

When Harm Is Polite: Coercion Markers and Low Toxicity in Conversion-Therapy Discourse

Adhiraj Chhoda

chhodaadhiraj@gmail.com

Abstract

Ten years of online-harms research have largely operationalized harm through abusive tone: slurs, hostility, and platform-visible toxicity. Yet some harms are carried by polite, institutional, role-mediated discourse that standard toxicity classifiers consistently miss. We study one such case through a corpus of 2,388 publicly available conversion therapy (CT) documents (119,976 sentences) compared with 26 APA-aligned affirmative-therapy controls (16,948 sentences). We test a semantic moralization hypothesis using contextual association tests; the signed statistic ($d = -5.11$, $p = 0.0002$) indicates that identity terms in CT are on average more similar to neutral-psychology attributes than to morality attributes, contradicting the hypothesis. A transparent rule set based on the BITE model of coercive control finds all four dimensions elevated in CT, with emotional control highest. We test two further alternative mechanisms—a staged emotional script and grammatical agency-deletion—and find neither at corpus scale. Three toxicity classifiers (Detoxify, ToxiGen-HateBERT, ToxiGen-RoBERTa) do not register the elevated coercion markers: Detoxify and RoBERTa flag $\leq 1.4\%$ of CT sentences; HateBERT flags 11.5% but triggers on identity-group vocabulary rather than coercive structure. These findings argue for online-harms taxonomies that incorporate relational context and semantic constraint alongside surface abuse, and we release derived features and code.

1 Introduction

A decade of work at WOAHA has steadily widened the scope of detectable harm. Early systems keyed on explicit slurs; later work reached coded language (Magu et al., 2017), implicit hate (ElShrief et al., 2021), coded rhetoric and dogwhistles (Mendelsohn et al., 2023), and masked trans-exclusionary rhetoric (Lu and Jurgens, 2022). The

field’s tenth-anniversary call asks where the enduring gaps remain, and points in particular to systemic and structural harms, to user agency, and to forms of harm that narrow phenomenon-by-phenomenon detection has tended to miss. This paper takes up one such gap: *institutional, role-mediated harm*, in which a power-holding actor systematically redefines ordinary vocabulary to constrain a target’s self-understanding, without ever using language that a toxicity or hate-speech classifier would flag.

Conversion therapy (CT) is a clear instance. Practitioners redefine identity terms (“attraction,” “orientation,” “desire”) into a moral-pathological register (“unwanted same-sex attraction,” “gender-identity deficit,” “sexual brokenness”) within a therapeutic relationship in which the practitioner holds authority over the client. The language is procedural, clinical, and often warm. Because the harm is not signaled by abusive tone, classifiers trained on overtly hostile text are unlikely to register it; we test this expectation directly with three detectors spanning the Jigsaw (Detoxify), HateBERT, and ToxiGen families. The stakes are well documented. The APA Task Force (American Psychological Association, Task Force on Appropriate Therapeutic Responses to Sexual Orientation, 2009) concluded that sexual orientation change efforts lack evidence of efficacy and carry risk of harm; a cross-sectional study of 4,426 sexual and gender minority adults linked CT exposure to elevated depression, PTSD, and suicidality (Tran et al., 2024), and population samples report nearly doubled odds of lifetime suicidal ideation after exposure (Blosnich et al., 2020). The American Academy of Pediatrics (American Academy of Pediatrics, 2018) holds that framing a diverse identity as a condition to be treated is itself harmful, independent of technique. That position motivates our question: what does the harmful framing look like as a measurable property of language?

We test three hypothesized mechanisms of CT discourse. A contextual association test does not support the semantic moralization hypothesis; however, the BITE model of coercive control (Hassan, 2015, 2020) reveals elevated markers across all four dimensions. We also test and refute at corpus scale a staged emotional script and grammatical agency-deletion. Finally, three toxicity classifiers pass CT text at rates that highlight a detection blind spot for polite, role-mediated harm.

Our contributions are:

1. **A harm type for the WOA taxonomy.** We characterize role-mediated harm—polite institutional discourse that constrains identity through meaning rather than tone—and show its coercive-control markers are measurable.
2. **Direct evidence of a detection blind spot.** Three toxicity classifiers from independent model families fail to register elevated coercion markers: two pass $\geq 98.6\%$ of CT sentences; the third triggers on identity-group vocabulary but does not distinguish coercive structure from benign mentions.
3. **A measurement approach with reported nulls.** We test three mechanistic hypotheses—semantic moralization, staged emotional scripting, and agency deletion—and report all three as unsupported, replacing loose analogies with specific, falsifiable claims.

We do not claim CT is reducible to a single mechanism, nor that our measures settle clinical questions. We characterize the structure of public CT discourse and the blind spot it exposes in toxicity-centered measurement.

2 Related Work

2.1 From Explicit to Disguised Abuse

Abusive-language detection has moved from lexicon matching toward models that recover hostility encoded beneath the surface: implicit hate (ElShrief et al., 2021), the “hate code” of substituted terms (Magu et al., 2017), and dogwhistles whose harmful reading is available only to an in-group (Mendelsohn et al., 2023). A recurring finding is that seemingly non-hostile language can still cause harm; Lu and Jurgens (2022) show this for trans-exclusionary rhetoric that presents as reasoned and even sympathetic. Surveys of hate-speech detection (Rawat et al., 2024) note that transformer classifiers trained on overtly hostile text systematically miss

harm expressed in measured or clinical registers, and the broader harmful-content literature shares this blind spot for disguised harm (Perez-Rosas et al., 2018; Torregrosa et al., 2023). Persuasion-technique detection supplies relevant taxonomies, including loaded language, appeal to authority, and false dilemma (Dimitrov et al., 2021; Piskorski et al., 2023), and computational framing analysis formalizes how texts make particular construals salient (Entman, 1993; Card et al., 2015). Work on dehumanization models the denial of agency and the reframing of personhood as defect (Mendelsohn et al., 2020). None of this infrastructure has been directed at institutional therapeutic discourse, where persuasion is delivered under the guise of care.

2.2 Coercive Control and the BITE Model

A long tradition in the psychology of coercive persuasion describes how institutional actors overwrite a target’s self-concept. Lifton (1961) names “loading the language,” the compression of experience into thought-terminating clichés and insider terms that short-circuit independent reasoning. Schein (1961) and Singer and Lalich (1995) add process and condition models. Hassan (2015) consolidates this line into the BITE model, organizing undue influence into Behavior, Information, Thought, and Emotional control. Critical discourse analysis advances a complementary, specifically syntactic hypothesis: that coercive identity language deletes agency through nominalization and agentless passives (Billig, 2008; Lingle, 2018). Goffman (1974) frames the same move at the level of interpretation, where relocating a descriptor into a new frame changes what it means and what response it licenses. These accounts leave observable traces in text, which makes them operationalizable; we treat them as *a priori* criteria rather than post hoc patterns.

2.3 Conversion Therapy and Documented Harm

The clinical record establishes that CT causes harm (American Psychological Association, Task Force on Appropriate Therapeutic Responses to Sexual Orientation, 2009; Przeworski et al., 2021; Tran et al., 2024; Salway et al., 2020), with population data indicating not only pre-existing vulnerability, but the practice itself, as contributing to suicidality (Blosnich et al., 2020). Consumer and survivor studies describe the harm as substantially linguistics-

	CT corpus	Control
Documents	2,388	26
Paragraphs (CT total)	18,891	
Sentences	119,976	16,948

Table 1: Analytical corpus after extraction and filtering. The control is a small directional reference (see Limitations).

tic: false efficacy claims, pathologizing characterizations, and shame-based messaging (Shidlo and Schroeder, 2002). An estimated 698,000 U.S. adults have undergone CT, most often from religious rather than licensed providers (Mallory et al., 2019). The clinical literature documents that harm occurs but does not analyze the discourse structures through which the framing operates. We take up that gap with computational tools.

3 Data

CT corpus. We compiled 2,388 publicly available CT documents spanning practitioner manuals, organizational websites, the *Journal of Human Sexuality* (NARTH/ATCSI), archived Exodus International and Love in Action materials recovered from the Wayback Machine, institutional and UN reports, and program documents. After extraction, deduplication, and paragraph-level filtering of boilerplate and front matter, the corpus contains 119,976 sentences across 18,891 paragraphs (15,501 substantive after boilerplate removal). All materials are publicly accessible; the corpus contains no session recordings or confidential notes. We follow WOA’s reporting policy by paraphrasing rather than reproducing harmful content.

Control corpus. 26 APA-aligned affirmative-therapy documents (16,948 sentences), drawn from national association guidelines (APA, BPS, APS, CPA, RCPsych), WPATH standards of care, SAMHSA resources (Substance Abuse and Mental Health Services Administration, 2015), and open-access affirmative-therapy articles. These texts address the same population and clinical domain under the opposed orientation that identity is not pathological. The control is smaller because affirmative material of the same practitioner-facing register is far less abundant in public archives (Table 1); we treat it as a directional reference and anchor our strongest claims in within-CT structure.

4 Methods

All components share a preprocessing stage built on spaCy’s `en_core_web_trf` model (Honnibal et al., 2020), which tokenizes, lemmatizes, tags, and dependency-parses both corpora; documents are segmented into paragraphs and sentences with source metadata retained. Random seeds are fixed at 42.

Topic modeling. We recover the dominant rhetorical repertoire with BERTopic (Grootendorst, 2022) over `all-mpnet-base-v2` paragraph embeddings, using LDA (Blei et al., 2003) with identical preprocessing as a robustness baseline. Topic count is selected by a c_v coherence sweep over 8–25 topics. A custom vectorizer with an n-gram range of (1,3) retains multi-word CT constructions such as “same-sex attraction.”

Contextual association. For each target term with at least three occurrences, we collect the first 100 matching sentences in corpus order (or all matches if fewer), encode each with `all-mpnet-base-v2`, and average normalized sentence embeddings to obtain a corpus-specific contextual representation; the measure is over the contexts in which identity terms appear, not isolated word vectors. Inspired by the Word Embedding Association Test (WEAT; Caliskan et al. 2017) and Sentence Encoder Association Test (SEAT; May et al. 2019), the implemented one-target statistic is

$$s(t) = \text{mean}_{a \in A} \cos(t, a) - \text{mean}_{n \in N} \cos(t, n),$$

where A is morality or pathology and N is neutral psychology. We report $d = \text{mean}_t s(t) / \text{sd}_t s(t)$ (sample standard deviation, $n-1$ degrees of freedom). Negative values indicate greater average similarity to neutral psychology than to the attribute set. Only terms meeting the three-occurrence threshold in each corpus are included; the eligible sets differ between corpora (Appendix A). Cross-corpus cosine distance measures term-level shifts. A separate sensitivity analysis samples up to 200 contexts randomly (seed 42) with a four-word minimum filter and reports document-level bootstrap 95% CIs (500 resamples).

BITE-aligned rule set. A transparent rule set operationalizes Hassan’s four dimensions through 22 interpretable feature types (keyword lists, lemma

matches, regex, and morphological filters; Appendix B). A sentence activates a dimension if any of its features fire, subject to clinical-exclusion patterns that suppress hits in descriptive or citational contexts. Because the BITE lexicon was derived from CT texts, vocabulary overlap inflates CT scores by construction; the CT-control contrast partly reflects this circularity and cannot independently validate coercive control. We report activation rates as an interpretable indicator of coercive-control register, not a validated classifier, and flag systematic annotation with multiple raters as future work.

Alternative mechanisms. (1) Emotional sequence: we map paragraph emotion with a GoEmotions RoBERTa classifier (Demszky et al., 2020; Lowe, 2023) onto a five-phase coercive cycle (shame, fear, hope, conditional acceptance, renewed shame) and test for the staged ordering. (2) Syntactic framing: using dependency parsing and PassivePy (Sepehri et al., 2022), we measure passive-voice rates, agent omission, nominalization density, and whether identity terms occupy agent or patient roles.

Toxicity baseline. To test whether off-the-shelf detection registers the discourse, we score a random sample of 3,000 CT and 3,000 control sentences (fixed seed) with three independent classifiers: Detoxify (Hanu and Unitary team, 2020) (the original model, trained on the Jigsaw data underlying Perspective API), ToxiGen-HateBERT (HateBERT fine-tuned on the ToxiGen adversarial benchmark; Hartvigsen et al. 2022), and ToxiGen-RoBERTa (RoBERTa-large fine-tuned on ToxiGen). We report score distributions and flag rates at the standard 0.5 threshold for each.

Statistics. Embedding measures use permutation tests ($n = 10,000$ two-sided attribute-label shuffles); per-sentence BITE activation uses Yates-corrected Pearson’s χ^2 on activation counts; continuous syntactic and emotion measures use the Mann–Whitney U test, with Cohen’s d for effect size. Because sentences are nested within documents, the effective sample size is smaller than the raw sentence count, and sentence-level p -values overstate significance. We therefore report effect sizes and direction as the primary evidence, treat p -values as ordinal indicators rather than precise tail probabilities, and supplement with document-level bootstrap confidence intervals that respect the

Top terms	Interpretation
change, SOCE, APA, therapy	SOCE/APA efficacy debate
God, life, love, Jesus, church	Religious redemption
children, parents, same-sex	Developmental/parental etiology
unwanted, attractions, clients	Unwanted-attraction framing
homosexuality, Exodus, church, gay	Ex-gay ministry
gender, identity, transgender	Gender identity

Table 2: Representative LDA topics (of 19 after filtering) in the CT corpus. SOCE: sexual orientation change efforts.

clustering structure. Reported p -values are nominal unless otherwise noted.

5 Results

We report the association evidence first, then the coercive-control analysis, then the two refuted alternatives, then the toxicity baseline. The headline pattern is that elevated coercion markers coexist with near-floor toxicity scores.

5.1 Thematic structure

LDA reaches peak coherence at 19 topics ($c_v = 0.75$); BERTopic produces a finer-grained partition that agrees on the major clusters. The repertoire (Table 2) spans a developmental-etiology narrative, a religious-redemption frame, a clinical framing of orientation as an “unwanted” condition, and a policy register contesting the evidence against change. These themes recur across books, websites, and pamphlets rather than concentrating in one source, indicating a shared rhetorical repertoire rather than the idiom of a single author.

5.2 Identity–attribute associations

The moralization hypothesis predicts that identity terms in CT should associate more strongly with morality attributes than with neutral-psychology attributes. The first-100 contextual association statistic is $d = -5.11$ (permutation $p = 0.0002$): on average, identity-term contexts are *more* similar to neutral-psychology attributes (well-being, authentic, congruence) than to morality attributes (sin, redemption, purity). This contradicts the moralization hypothesis; the signed statistic does not support the claim that CT relocates identity into a moralizing semantic field. Identity–pathology

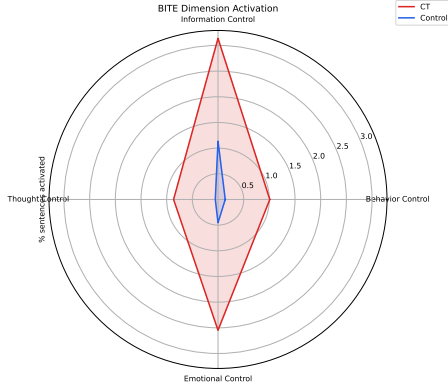


Figure 2: Four-dimensional BITE profile showing document-mean activation percentages. Table 4 reports pooled sentence-level rates.

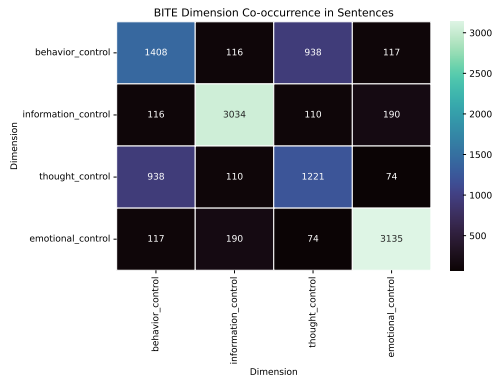


Figure 3: Within-sentence co-occurrence of BITE dimensions in CT. Behavior and thought control co-occur most often (938 sentences), vs. 190 for information and emotional.

No staged emotional cycle. The canonical shame–fear–hope–conditional–acceptance sequence appears zero times in either corpus, and the relaxed cycle is no more frequent in CT than in the control. What the emotion analysis shows instead is *ambient shame*: shame and guilt occupy 38.5% of CT paragraphs (neutral 49.6%, hope 6.0%, conditional love 3.2%, fear 1.5%) as a steady background rather than the opening beat of a script. Shame is also high in the control (32.6% of paragraphs; document-level difference not significant, $p = 0.17$), so the harm does not depend on a therapist executing a sequence in order (Figure 4).

No syntactic evasion. CT is not grammatically evasive. Passive-voice rates are not higher in CT and are if anything lower than the control (9.9% vs. 12.3%, Cohen’s $d = -0.25$); agent omission within passives is lower (64.1% vs. 90.6%) and nominalization density is lower (8.9 vs. 14.6 per

Dimension	CT%	Ctrl%	Δ pp [95% CI]
Emotional	2.61	0.53	2.08 [1.78, 2.41]
Information	2.53	0.54	1.99 [1.61, 2.35]
Behavior	1.17	0.16	1.01 [0.89, 1.15]
Thought	1.02	0.08	0.94 [0.85, 1.05]

Table 4: Pooled BITE sentence activation: 119,976 CT and 16,948 control sentences. A sentence counts once per dimension. CT–control differences (Δ) in percentage points (pp) with 10,000-resample document bootstrap CIs. The CT-derived lexicon inflates CT rates by construction.

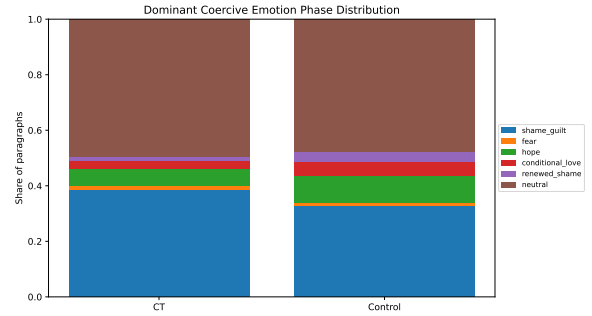


Figure 4: Emotion-phase distribution by corpus. Shame/guilt dominates CT but is also high in the control; no sequential cycle emerges.

1,000 words). Both corpora cast identity terms as grammatical patients at nearly identical rates (CT 80.2%, control 83.3%). The critical-discourse hypothesis that coercive identity language hides agency (Billig, 2008; Lingle, 2018) is not borne out at corpus scale; practitioners state claims plainly (“You have a gender-identity deficit”). The coercion is in what the words mean, not in who is shown to act (Figure 5).

5.5 Toxicity classifiers do not register the discourse

We scored 3,000 CT and 3,000 control sentences (fixed seed) with three independent classifiers (Table 5). Detoxify reports a CT median toxicity of 0.001 (mean 0.033); only 1.43% of sentences exceeded the 0.5 flag threshold, and the identity-attack score flags 0.1%. ToxiGen-RoBERTa is similarly floor-level (1.07% flagged). ToxiGen-HateBERT is the partial outlier: it flags 11.5% of CT sentences (vs. 3.1% of controls), reflecting its sensitivity to identity-group language in the ToxiGen training set. Yet even HateBERT’s flags are driven by vocabulary overlap rather than structural recognition of the coercive mechanism: the flagged sentences are those containing identity terms in any context, not

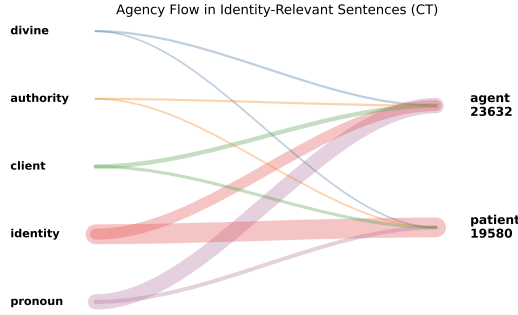


Figure 5: Agent and patient role assignment for identity terms. Both corpora cast identity terms as patients at similar rates; CT shows no distinctive agency-deletion pattern.

Classifier	Corpus	Med.	Mean	p90	%flag
Detoxify	CT	0.001	0.033	0.068	1.43
	Control	0.001	0.025	0.043	1.17
HateBERT	CT	0.006	0.131	0.581	11.50
	Control	0.001	0.047	0.100	3.10
RoBERTa	CT	0.001	0.012	0.003	1.07
	Control	0.001	0.003	0.002	0.10

Table 5: Toxicity/hate-speech score distributions over 3,000 sampled sentences per corpus. Detoxify reports Jigsaw toxicity; HateBERT and RoBERTa report ToxiGen-trained hate-speech probability. %flag = share exceeding the 0.5 threshold.

those exhibiting BITE-pattern activation. In other words, HateBERT detects the topic, not the harm.

The elevated BITE activation of the previous section is therefore invisible to detectors spanning the Jigsaw, HateBERT, and RoBERTa model families, even when one of them partially triggers on surface vocabulary. This is the gap the paper targets: not that CT discourse is never flagged, but that toxicity-centered detection, across the architectures and training regimes tested here, lacks features that respond to role-mediated coercive-control markers.

6 Discussion

The moralization hypothesis was not supported: the signed association statistic favors neutral psychology over morality in CT identity-term contexts. Combined with the unsupported emotional-staging and agency-deletion hypotheses, all three tested mechanisms are negative at corpus scale. The positive findings are elevated BITE coercion markers across all four dimensions and a demonstrated toxicity blind spot.

These findings point to a form of discourse that

falls outside current online-harms taxonomies. As the toxicity baselines show, two of three classifiers pass this text at rates indistinguishable from affirmative-care guidelines; the third triggers on identity vocabulary but does not discriminate coercive structure from benign uses of the same terms. The BITE-aligned rules, while limited by their CT-derived lexicon, identify a coercive-control register that toxicity models do not reach. The two nulls on emotional staging and agency deletion matter because they replace a loose “brainwashing-script” analogy with a more specific negative: at corpus scale, the measurable signatures are lexical, not procedural or syntactic, though this does not exclude their presence in individual encounters.

This extends WOA’s scope in two directions that bear on the tenth-anniversary theme. First, some harms may require role- and relation-aware modeling. Content-only classification does not currently model the authority relation between a credentialed practitioner and a dependent client, yet that relation is part of what makes the discourse harmful; the same words between peers carry different weight than words spoken by an authority to someone who depends on them. Practitioner–client power imbalance is an interpretive frame in this paper, not a measured finding: we did not tag speaker roles, quantify direct address to clients, or measure authority and dependency. Second, harmful language need not be aggressive. [Lu and Jurgens \(2022\)](#) made a related observation for seemingly-positive exclusionary rhetoric; we extend it to institutional therapeutic settings and give it a measurement through the BITE rule set.

For future work, a genre-matched control of pastoral counseling on grief, marriage, and addiction would help separate CT content from religious register. The moralization hypothesis itself requires independent annotation, balanced contexts, and validated source boundaries. Persuasion-technique ([Dimitrov et al., 2021](#); [Piskorski et al., 2023](#)) and framing-analysis systems ([Card et al., 2015](#)) are natural baselines beyond toxicity that this study did not test.

7 Conclusion

We tested three hypothesized mechanisms in conversion-therapy discourse—semantic moralization, staged emotional scripting, and grammatical agency-deletion—and found none supported at corpus scale. The positive findings are elevated BITE

coercion-control markers, with emotional control highest, coexisting with near-floor toxicity scores across three independent classifier families. As online-harms research reflects on its first decade, this case argues for taxonomies and detectors that model relational context, not only abusive tone, and for the value of reporting unsupported hypotheses alongside positive results.

Limitations

Our control corpus is small (26 documents, 16,948 sentences) relative to the CT corpus, and the two differ in genre, register, and institutional purpose. These confounds mean the cross-corpus association tests should be read as consistent with, not proof of, a CT-specific mechanism; a stronger design would add genre-matched controls (non-CT religious counseling, general clinical pathology texts). The BITE lexicon was derived from CT texts, so vocabulary selection inflates CT scores by construction; its single-author self-check supplies no independent validation or agreement estimate.

The signed association statistic contradicts the moralization hypothesis. Within the CT corpus, source-family effects range from $d = -9.64$ to $+0.33$, and 86.8% of selected identity-term context occurrences in the first-100 analysis come from one aggregate record, limiting representativeness. The first-100 headline estimates depend on corpus order; the sensitivity analysis uses a separate random-200 protocol and its bootstrap intervals are conditional on that fixed context sample.

The corpus is published material, not session transcripts; we characterize the prescriptive discourse of the CT movement, not individual therapeutic encounters. A fraction of feature hits fall on residual boilerplate and reference lists; because this inflates both corpora it does not threaten the contrasts, but absolute rates are lower bounds on signal quality. The GoEmotions classifier was trained on Reddit comments, a domain distant from religious prose, and the BITE measure is a rule set rather than a learned classifier validated against expert annotation. Our toxicity baseline covers three model families (Detoxify, ToxiGen-HateBERT, ToxiGen-RoBERTa); a fuller study would add Perspective API directly, prompt-based evaluation with large language models, and persuasion-technique or framing-analysis systems (Dimitrov et al., 2021; Piskorski et al., 2023; Card et al., 2015). The corpus is English and predominantly North American.

Ethics Statement

All analyzed materials are publicly available; the corpus contains no private session content. We do not reproduce harmful content verbatim; examples are paraphrased, and we will release derived features and code rather than the raw corpus to avoid improving the discoverability of CT materials. The corpus includes materials that may distress LGBTQ+ readers; our analysis characterizes their linguistic structure for harm-detection research, not to amplify their claims. We distinguish coercive institutional discourse from religious language as such: many markers in the corpus are religious, and our claim concerns how identity terms are constrained within an authority-bearing relationship, not religious speech in general. A detector built on these features must be evaluated for, and guarded against, flagging religious register per se, and against pathologizing the affirmative and survivor discourse that names CT framings in order to reject them.

References

- American Academy of Pediatrics. 2018. [Ensuring comprehensive care and support for transgender and gender-diverse children and adolescents](#). *Pediatrics*, 142(4):e20182162.
- American Psychological Association, Task Force on Appropriate Therapeutic Responses to Sexual Orientation. 2009. Report of the American Psychological Association task force on appropriate therapeutic responses to sexual orientation.
- Michael Billig. 2008. [The language of critical discourse analysis: The case of nominalization](#). *Discourse & Society*, 19(6):783–800.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- John R. Blosnich, Emmett R. Henderson, Robert W. S. Coulter, Jeremy T. Goldbach, and Ilan H. Meyer. 2020. [Sexual orientation change efforts, adverse childhood experiences, and suicide ideation and attempt among sexual minority adults, United States, 2016–2018](#). *American Journal of Public Health*, 110(7):1024–1030.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Dallas Card, Amber E. Boydston, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. [The media frames](#)

- corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 438–444.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. [SemEval-2021 task 6: Detection of persuasion techniques in texts and images](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.
- Robert M. Entman. 1993. [Framing: Toward clarification of a fractured paradigm](#). *Journal of Communication*, 43(4):51–58.
- Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper & Row, New York.
- Maarten Grootendorst. 2022. [BERTopic: Neural topic modeling with a class-based TF-IDF procedure](#). *arXiv preprint arXiv:2203.05794*.
- Laura Hanu and Unitary team. 2020. Detoxify. <https://github.com/unitaryai/detoxify>.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.
- Steven Hassan. 2015. *Combating Cult Mind Control: The Guide to Protection, Rescue and Recovery from Destructive Cults*, 4th edition. Freedom of Mind Press.
- Steven A. Hassan. 2020. *The BITE Model of Authoritarian Control: Undue Influence, Thought Reform, Brainwashing, Mind Control, Trafficking and the Law*. Ph.D. thesis, Fielding Graduate University.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength natural language processing in Python](#). Software.
- Robert Jay Lifton. 1961. *Thought Reform and the Psychology of Totalism: A Study of “Brainwashing” in China*. W. W. Norton & Company, New York.
- William A. Lingle. 2018. *Nominalizations, Agentless Passives, and Social Actor Mystification: Newspaper Editorials on the Greek Financial Crisis*. Ph.D. thesis, University of Birmingham.
- Sam Lowe. 2023. [roberta-base-go_emotions: A roberta model fine-tuned on goemotions](#). Hugging Face Model Hub.
- Christina Lu and David Jurgens. 2022. The subtle language of exclusion: Identifying the toxic speech of trans-exclusionary radical feminists. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 79–91.
- Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM)*, pages 608–611.
- Christy Mallory, Taylor N. T. Brown, and Kerith J. Conron. 2019. [Conversion therapy and LGBT youth \(update\)](#). Technical report, Williams Institute, UCLA School of Law, Los Angeles, CA.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 622–628.
- Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15162–15180.
- Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. [A framework for the computational linguistic analysis of dehumanization](#). *Frontiers in Artificial Intelligence*, 3:55.
- Verónica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3391–3401.
- Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.
- Amy Przeworski, Emily Peterson, and Alison Piedra. 2021. [A systematic review of the efficacy, harmful effects, and ethical issues related to sexual orientation change efforts](#). *Clinical Psychology: Science and Practice*, 28(1):81–100.

Rahul Rawat and 1 others. 2024. [Hate speech detection in social media: Techniques, recent trends, and future challenges](#). *WIREs Computational Statistics*, 16(2):e1648.

Travis Salway, Olivier Ferlatte, Dionne Gesink, and Nathan J. Lachowsky. 2020. [Prevalence of exposure to sexual orientation change efforts and associated sociodemographic characteristics and psychosocial health outcomes among Canadian sexual minority men](#). *The Canadian Journal of Psychiatry*, 65(7):502–509.

Edgar H. Schein. 1961. *Coercive Persuasion: A Socio-Psychological Analysis of the “Brainwashing” of American Civilian Prisoners by the Chinese Communists*. W. W. Norton, New York.

Amir Sepehri, Roodabeh S. Mirshafiee, and David M. Markowitz. 2022. [PassivePy: A tool to automatically identify passive voice in text](#). GitHub repository.

Ariel Shidlo and Michael Schroeder. 2002. [Changing sexual orientation: A consumers’ report](#). *Professional Psychology: Research and Practice*, 33(3):249–259.

Margaret Thaler Singer and Janja Lalich. 1995. *Cults in Our Midst: The Continuing Fight Against Their Hidden Menace*. Jossey-Bass, San Francisco.

Substance Abuse and Mental Health Services Administration. 2015. Ending conversion therapy: Supporting and affirming LGBTQ youth. HHS Publication No. (SMA) 15-4928.

Javier Torregrosa and 1 others. 2023. [A survey on extremism analysis using natural language processing: Definitions, literature review, trends and challenges](#). *Journal of Ambient Intelligence and Humanized Computing*, 14:9869–9905.

Nguyen K. Tran, Elle Lett, and 1 others. 2024. Conversion practice recall and mental health symptoms in sexual and gender minority adults in the USA: a cross-sectional study. *The Lancet Psychiatry*, 11(11):879–889.

A Association-Test Term Sets

The WEAT/SEAT-style tests measure how identity terms associate with three attribute sets relative to one another. The sets were fixed in advance from the clinical and religious vocabulary of the domain, not tuned to the result.

Identity (target). homosexuality, gay, lesbian, orientation, attraction, desire, identity, same-sex, bisexual, queer

Morality (attribute). sin, temptation, redemption, purity, holy, righteousness, repentance, fallen, sanctification, obedience

Pathology (attribute). disorder, disease, addiction, brokenness, wound, dysfunction, compulsion, condition, abnormal, deviant

Neutral psychology (contrast). well-being, self-acceptance, resilience, growth, authentic, congruence, integration, flourishing

All listed terms meet the three-occurrence threshold in CT. In the control, all ten identity terms remain, but the morality set retains only *sin* (three sentences); the neutral set excludes *flourishing*; pathology retains disorder, disease, addiction, dysfunction, condition, and abnormal. Missing terms are omitted before computing means and permutations. Thus the control morality result ($d = +0.10$, $p = 0.85$) contrasts sin with seven neutral terms and is not a ten-term morality comparison matched to CT.

B BITE Rule Set

Each dimension is operationalized as a set of interpretable features (Table 6); a sentence activates a dimension if any feature fires, subject to clinical-exclusion patterns. The most frequent features overall are loaded language, prescriptive modals, and shame/guilt induction.

A single-author review examined 50 flagged and 50 unflagged CT sentences for each of the two most-active dimensions. This was a self-check: there was no second annotator or inter-annotator agreement statistic, so these judgments do not validate the rules.

C Within-CT Provenance

In response to the question of whether contextual factors influence the observed patterns, Table 7 reports the first-100 association statistic within five provenance families identified by filename. Effects range from $d = -9.64$ (JHS) to $+0.33$ (Desert Stream). The JHS–Desert Stream contrast is $\Delta d = -9.96$ (document permutation $p = 0.001$). In the

Dimension	Features
Information	loaded language; thought-terminating clichés; us-vs-them; source distrust; testimony as proof
Emotional	shame/guilt induction; fear induction; love bombing; conditional acceptance; hope through change; confession–relief cycle
Behavior	imperatives; prescriptive modals; accountability language; behavioral rules; isolation restrictions; program commitment
Thought	black-and-white framing; prescriptive modals; redefining reality; identity denial; root-cause framing

Table 6: Features comprising each BITE dimension. Activation rates in Table 4.

full first-100 run, 837 of 964 selected identity-term context occurrences (86.8%) come from the single aggregate ATCSI website record; eight identity terms draw all 100 contexts from it. Validated genre and audience labels are absent; these comparisons remain exploratory.

Stratum	Records	Sentences	$d_{I,M}$	p
All CT	2,388	119,976	-5.11	0.0002
JHS records	1,058	53,324	-9.64	0.0001
Reparative clinical	74	4,068	-0.35	0.824
Exodus sources	478	30,563	-2.91	0.010
Desert Stream	517	13,759	+0.33	0.861
Testimonial networks	259	11,764	-1.17	0.259

Table 7: Within-CT provenance analysis. $d_{I,M}$: first-100 morality-minus-neutral association; negative values favor neutral psychology. p -values: 10,000 two-sided attribute-label permutations, nominal.