsilon^2 = 0.0003$). The effect size is substantively indistinguishable from zero, era accounts for 0.03% of variance, making this a positive finding rather than a null result. Content moderation systems that use negative affect as a proxy for harm would detect no meaningful change across the study window despite the substantial register migration documented above.

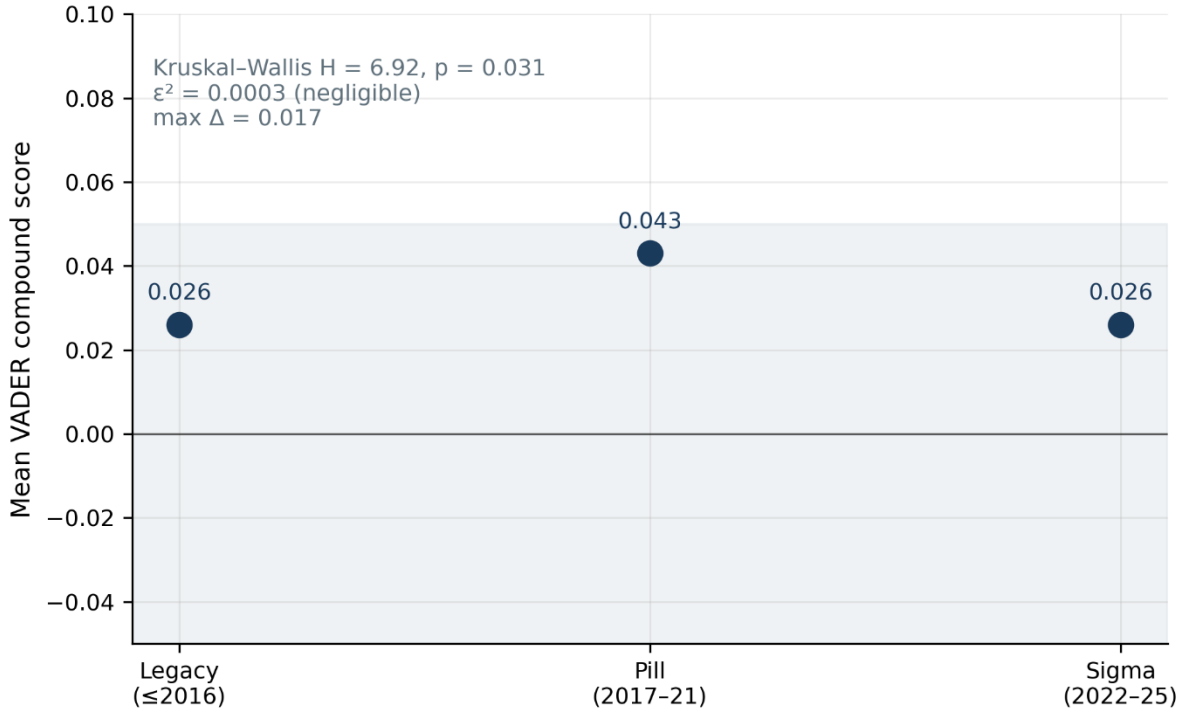

Figure 4. Mean VADER compound sentiment score by era. Mean sentiment remains practically stable across the study period, with all three eras sitting close to the neutral range. Although the cross-era difference is statistically detectable, the effect size is negligible, indicating that affective tone changes very little as vocabulary migrates.

Note. The shaded band marks the neutral affective range. Kruskal-Wallis $H = 6.92$, $p = 0.031$, $\varepsilon^2 = 0.0003$; the maximum cross-era difference is 0.017 compound-score points.

4.2 Detection Degradation (RQ3)

Temporal failure refers to performance degradation caused by register change across time. Structural failure refers to the limits of surface-level detection when ideology is communicated through implication, presupposition, or illocutionary force rather than explicit lexical markers.

A Legacy-trained logistic regression classifier achieves strong in-sample performance (train F1 = 0.976) but generalises less strongly to Sigma-era titles (test F1 = 0.733; test AUC = 0.702). The train-to-test F1 drop is 24.4 percentage points, with a bootstrap 95% confidence interval of [22.0, 27.0] percentage points. Cross-validation within the Legacy training set gives a lower baseline (CV F1 = 0.744), so the conservative CV-to-test drop is only 1.2 percentage points and its confidence interval crosses zero. The result should therefore be interpreted cautiously: the model's fitted Legacy boundary transfers poorly to Sigma-era content, but some of the apparent degradation reflects limited within-era generalisation rather than temporal shift alone. A symmetric test, training on Sigma content and evaluating on Legacy posts, produces a smaller but still visible drop of 14.4 percentage points (test AUC = 0.747).

Table 5. Temporal generalisation classifier performance: main and symmetric tests

Direction	Train F1	CV F1	Test F1	F1 drop (pp)	Test AUC
Legacy to Sigma (main)	0.976	0.744	0.733	24.4	0.702
Sigma to Legacy (symmetric)	0.854	N/A	0.710	14.4	0.747

Note. The Legacy to Sigma confidence interval refers to the train-to-test F1 drop, estimated by bootstrap resampling of the Sigma test set. The conservative CV-to-test comparison is much smaller: 1.2 percentage points, 95% CI [-1.2, 3.8]. The symmetric Sigma to Legacy test is reported descriptively as a robustness check.

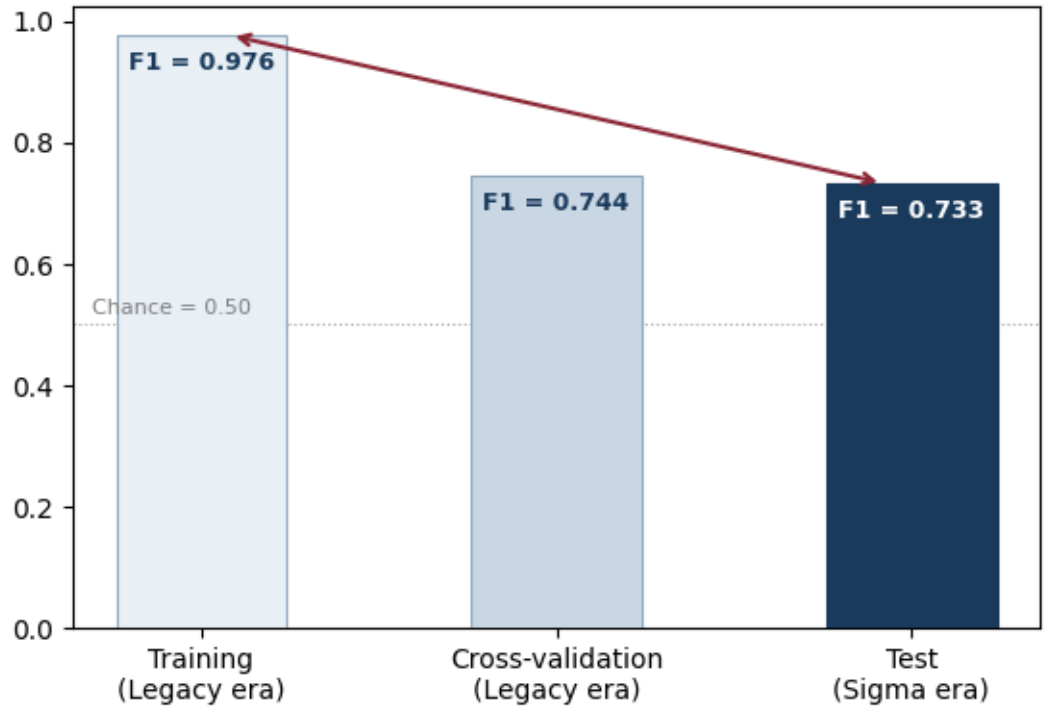

Figure 5. Temporal generalisation performance across training, cross-validation, and Sigma-era test conditions. The fitted Legacy-era classifier performs strongly on training data, but performance is lower under cross-validation and remains lower when applied to Sigma-era content. The train-to-test F1 drop

is 24.4 percentage points, while the CV-to-test comparison is much smaller, indicating that the temporal transfer finding should be interpreted cautiously.

Note. Train F1 = 0.976; CV F1 = 0.744; Sigma test F1 = 0.733. Bootstrap 95% CI for the train-to-test drop: [22.0, 27.0] percentage points. Test AUC = 0.702.

Feature-weight analysis identifies the mechanism driving temporal failure. The classifier has learned to treat interrogative and advice-seeking syntactic forms, how, self, are you, as strong indicators of non-ideological content, reflecting their prevalence in r/AskMen and r/selfimprovement. As Sigma-era manosphere titles increasingly adopt the same conversational, advice-seeking framing, they are misread as control content even when carrying manosphere ideology. Register migration at the level of syntactic form, not just vocabulary, produces degradation.

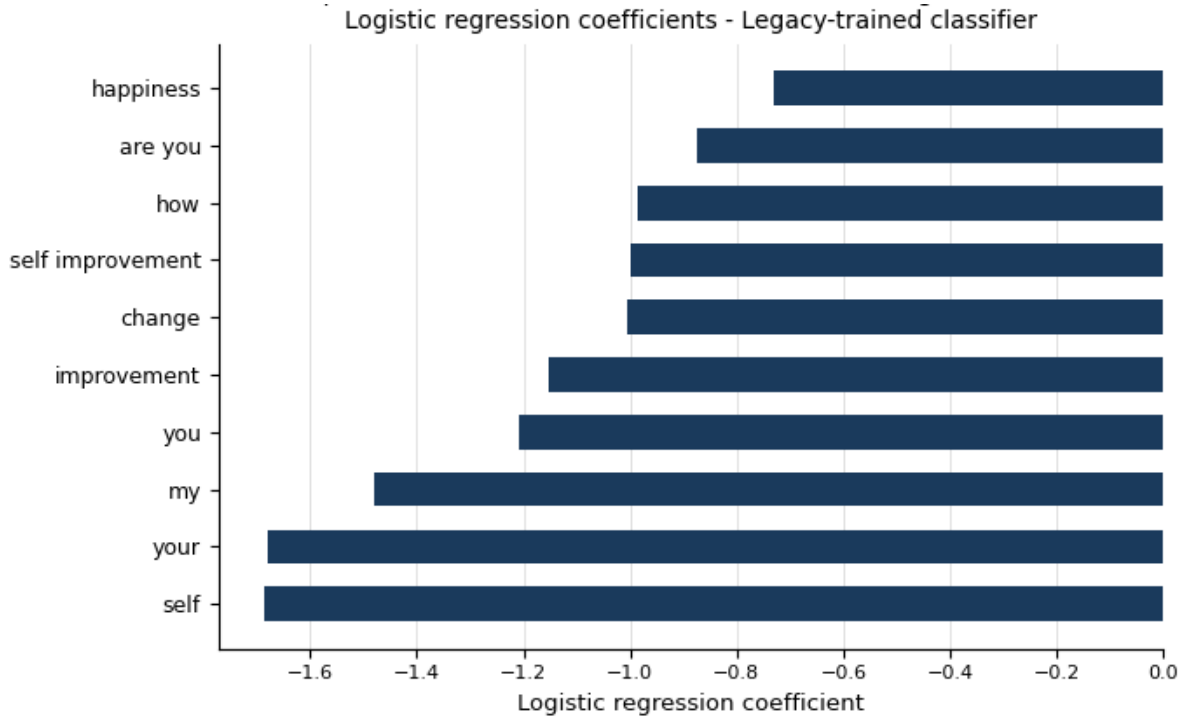

Figure 6. Features most strongly associated with control classification. The features most predictive of control content include self, your, my, you, improvement, change, self-improvement, how, are you, and happiness. This supports the interpretation that Sigma-era manosphere posts become harder to classify when they adopt the self-help and advice-seeking language that the Legacy-trained classifier associates with benign control communities.

Note. More negative coefficients indicate stronger association with control classification. The figure therefore shows which surface features pull posts away from manosphere classification in the temporal generalisation model.

Trope classification indicates differential detectability across ideological categories. Prescription is the most reliable category (F1 = 0.842, κ = 0.777), suggesting that prescriptive self-improvement ideology retains enough stable surface signal to classify. Rejection is also detectable by F1 (0.769), but its κ value (0.617) remains below the stronger reliability threshold, so it is interpreted cautiously. Crisis and Hate are excluded from prevalence analysis because their validation scores are too weak.

Indicatively, Rejection prevalence increases from Legacy (0.390, 95% CI [0.337, 0.446]) to Sigma (0.507, 95% CI [0.450, 0.563]). Although the confidence intervals do not overlap, this pattern is interpreted cautiously because Rejection did not meet the stronger validation threshold.

Table 6. Trope prevalence by era with Wilson 95% confidence intervals (n = 300/era, balanced sample)

Trope	Legacy (≤2016)	Pill (2017–21)	Sigma (2022–25)	Δ L to S
Prescription	0.213	0.290	0.253	+0.040
Rejection	0.390	0.383	0.507	+0.117*

*Note. *Wilson 95% CIs do not overlap between Legacy and Sigma. Prescription is treated as validated (F1 = 0.842, κ = 0.777). Rejection is reported cautiously as indicative (F1 = 0.769, κ = 0.617). Crisis and Hate are excluded from prevalence interpretation because classifier performance did not meet the reliability threshold.*

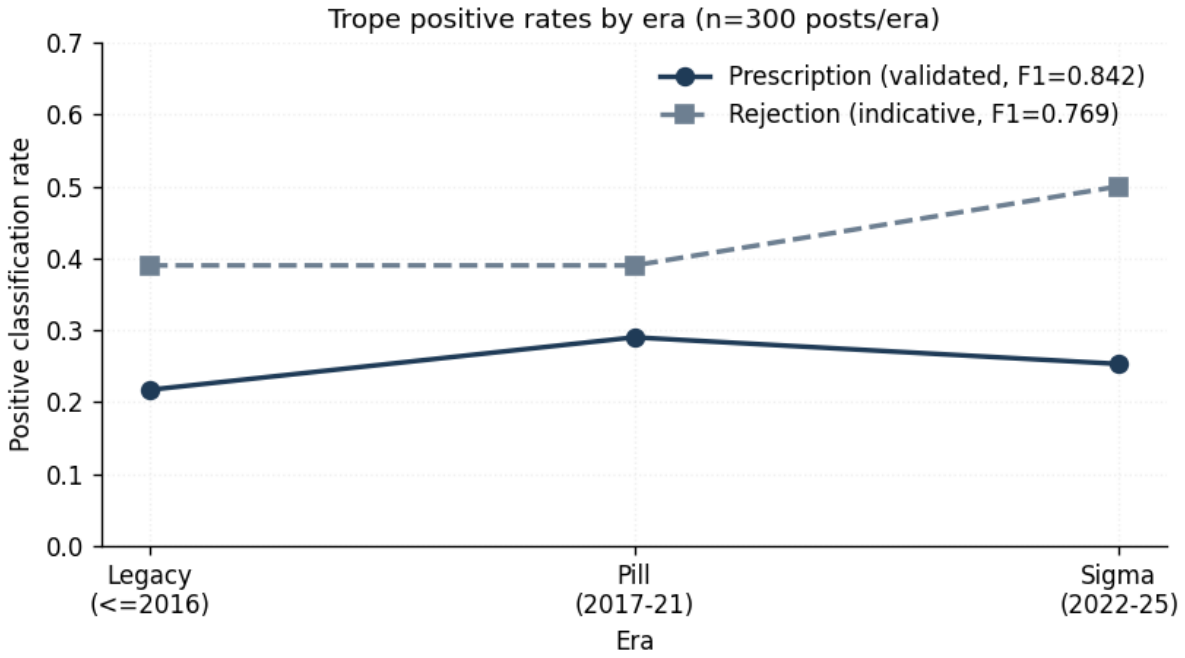

Figure 7. Trope prevalence by era. Prescription remains broadly stable across eras, while Rejection increases from Legacy to Sigma. This pattern is consistent with the wider claim that manosphere ideology persists beneath register change, but that different tropes vary in how reliably they can be detected.

Note. Prescription is treated as validated; Rejection is reported as indicative only. Shaded bands show Wilson 95% confidence intervals. Crisis and Hate are excluded because classifier performance did not meet the threshold for reliable prevalence interpretation.

4.3 Rhetorical Strategies: Structural Detection Failure (RQ2 & RQ3)

Reflexive thematic analysis of the twelve purposively selected cases generated three themes through iterative coding across two analytical passes. Initial Pass 1 codes, locutionary surface, illocutionary function, presupposition structure, were treated as raw material for Pass 2 rather than as themes in themselves, consistent with Braun and Clarke's (2006, 2019) requirement that thematic analysis remain inductive at the level of pattern rather than deductive at the level of category.

Early coding separated “self-improvement advice” from “status prescription,” but these were collapsed after repeated comparison showed that the distinction often described surface register rather than underlying function. This revision is why Theme 3 is relatively frequent: it captures the mechanism most compatible with Sigma-era wellness language, not a pre-set category imposed before coding.

Table 7. Rhetorical themes: definitions, case distribution, and era coverage

Theme	Definition	Cases (n)	Eras present
1. Epistemic authority as mask	Surface register of empirical analysis or academic debate borrowed to naturalise ideological assumptions as settled findings. Ideology is carried by presupposition rather than assertion.	2, 3, 7, 11 (n=4)	Legacy, Pill
2. Fairness and symmetry inversion	Progressive or feminist rhetoric repurposed to reverse its political direction. Anti-feminist ideology encoded through the rhetorical structure of feminist argument.	1, 5, 8 (n=3)	Legacy, Sigma
3. Softened or displaced prescription	Hierarchical ideological content encoded through wellness, productivity, or individualised self-improvement register. Coercive frame removed; prescriptive function preserved.	4, 6, 9, 10, 12 (n=5)	Legacy, Pill, Sigma

Note. Theme 3 is the only theme present across all three eras. Theme 1 is absent from Sigma; Theme 2 is absent from Pill.

Case 1 (r/TheRedPill, Legacy era) illustrates Theme 2. The post poses a rhetorical question linking the convention of men paying for dates to demands for gender equality, implying that feminist goals are inconsistent if traditional dating costs persist. VADER score: −0.026. SBERT similarity to Rejection anchors: 0.561. The classifier assigned this to the non-ideological class. Pass 1 analysis: locutionarily, this is a fairness question, no hostile vocabulary, neutral syntax. Illocutionarily, it performs a delegitimisation of feminist claims by

presupposing that feminism is a demand for advantage rather than equality, encoding that presupposition as the conclusion any reasonable person would reach. The ideology is in what the question takes for granted, not in what it asserts. Pass 2 (reflexive thematic): Theme 2, progressive rhetoric repurposed to reverse its political direction. No retraining on updated vocabulary resolves this gap.

Case 2 (r/PurplePillDebate, Pill era) illustrates Theme 1. A title presents a claim about female behaviour framed as a research finding. VADER score: 0.000. SBERT similarity to Crisis anchors: 0.759. The classifier assigned this to the non-ideological class. Pass 1: locutionarily, a report of empirical evidence, objective register, no slurs, neutral syntax. Illocutionarily, it naturalises male victimhood by embedding the presupposition that women systematically mis-evaluate men as a discovered fact. The 'studies show' construction confers epistemic authority that forecloses critique before it can begin. Pass 2: Theme 1. A classifier finds a data claim; a human reader finds a grievance legitimised as science.

Cases 9 and 10 (r/TheRedPill, Legacy and Sigma respectively) together illustrate Theme 3 and the collocate-shift mechanism across a decade. Case 9 frames self-discipline as a time-limited challenge to give up personal vices, drawing on a moral-religious register of temptation and self-control. Case 10 frames discipline as outdated and reframes the same underlying behaviour as productivity optimisation, centered on eliminating procrastination. The ideological prescription (male self-mastery as resistance to feminised culture) is constant; the surface context has entirely migrated. Pass 1: the illocutionary force, perform discipline or forfeit masculine status, is identical across the decade. Pass 2: Theme 3, softened prescription.

The remaining cases were not treated as separate representative mini findings, but as checks on the three mechanisms identified through the detailed analysis. Cases 3, 7 and 11 reinforced Theme 1 by showing how claims about male disadvantage were stabilised through quasi-evidential or explanatory registers. Cases 5 and 8 extended Theme 2 by showing that fairness language could carry hostile or exclusionary implications while remaining affectively neutral. Cases 4, 6 and 12 extended Theme 3 by showing that prescription often appeared as self-management or optimisation advice rather than explicit ideological instruction. This distribution explains why Theme 3 appears most frequently in the corpus: it is the mechanism most compatible with Sigma-era wellness language, not an artefact of deductive coding.

4.4 Integrated Findings: The Machine Blindness Mechanism

The quantitative and qualitative strands converge on a single explanatory mechanism. Jaccard shows turnover (0.6276) that is directional rather than unusually large (controls: 0.7751); SBERT shows rising ideological proximity despite that turnover; VADER shows affective tone unchanged and consistent with sentiment-based detection finding no signal.

The temporal generalisation test shows transfer difficulty under register shift. The 24.4pp train-to-test F1 drop is consistent with Ushio and Camacho-Collados's finding that hate speech detection can degrade under temporal shift, but the much smaller CV-to-test comparison means this result is best read as evidence of transfer difficulty rather than as a clean estimate of temporal drift alone.

Table 8 is therefore presented not as an additional empirical result, but as a translational synthesis of the preceding findings for safeguarding practice.

Table 8. Safeguarding Diagnostic Framework for Recognising Manosphere Content

Failure mode	Linguistic appearance	Underlying mechanism	Risk for practitioners	Recommended response
Temporal blindness	Wellness, productivity, accountability language	Register mutation	Harmful content mistaken for generic self-improvement	Focus on function rather than keywords
Structural blindness	Neutral questions, fairness claims, advice framing	Presupposition and illocutionary force	Ideological content hidden within apparently benign language	Examine assumptions and implied meanings
Epistemic authority masking	Studies, statistics, 'research shows' framing	Ideological claims presented as factual conclusions	Credibility granted to ideological assertions	Evaluate evidence and framing separately
Fairness inversion	Equality rhetoric used to reverse critique	Progressive language repurposed for grievance narratives	Content appears reasonable at first glance	Analyse who is positioned as victim and why
Softened prescription	Self-improvement and discipline discourse	Gender hierarchy implied rather than stated	Advice interpreted as ideologically neutral	Consider what normative assumptions underpin the advice

5. Discussion and Conclusion

5.1 Interpretation Against Prior Literature

The findings extend and specify the account of manosphere mainstreaming in Gerrand et al. (2025) and Haslop et al. (2024). What those studies document at the level of influencer strategy and audience affect, this study examines at the corpus level across thirteen years: the wellness register appears to constitute a substantive shift rather than cosmetic variation, one that produces measurable degradation in Legacy-calibrated detection infrastructure. The 24.4pp train-to-test F1 drop is consistent with Ushio and Camacho-Collados's finding that hate speech detection can degrade under temporal shift. However, the much smaller CV-to-test comparison means this result is best interpreted as evidence of transfer difficulty under register shift rather than as a clean estimate of temporal drift alone.

Within the critical realist framework, this finding confirms that the real domain, manosphere ideology, remains structurally stable even as the actual domain of observable language undergoes substantial surface transformation. The monotonically rising proximity to academic definitions of manosphere ideology, from 0.3483 to 0.3843 on Ging anchors despite 63% vocabulary turnover, is consistent with ideological sharpening: Sigma-era content articulates core positions in the vocabulary of wellness rather than abandoning them. The r/ExRedPill comparison confirms that SBERT captures articulation rather than endorsement.

This paper therefore does not claim that Sigma-era discourse is uniquely lexically unstable, nor that semantic proximity measures belief or endorsement directly. Its narrower claim is that the direction and consequences of register mutation matter: manosphere content can move toward mainstream wellness language while retaining measurable semantic alignment with ideological anchor categories.

5.2 Implications for Detection and Safeguarding

For automated moderation, the findings distinguish two problems that require different solutions. The temporal failure mode, Legacy-trained classifiers degrading on Sigma content, is addressable in principle by retraining on contemporary data, monitoring register drift, and updating detection infrastructure on shorter cycles than current practice suggests. The structural failure mode, Rejection-trope content evading detection because the ideology is carried by presupposition and illocutionary force, is unlikely to be solved by vocabulary retraining alone. Surface-based classifiers are therefore likely to remain vulnerable unless supplemented by models or human review processes capable of evaluating pragmatic function, context, and discourse-level patterns; this remains an open problem in computational linguistics (Ziems et al., 2024; Grimmer et al., 2022).

For school safeguarding, the implications are analogous. The KCSiE 2026 proposed revisions and RSHE statutory guidance place renewed emphasis on recognising misogyny and harmful online influence in schools. The practitioner diagnostic tool developed from this research operationalises both failure modes as a recognition framework that does not depend on keyword recall. This positions the intervention at the pre-recognition stage of safeguarding practice: the point at which staff may notice a shift in language or behaviour

but lack a framework for interpreting its ideological function. Recognition therefore shifts from asking only, ‘is this a red-flag word?’ to asking what assumptions about women, men, entitlement, blame, and authority the language makes available.

5.3 Limitations, Future Directions, and Conclusion

Four limitations constrain interpretation. First, the analysis covers post titles only from four Reddit communities; comments, other subreddits, and platforms where Sigma-era content is now more prominent, TikTok, Telegram, YouTube, may show different dynamics, and the findings should not be generalised beyond the studied context without replication. Second, the gold standard was coded by a single researcher; κ values measure classifier-coder agreement rather than intercoder reliability, and F1 scores should be read as indicators of relative trope difficulty rather than absolute accuracy benchmarks. Third, r/AllPillDebate contributes 17.6% of Sigma-era titles but is absent from Legacy and Pill, introducing a potential era-specific confound; a robustness check excluding this subreddit is therefore the priority for future work. Fourth, SBERT similarity provides evidence of semantic and ideological alignment with anchor statements, not direct evidence of individual belief, endorsement, or intent.

The value of this study is not a list of current terms to flag. It is a framework for understanding that ideology travels through language at the level of illocutionary force rather than lexical surface. The central finding can be stated precisely: surface vocabulary and underlying ideology are systematically dissociated in Sigma-era manosphere discourse, producing two distinct and non-reducible failure modes in automated detection, one addressable in principle by retraining on contemporary data, and one likely to remain vulnerable when detection relies only on surface-level features regardless of training horizon.

The contribution of this study is diagnostic rather than predictive. For computational linguistics and NLP, it shows that temporal robustness is necessary but not sufficient when the construct of interest is pragmatic ideology rather than lexical hostility. For gender and masculinity studies, it offers a way of operationalising manosphere mainstreaming that is sensitive to register, affect, and illocutionary force. For platform governance and safeguarding, it reframes recognition as a question of what language is doing in context, not simply which words it contains. Across these domains, the central claim is that surface vocabulary and ideological function can become systematically dissociated, producing detection failures that keyword-based or historically calibrated systems are poorly equipped to resolve.

References

Aggarwal, J., & Stevenson, S. (2024). Style-shifting behaviour of the manosphere on Reddit. *In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 21976-21989). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.1226/

Austin, J. L. (1962). How to do things with words. Oxford University Press.

Bhaskar, R. (1978). A realist theory of science. Harvester Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health, 11(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Bridges, T., & Pascoe, C. J. (2014). Hybrid masculinities: New directions in the sociology of men and masculinities. Sociology Compass, 8(3), 246–258. https://doi.org/10.1111/soc4.12134

Bunt, H. L., Goddard, A., Reader, T. W., & Gillespie, A. (2025). Validating the use of large language models for psychological text classification. Frontiers in Social Psychology, 3, 1460277. https://doi.org/10.3389/frsps.2025.1460277

Connell, R. W., & Messerschmidt, J. W. (2005). Hegemonic masculinity: Rethinking the concept. Gender & Society, 19(6), 829–859. https://doi.org/10.1177/0891243205278639

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955

de Kock, C. (2024a). Jointly modelling the evolution of community structure and language in online extremist groups. arXiv. https://arxiv.org/abs/2409.19243

de Kock, C. (2024b). LISTN: Lexicon induction with socio-temporal nuance. arXiv. https://arxiv.org/abs/2409.19257

Department for Education. (2025). *Relationships education, relationships and sex education (RSE) and health education: Statutory guidance.* HM Government. https://www.gov.uk/government/publications/relationships-education-relationships-and-sex-education-rse-and-health-education

Department for Education. (2026). *Keeping children safe in education: 2026 proposed revisions.* HM Government. https://www.gov.uk/government/consultations/keeping-children-safe-in-education-proposed-revisions-2026

Evans, A., & Ringrose, J. (2025). More-than-human, more-than-digital: Postdigital intimacies as a theoretical framework. Social Media + Society, 11(1). https://doi.org/10.1177/20563051251317779

Farrell, T., Araque, O., Fernandez, M., & Alani, H. (2020). On the use of jargon and word embeddings to explore subculture within Reddit's manosphere. *In Proceedings of the 12th ACM Conference on Web Science (WebSci '20)* (pp. 221-230). ACM. https://doi.org/10.1145/3394231.3397912

Franzke, A. S., Bechmann, A., Ess, C. M., & Zimmer, M. (Eds.). (2020). Internet research: Ethical guidelines 3.0. Association of Internet Researchers. https://aoir.org/reports/ethics3.pdf

Gerrand, V., Ging, D., Roose, J. M., & Flood, M. (2025). Mapping the neo-manosphere(s): New directions for research. Men and Masculinities, 28(5), 443–464. https://doi.org/10.1177/1097184X251350277

Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press. https://doi.org/10.12987/9780300235029

Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. Men and Masculinities, 22(4), 638–657. https://doi.org/10.1177/1097184X17706401

Gothard, K., Dewhurst, D. R., Minot, J. R., Adams, J. L., Danforth, C. M., & Dodds, P. S. (2021). The incel lexicon: Deciphering the emergent cryptolect of a global misogynistic community. arXiv. https://arxiv.org/abs/2105.12006

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022*). Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Haslop, C., Ringrose, J., Cambazoglu, I., & Milne, B. (2024). Mainstreaming the manosphere's misogyny through affective homosocial currencies. Social Media + Society, 10(1). https://doi.org/10.1177/20563051241228811

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. Trends in Cognitive Sciences, 18(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Krendel, A. (2020). The men and women, guys and girls of the 'manosphere': A corpus-assisted discourse approach. Discourse & Society, 31(6), 607–630. https://doi.org/10.1177/0957926520939690

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Marwick, A., & Lewis, R. (2017). Media manipulation and disinformation online. Data & Society Research Institute. (pp. 1–104). https://www.datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

Ofcom. (2024*). Illegal content risk assessment guidance and risk profiles*. Ofcom. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/online-safety-regulatory-documents

Pinkett, M. (2026). *Unmasking the manosphere: Tackling the misogyny crisis in schools.* Routledge.

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S., & Zannettou, S. (2021). The evolution of the manosphere across the Web. *Proceedings of the International AAAI Conference on Web and Social Media, 15*(1), 196–207. https://doi.org/10.1609/icwsm.v15i1.18053

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge University Press.

Skinner, Q. (2002). Visions of politics: Volume 1, Regarding method. Cambridge University Press.

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. Political Communication, 37(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Tanner, S., & Gillardin, F. (2025). Toxic communication on TikTok: Sigma masculinities and gendered disinformation. Social Media + Society, 11(1). https://doi.org/10.1177/20563051251313844

Törnberg, P. (2024). How to use large-language models for text analysis. In *Doing research online*. SAGE Research Methods. https://doi.org/10.4135/9781529683707

Ushio, A., & Camacho-Collados, J. (2024). A systematic analysis on the temporal generalization of language models in social media. *In Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysi*s (pp. 52–62). https://aclanthology.org/2024.wassa-1.5/

Women's Aid. (2023). *Influencers and attitudes: How will the next generation understand domestic abuse?* Women's Aid Federation of England. https://www.womensaid.org.uk/wp-content/uploads/2023/12/CYP-Influencers-and-Attitudes-Report.pdf

YouGov. (2023, September 27). *One in six boys aged 6–15 have a positive view of Andrew Tate.* https://yougov.co.uk/society/articles/47419-one-in-six-boys-aged-6-15-have-a-positive-view-of-andrew-tate

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? Computational Linguistics, 50(1), 237–291. https://doi.org/10.1162/coli_a_00502

AI declaration

I used Grammarly for spelling and grammar correction. I used generative AI tools, including ChatGPT, to support structural refinement, clarity checking, and critical review against the assessment criteria and for limited visual refinement of the infographic/workflow diagram. I used Perplexity to help scope relevant academic and non-academic sources; all sources included in the final submission were independently checked and critically evaluated by me.

ChatGPT was used for occasional support with coding, including debugging Python notebooks. It was also used to help improve APA-style referencing in the reference list. Large language models were additionally used as part of the research workflow for trope classification, as described in the methodology, and their outputs were validated against a hand-coded test set.

All final analysis, interpretation, source evaluation, and wording are my own.