Fraud, Fairness, and Framing: How State Legislators Tweet about Elections

Isaac Pollert[1], Bruce Desmarais[1], Muhammed Cifci[1], Gary Fong[1], Ishita Gopal[1]

[1]Department of Political Science, Pennsylvania State University

April 21, 2026

Abstract

The 2020 presidential election triggered an unprecedented volume of elite commentary on election integrity, but state legislators—who hold direct authority over election administration—have received far less scholarly attention than federal officials. Using over 90,355 election-related tweets from 3,235 state legislators in 2020–2021, we model both the *volume* of election integrity tweets (legislator-week counts, hierarchical negative binomial) and the *binary decision* to post at least one tweet challenging electoral legitimacy (legislator-level hierarchical linear probability model, with a Firth penalized logistic sensitivity). We find a clean division: Democrats dominate the volume of election-integrity discourse, and engagement received in the previous week predicts next-week posting; but whether a legislator ever posts delegitimizing content is overwhelmingly predicted by party. Republicans are roughly 1.79 times as likely as otherwise-similar Democrats to have posted at least one challenge tweet during the study period. Within the Republican caucus, no measured structural predictor—vote share, state democratic performance, legislative professionalism, or state-level fraud cases—distinguishes challenge-posters from non-posters, an honest null that does not pin down a specific mechanism but that is inconsistent with local electoral incentives being the primary driver. Results are robust across eight alternative specifications. Our

findings document how delegitimizing rhetoric has diffused to the state legislative level and clarify what is and is not a plausible mechanism for it.

Keywords: Elections; State Politics; Social Media; Election Integrity; Elite Rhetoric; Partisan Polarization

1 Introduction

The 2020 presidential election became a flashpoint for competing claims about electoral legitimacy, with consequences that continue to reverberate through American democracy. Election integrity is central to the stability of democratic systems (Przeworski, 2018): it sustains public trust in settling disputes through institutional means, particularly among electoral losers who might otherwise resort to post-election violence (Donno, Roussias and Scharpf, 2022) or non-compliance with public regulations (Schnaudt, 2025). Yet Americans now hold sharply divergent views of what "election integrity" means, with Republicans focused on voter fraud and Democrats emphasizing voting accessibility. These divergent frames have become increasingly polarized, with significant consequences for democratic governance.

State legislators play a particularly consequential role in this debate: they have direct authority over election administration and voting rules, yet their rhetoric on elections has received far less scrutiny than that of federal officials. Increasingly, state legislators communicate directly with the public through social media. How they frame elections—whether as fair or as fraught with fraud—shapes how the public views democratic institutions and the legitimacy of electoral outcomes. Prior work has shown that elite messaging about election fraud can shift public perceptions of electoral integrity (Donno, Roussias and Scharpf, 2022; Berlinski et al., 2021), making the question of who posts what on social media both consequential and theoretically important.

This study examines how state legislators discussed election integrity on Twitter during 2020–2021. We ask two questions. First, what structural and behavioral factors shape the *volume* of a legislator's election-integrity discourse? Second, what predicts whether a legislator ever posts content that *challenges* electoral legitimacy? These two outcomes turn out to have very different drivers, and separating them is central to the paper.

We use a dataset of 90,355 election-related tweets from 3,235 state legislators (of whom 2,026 are Democrats and 1,210 are Republicans). Tweets were classified with a two-stage

supervised classifier; 5,492 tweets (6.08% of the sample) are classified as challenging electoral legitimacy. We model weekly tweet volume with a hierarchical negative binomial model, and the binary decision to ever post a challenge tweet with a hierarchical linear probability model (with a Firth penalized logistic sensitivity). Both specifications include random intercepts for state; the count model also includes random intercepts for legislator and week fixed effects.

We find that election-integrity tweet volume is shaped primarily by partisanship and engagement: Democratic legislators tweet more about election integrity than Republicans (-0.133, IRR $= 0.875$, $p < .001$), and both lagged likes (0.028, $p < .001$) and lagged retweets (0.081, $p < .001$) positively predict subsequent posting. None of the state-level structural variables—democratic performance, legislative professionalism, or documented fraud cases—significantly predict tweet volume once legislator and state clustering are absorbed. For the binary decision to ever post a challenge tweet, party is dominant: being Republican is associated with a 15.6-percentage-point higher probability of ever posting challenge content (0.156, $p < .001$; Firth logit odds ratio $= 1.79$). None of the measured structural or electoral predictors significantly predict challenge posting, a null we report honestly without overclaiming a particular mechanism behind it.

2 Background and Hypotheses

2.1 Two outcomes, not one

A first-order point about state legislative discourse on election integrity is that "how much" and "how negative" are distinct questions with distinct drivers. A legislator who tweets frequently about election administration, voting rights, or ballot access can be tweeting supportively about electoral institutions; a legislator who tweets rarely but, when they do, alleges fraud or rigging is delegitimizing those same institutions. Treating these as one outcome would obscure what the data can actually say. It would also obscure what turns out to be the most striking finding in this paper: that the volume of election-integrity discourse is

a predominantly *Democratic* phenomenon, while the decision to post delegitimizing content is a predominantly *Republican* one. Those two facts are in tension only if one assumes (as much of the public debate implicitly does) that the party producing more election talk is the party producing more election-undermining talk. Our data show the opposite.

We therefore separate the *intensity* of election-integrity discourse (count of election-related tweets per legislator-week) from the *decision to delegitimize* (a legislator-level binary: did this legislator ever post a challenge tweet during the study period?). The count outcome is modeled at the legislator-week level to capture within-legislator temporal variation and to align with the lag structure of our engagement variables. The binary outcome is modeled at the legislator level because delegitimization is a qualitative threshold that is meaningful whether it happened once or fifty times during the sample. Conceptually, the two outcomes pick up different decisions: the count outcome is about how much bandwidth a legislator spends on election-related topics overall, while the binary outcome is about whether that legislator crosses a normative line into challenging the legitimacy of electoral institutions. These are behaviorally distinct, they have different drivers in our data, and they have different implications for democratic health. A finding that election-integrity discourse volume is Democratic matters for how we think about which party is paying attention to elections as a topic; a finding that delegitimization is Republican matters for which party is producing the elite supply of rhetoric that could undermine public trust in elections. Our paper reports both findings and keeps them separate.

2.2 Partisan asymmetry in election rhetoric

Research on elite rhetoric about elections has grown substantially since 2020, though attention has concentrated on federal officials. LaPlant, Lee and LaPlant (2024) find that Republican House members were more likely to spread "Stop the Steal" claims and vote against certification, with prior Trump support predicting this behavior. We extend that line of work to the far larger and more diverse population of state legislators, who hold direct

authority over election administration and have received less scholarly attention.

A substantial body of research documents partisan asymmetry in beliefs about election fraud. Berlinski et al. (2021) show that Republican voters are substantially more likely than Democrats to believe that fraud affected the 2020 election outcome, even after exposure to corrections. Edelson et al. (2021) document similar patterns among the mass public. Norris (2019) shows that perceptions of election quality diverge sharply along partisan lines even where institutional conditions are similar. These patterns of asymmetric belief create fertile conditions for partisan divergence in elite rhetoric: if Republican state legislators share their constituents' skepticism about election integrity, they should be systematically more likely to communicate that skepticism through social media.

> *H1 (volume).* Democratic and Republican state legislators differ in how much they post about election integrity. Our prior, based on the broader framing of election integrity around voting access on the Democratic side and around episodic fraud concerns on the Republican side, is that Democrats post *more* about election integrity in total.

> *H2 (delegitimizing content).* Republican state legislators are more likely than otherwise-similar Democrats to have posted at least one tweet challenging electoral legitimacy. The partisan gap in delegitimizing content is the dominant feature of the data.

2.3 Engagement feedback

Social media behavior is not only strategic; it is also *reactive*. When a legislator's tweets receive attention—likes and retweets—that feedback signal can reinforce further posting on the same topic. Laboratory and observational work in social media dynamics has shown that emotionally charged or morally charged content travels further (Brady et al., 2017) and that creators respond to engagement signals in subsequent posting. We test this mechanism

with lagged engagement: likes and retweets on a legislator's election tweets in week $t-1$ predicting the count of election tweets in week t. The one-week lag rules out mechanical contemporaneous coupling.

> *H3 (engagement feedback).* Legislators whose election tweets received more attention in the previous week (likes and retweets at $t-1$) will post more election tweets in week t.

2.4 Within-party variation: a null to report honestly

A natural question, especially given the large between-party gap in delegitimizing content, is whether within-Republican variation in challenge posting can be explained by local electoral incentives. One plausible mechanism is primary vulnerability: in safe Republican general-election seats, legislators may face their main electoral threat from ideologically extreme primary electorates and therefore have stronger incentives to signal partisan loyalty through election challenges. A competing mechanism is local conditions: legislators in states with weaker democratic performance or with more actual fraud cases may have more reason—or more cover—to post delegitimizing content. Finally, legislative professionalism could capture variation in the institutional and ideological context of rhetoric.

We do not register directional hypotheses for these within-Republican mechanisms, because their directional predictions depend on mechanisms we cannot directly measure. Instead, we estimate a within-Republican model with vote share, state democratic performance, legislative professionalism, and documented fraud cases as predictors, and we report the result honestly. If none of these structural predictors is significant, that is informative as a null: it narrows the space of plausible mechanisms but does not, on its own, demonstrate any particular alternative.

We complement this with a within-Democrat model of election-integrity tweet *volume*. The directional prediction here is more defensible: if Democratic legislators' election-integrity discourse responds to institutional conditions in their state, we should see Democrats in

states with weaker democratic performance, or with more documented fraud cases, tweeting more about election integrity.

3 Data and Methods

3.1 Tweet collection and legislator identification

We compiled a comprehensive dataset of tweets from state legislators (state senators and state representatives) across all fifty states over the 2020–2021 period. Legislators and their Twitter accounts were identified through multiple sources, including state legislative websites, Ballotpedia, OpenSecrets, and the National Conference of State Legislatures (NCSL), to maximize coverage.

Tweets were collected with the Twitter Academic Research API, using a keyword filter for election-related content ("election," "voting," "vote," "ballot," "electoral," "voter," "poll," "election integrity," "election security," "voter fraud"). The collection window spans posts with a `created_date` from January 1, 2020 through December 31, 2021. After filtering to legislators active in 2020 or later, the raw corpus of 90,712 tweets is reduced to 90,649 election-related tweets from active legislators, of which 90,355 are from Democratic or Republican legislators. Roughly 63.2% of those are further dropped at the modeling stage due to missing values on at least one standardized predictor (most commonly state-level democratic performance or legislative professionalism in the small number of territories or chambers for which our state-level data sources do not supply values). The final count-model sample is 33,288 legislator-weeks and the final challenge-model sample is 3,237 legislators (2,026 Democrats, 1,210 Republicans).

An important distinction is between *original* tweets and *retweets*. We identify retweets by matching the "RT" prefix pattern (`^RT[ :@]`) in three available text fields (`text_corpse`, `text`, and `hydrated_text`) and flagging a tweet as a retweet if any of the three match. This catches the classic manually-prepended retweet format as well as "RT:" variants, but we note

a remaining limitation: Twitter's API also records native retweets (where the user clicked the retweet button without adding a textual prefix), and these are not always marked with a detectable prefix in the text fields we have. The robustness check using "originals only" (Section 4.5, column 6) therefore retains some native retweets we cannot identify, which if anything biases that check against finding a difference from the main specification. Table 2 shows the share of each party-by-stance cell that our detection flags as original vs. retweeted.

3.2 Tweet classification

Classification proceeds in two sequential stages. *Stage 1* is the keyword filter described in Section 3.1: it produces the pool of *election-relevant candidates* by retaining tweets that contain at least one election-related keyword. This filter is high-recall by design but not high-precision: many retained tweets use the keywords in ways that are not about electoral legitimacy (e.g., policy tweets that happen to mention "vote," or tweets about unrelated elections), so a second-pass classifier is needed to separate tweets that *challenge* electoral legitimacy from those that merely mention election-related topics.

Stage 2 is a binary text classifier applied to the keyword-filtered pool. We fine-tuned **PoliBERTweet** (Kawintiranon and Singh, 2022)—a BERT (Devlin et al., 2019) variant pre-trained on a large corpus of political tweets—for the binary task of distinguishing tweets that challenge electoral legitimacy from tweets that do not. Fine-tuning and inference used the Hugging Face `transformers` library (Wolf et al., 2020); we initialized from `kornosk/polibertweet-mlm`, tokenized with the corresponding tokenizer (sequences padded and truncated to 128 tokens), and trained a classification head with standard cross-entropy loss. Model selection during training used the area under the precision-recall curve (AUPRC) on a held-out validation slice, which is the appropriate selection metric for imbalanced binary classification tasks in which the positive class is rare.

The initial training set consisted of 500 hand-labeled tweets drawn from the keyword-filtered pool, supplemented by ten rounds of active learning (Settles, 2009). In each round,

the 50 tweets for which the current model was most uncertain (predicted probability closest to 0.5) were drawn from the remaining unlabeled pool, hand-coded by the primary coder, and added to the training set only. A held-out test set of 200 tweets was drawn randomly at the beginning of training and was excluded from the active-learning pool throughout, so the final performance metrics are computed on examples the classifier never saw during iterative retraining. The final Stage 2 classifier achieved 92% accuracy on the held-out test set, Cohen's $\kappa = .87$ (Cohen, 1960) against a second independent coder, and precision, recall, and F_1 of 0.88, 0.86, and 0.87 for the rarer challenge class.

3.3 Variables

Dependent variables. For the *volume* analysis we use the legislator-week count of election-related tweets. For the *challenge* analysis we use a legislator-level binary indicator equal to 1 if the legislator ever posted a tweet classified as challenging electoral legitimacy during the study period, and 0 otherwise.

Party. Republican legislators are coded 1 and Democrats 0. Independent legislators are excluded from party-specific analyses.

Lagged engagement. Likes and retweets on election-related tweets enter the count model as one-week lagged predictors of tweet volume. At the legislator-week panel level, these are the sum of likes and retweets received on election tweets posted in the previous calendar week, scaled by 1,000 and then standardized. We use a one-week lag rather than a shorter or longer interval because the weekly cadence aligns with both the legislator-week panel structure and typical media-cycle turnover: engagement received within a day of posting would risk mechanical contemporaneous coupling, while a multi-week lag would obscure the short-term reinforcement dynamic we are trying to detect. Section 4.5 includes a two-week-lag robustness check to show that the feedback pattern does not depend on the one-week choice. Engagement also enters the challenge model, but measured on *challenge tweets specifically*: the legislator's mean-per-active-week lagged likes and retweets on tweets

classified as challenging electoral legitimacy. An *active week* is a week in which the legislator posted at least one tweet of the relevant type (at least one election tweet for the count-model aggregation, at least one challenge tweet for the challenge-model aggregation); inactive weeks are excluded from the mean so that a legislator who posts challenging content sporadically but with high engagement is not diluted by her silent weeks. The challenge LPM is legislator-level rather than legislator-week, so the engagement predictor must be a single per-legislator summary; the mean per active week is the natural level-to-level reduction from the weekly panel, matching the outcome to the mechanism (attention to delegitimizing content predicting challenge posting) and differing from the count model, where engagement is measured on all election-integrity tweets. Using the mean per active week—rather than the total or the mean per calendar week—prevents conflating activity duration (how many weeks the legislator was posting) with intensity (how much attention each active week drew).

Vote share. Vote percentage won in the most recent general election before the study period, scaled.

State-level predictors. Democratic performance is operationalized from the Perceptions of Electoral Integrity (PEI) project's sub-national index for U.S. states (Norris, 2019). Specifically, we use the 2020 state-level PEI index value (single composite score per state) from the file `dem_index.csv` provided with the replication materials, filtered to year 2020. Higher values indicate better expert-assessed democratic performance. The PEI index is itself a composite of expert ratings across eleven sub-areas of electoral integrity (electoral laws, procedures, boundaries, voter registration, party registration, campaign media, campaign finance, voting process, vote count, results, and electoral authorities); the file distributed to replicators contains the pre-aggregated 2020 index, not the component sub-scores. *Legislative professionalism* uses the Squire index (Squire, 2017), a composite of chamber salary, session length, and per-member staff; we merge the 2021-vintage state values from `legprof.csv`. *Fraud cases* is a state-level count of documented voter-fraud cases from the Heritage Foundation's election-fraud database, which we acknowledge is curated by a right-leaning institution

and may systematically overweight fraud claims in states whose election administration it has chosen to scrutinize. We include it as a predictor of *elite attention* to fraud rather than as a neutral measure of fraud prevalence, and we report a sensitivity note in the Discussion. States not appearing in the Heritage database are coded as zero, which treats absence as a structural zero rather than as informative missingness; this choice is conservative for the within-GOP null (it biases that null toward "no effect" only if Heritage systematically undercounts fraud claims in GOP-leaning states, which is unlikely).

Controls. We include legislator-level controls for chamber (House = 1, Senate = 0), race (White = 1), and gender (female = 1). The race control is a coarse binary: non-White legislators of all backgrounds are pooled, which obscures variation across racial groups that may be substantively important. We retain the binary because finer categorical breakdowns produce very thin cells in the challenge model (almost all challenge-posters are White), but we flag the coarseness as a real limitation of our control-variable specification. All continuous predictors are standardized ($M = 0, SD = 1$) for coefficient comparability.

3.4 Analytic approach

Our data have a nested structure: legislator-weeks are nested in legislators, and legislators are nested in states. For the weekly tweet-count outcome, we fit a **hierarchical negative binomial (NB)** model with random intercepts for state and for legislator, plus week fixed effects. Negative binomial regression (Cameron and Trivedi, 2013) is appropriate for overdispersed count data: the Pearson overdispersion from the fitted hierarchical NB model is $\hat{\phi} = 0.947$, computed conditional on the random effects (i.e., after the state and legislator variance components absorb between-cluster variation). The count model is estimated in `glmmTMB` (Brooks et al., 2017) with the `nbinom2` family. We report the estimated dispersion parameter $\theta = 4.72$ alongside the fixed-effect coefficients so the reader can evaluate how much residual variance remains after the random effects.

For the binary challenge outcome, we fit a **hierarchical linear probability model**

(LPM) at the legislator level, with a random intercept for state. Because each legislator contributes only one observation to this outcome, a legislator-level random effect is not identified; state clustering is the relevant nesting. We chose an LPM rather than a hierarchical logistic regression for a specific reason: an earlier version of this paper used a hierarchical logit (`glmer`) and found that the retweet coefficient was implausibly large and the standard errors on the other predictors were correspondingly inflated. The cause was *quasi-complete separation*: Republicans with high retweet engagement were nearly all classified as challenge-posters, which made the logit likelihood surface flat along the retweet direction and the MLE unstable. A linear-link model does not have this problem: its likelihood is quadratic, not flat along separation directions, so engagement variables can be included without inflating other coefficients. Coefficients from the LPM are interpretable as percentage-point changes in the probability of ever posting a challenge tweet. The LPM is estimated via `lmer` (Bates et al., 2015) with REML and a state random intercept. Critically, engagement in the challenge model is measured on *challenge tweets specifically*: the legislator's mean-per-active-week lagged likes and retweets on tweets classified as challenging electoral legitimacy. This matches the outcome to the mechanism—the question is whether attention to delegitimizing content predicts continued challenge posting—and differs from the count model, where engagement is measured on all election-integrity tweets (because the count outcome is all election tweets, not just challenge tweets). Both aggregations use the mean per active week to avoid conflating activity duration with intensity.

As a cross-check on the main LPM, we also fit a **Firth penalized logistic regression** (Firth, 1993; Heinze and Schemper, 2002) with the same predictor set (engagement included). Firth's penalty is explicitly designed to handle separation by penalizing the likelihood toward the origin, producing stable maximum-likelihood estimates where standard `glm` or `glmer` would fail. The Firth regression is non-hierarchical (`logistf` (Heinze and Schemper, 2002) does not support random effects), so it does not substitute for the main LPM, but it provides an odds-ratio-scale interpretation of the same specification. We report the Firth OR for party

alongside the LPM estimate in Section 4.2.3.

We report intraclass correlation coefficients (ICCs) from the fitted hierarchical models themselves, using `performance::icc()`, which applies the variance-partition approach appropriate for generalized linear mixed models (Nakagawa, Johnson and Schielzeth, 2017). This differs from the earlier version of this paper, which reported ICCs from auxiliary Gaussian null models on count data—a mis-specification for count outcomes that we have corrected.

Within-party models. For the *volume* outcome we estimate within-party negative binomial count models with cluster-robust standard errors by state (Cameron, Gelbach and Miller, 2011), one for each party. We do not split the *challenge* outcome by party: challenge-tweet engagement is structurally sparse (most legislators never post challenge tweets, so their challenge engagement is zero by definition), and splitting by party creates near-perfect collinearity between challenge engagement and the outcome within each subsample. The main hierarchical LPM, which includes party as a predictor, is the appropriate specification for the challenge outcome. Both within-party count models use the same structural predictor set (vote share, state democratic performance, legislative professionalism, documented fraud cases, chamber, race, gender) and include legislator-level mean-per-active-week engagement on all election tweets, with cluster-robust standard errors by state.

4 Results

4.1 Descriptive patterns

Our sample comprises 90,355 election-related tweets from 3,235 Democratic and Republican state legislators active in 2020–2021 (2,026 Democrats, 1,210 Republicans). Of these, 5,492 tweets (6.08% of the sample) are classified as challenging electoral legitimacy. The challenge rate differs dramatically by party: 2.14% of Democratic tweets and 18.89% of Republican tweets are challenge tweets, roughly a tenfold gap in tweet-level rates. At the legislator level, 22.0% of Democrats and 35.4% of Republicans posted at least one challenge tweet during the

study period; 73.0% of legislators posted no challenge tweets at all.

Table 1 presents summary statistics by party. Democrats, the larger group in the sample, average 34.1 election-integrity tweets per legislator (SD = 83.5, median 10); Republicans average 17.6 (SD = 79.5, median 4). Challenge tweets per legislator are 0.73 on average for Democrats and 3.32 for Republicans, reflecting the concentration of challenge content among a smaller set of Republican legislators.

Table 1: Summary Statistics for State Legislators by Party

	Democrats			Republicans		
Variable	Mean	SD	Median	Mean	SD	Median
Election-integrity tweets per legislator	34.1	83.5	10	17.6	79.5	4
Challenge tweets per legislator	0.73	4.61	0	3.32	27.74	0
Vote share (%)	70.6	23.2	66.8	67.9	20.0	63.5
Legislators		2,026			1,210	
Election-integrity tweets (total)		69,107			21,248	
Challenge tweets (total)		1,478			4,014	
% of tweets that challenge		2.14%			18.89%	
% of legislators who ever challenged		22.0%			35.4%	

Election-related tweets from state legislators, January 2020 — December 2021. Independent legislators omitted from party comparison.

The tweet-count outcome is heavily overdispersed at the legislator-week level: the variance-to-mean ratio is 7.74 for election-integrity tweet counts and 9.66 for challenge tweet counts. Both exceed the equidispersion assumption of Poisson regression by a large margin, motivating our use of negative binomial regression (see also Appendix A for the explicit Poisson–NB comparison).

Figure 1 plots biweekly tweet volume faceted by tweet stance and colored by party. The highest single-month volume falls in 2021-01 (11,015 tweets, 2.1% challenge rate), around Election Day and Capitol insurrection events. The peak *challenge rate* falls in 2020-12 (4,468 tweets, 16.1% challenge rate), coinciding with the intensification of post-election legal challenges. Challenge volume is not only lower than non-challenge volume—it also responds to different triggers.

Table 2: Retweets vs. Original Posts by Party and Tweet Stance

Party	Stance	Type	N	Med. RT	Med. Likes	% of Stance
Dem.	Challenge	Original	645	1	6	43.6
		Retweet	833	83	0	56.4
	Non-Challenge	Original	37,582	2	9	55.6
		Retweet	30,047	61	0	44.4
Rep.	Challenge	Original	1,693	2	8	42.2
		Retweet	2,321	669	0	57.8
	Non-Challenge	Original	9,754	1	4	56.6
		Retweet	7,480	98	0	43.4

Election-related tweets from state legislators, 2020–2021. "Original" rows are tweets the legislator authored; "Retweet" rows are tweets the legislator retweeted from another account. "Med. RT" is the per-tweet median retweet count: for originals it is the number of times the legislator's own tweet was retweeted, and for retweets it is the lifetime retweet count on the *underlying original* tweet (which is what the Twitter/X API reports as the retweet count of a retweet). That asymmetry explains why retweet rows have much higher median RT counts than original rows: retweets inherit the engagement of the source post. "Med. Likes" is 0 for retweets because the Twitter/X API reports 0 likes on retweet events, attributing the like count to the original post only. "% of Stance" is the share of tweets in that party × stance cell that are original vs. retweeted.

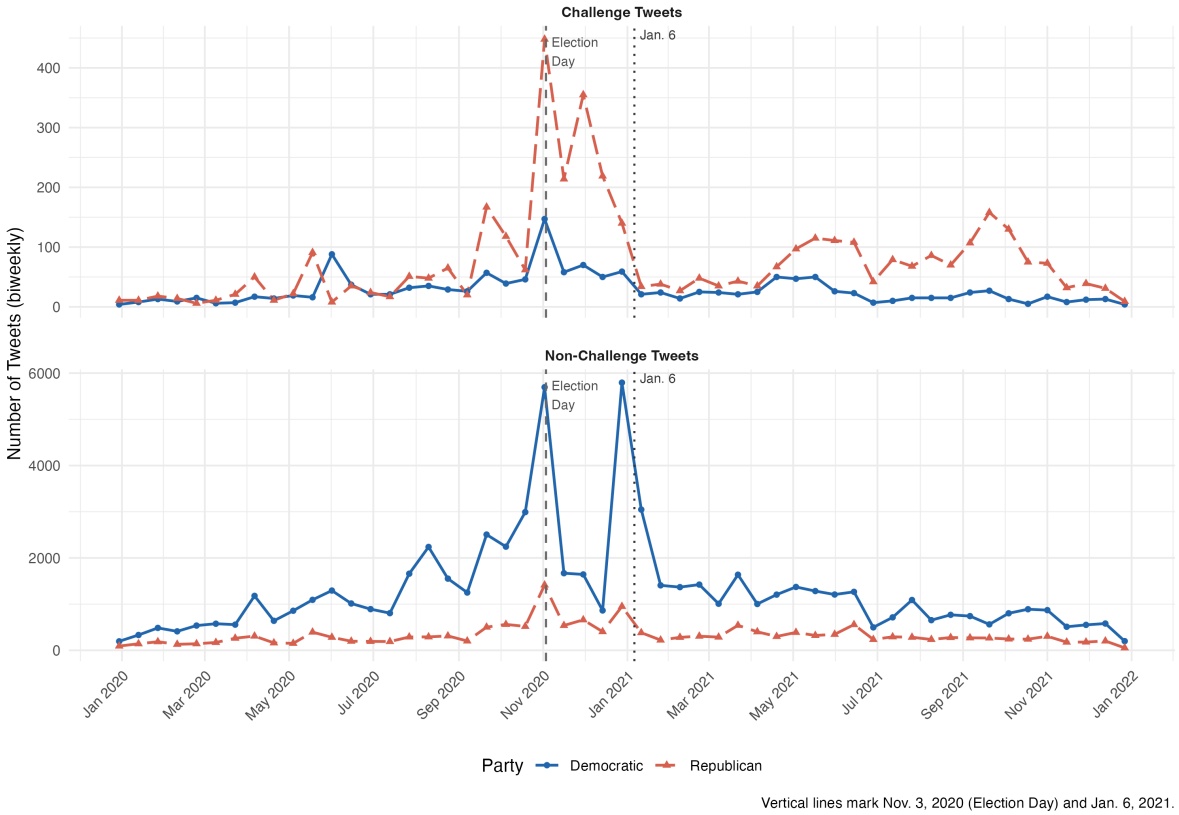

Figure 1: Biweekly volume of election-related tweets by party, 2020–2021. Top panel: challenge tweets. Bottom panel: non-challenge tweets. Vertical reference lines mark Nov. 3, 2020 (Election Day) and Jan. 6, 2021.

Engagement with tweets, finally, differs sharply by stance. The median engagement (likes plus retweets) of a challenge tweet is 96, compared with 18 for non-challenge tweets; at the 90th percentile the gap is 7,561 vs. 1,715, and at the 99th percentile 50,948 vs. 19,593. Challenge content attracts disproportionately large engagement in the tails—a pattern consistent with algorithmic amplification of adversarial content.

4.2 Main models

4.2.1 Intraclass correlations and justification for hierarchical specifications

The intraclass correlation coefficient (ICC) reports the share of total residual variance that is attributable to between-cluster differences rather than within-cluster variation. In a multilevel model, an ICC near zero means observations within a cluster (e.g., legislators within a state) are no more similar to each other than to legislators in other states, and a pooled single-level regression would suffice; an ICC above roughly 0.05–0.10 is typically taken as evidence that the within-cluster dependence is non-negligible and that ignoring the clustering would understate standard errors and risk biased inference (Nakagawa, Johnson and Schielzeth, 2017). Reporting the ICC therefore serves two purposes here: first, it documents how much of the outcome variation lives between states and between legislators once fixed predictors have been accounted for, and second, it justifies the hierarchical specification against a simpler pooled alternative.

Computed via the variance-partition approach appropriate for generalized linear mixed models (`performance::icc` applied to the fitted models), the adjusted ICC for the tweet-count outcome is 0.228—the joint share of residual variance attributable to the state and legislator random intercepts together—and the adjusted ICC for the challenge outcome is 0.034 (state random intercept only). The count-outcome ICC is substantial and clearly justifies the two-level (state + legislator) hierarchical specification: a meaningful fraction of the variation in weekly posting cannot be explained without the random intercepts. The challenge outcome clusters less strongly at the state level, but the state random intercept still

modestly improves the fit and is retained, and its presence changes the state-level predictor standard errors in a direction that is more conservative than the pooled alternative. We note that in the earlier version of this paper we reported ICCs computed from Gaussian null models (`lmer`) on count data, which was a mis-specification that the present numbers correct.

4.2.2 H1 and H3: Volume of election integrity discourse

Table 3 presents the hierarchical negative binomial model of legislator-week election-integrity tweet counts. Consistent with H1, the Republican indicator is negative and significant (-0.133, IRR $= 0.875$, $p < .001$): holding other predictors at their means, Democratic legislators post more election-integrity tweets per week than otherwise-similar Republicans.

Table 3: Hierarchical Negative Binomial Model—Predictors of Election-Related Tweet Count (Legislator-Week). Random intercepts for state and legislator. Week fixed effects included but not shown. $N = 33{,}288$ legislator-weeks.

	Weekly Tweet Count	
Variable	Coefficient	IRR
Likes $t-1$ (scaled)	0.028^{***} (0.004)	1.028
Retweets $t-1$ (scaled)	0.081^{***} (0.004)	1.084
Republican	-0.133^{***} (0.025)	0.875
House Member	$-0.044^{\dagger}$ (0.023)	0.957
White	0.023 (0.026)	1.023
Female	0.008 (0.022)	1.008
Vote Share (scaled)	-0.001 (0.012)	0.999
Democratic Performance (scaled)	-0.013 (0.021)	0.987
Leg. Professionalism (scaled)	-0.011 (0.026)	0.989
Fraud Cases (scaled)	0.016 (0.024)	1.017
Random effects	State ($\sigma = 0.130$); Legislator ($\sigma = 0.387$)	
AIC	121,604	
BIC	122,589	

Standard errors in parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. IRR = incidence rate ratio (e^b). Continuous predictors standardized ($M = 0, SD = 1$). State-level ICC = 0.228; legislator-level ICC = 0.200.

Consistent with H3, both lagged likes (0.028, $p < .001$) and lagged retweets (0.081, $p < .001$) are positive and significant: legislators who received more engagement on election

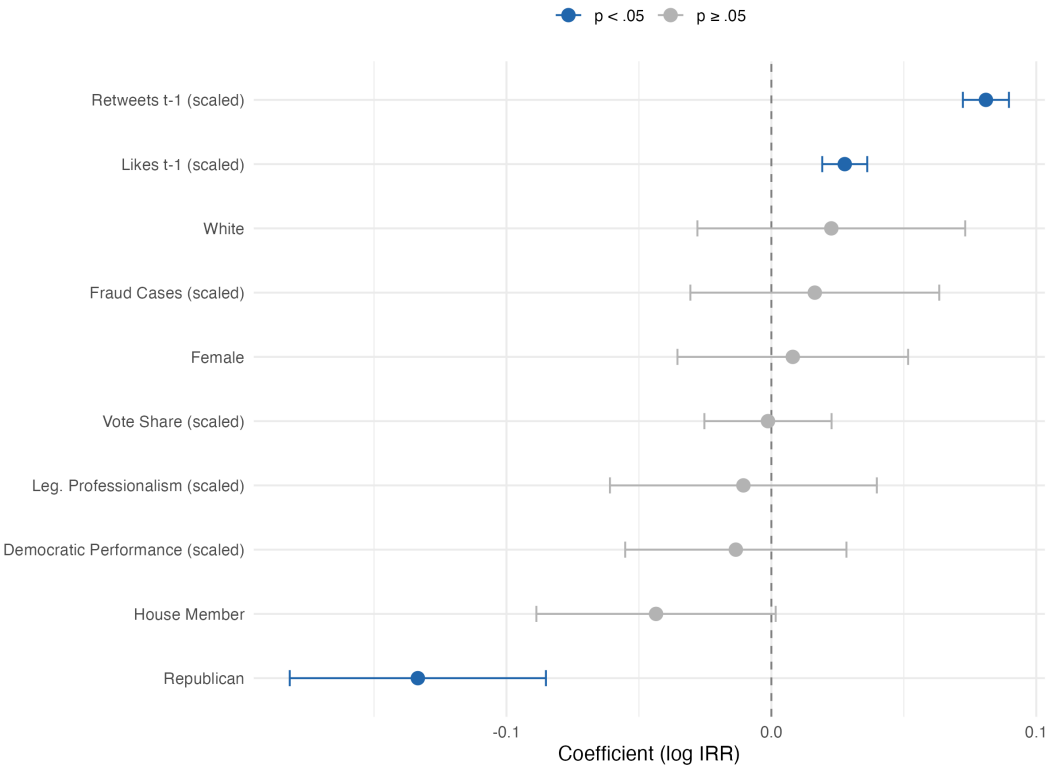

Figure 2: Coefficient plot for the hierarchical negative binomial count model. Points show point estimates with 95% confidence intervals. All continuous predictors standardized.

tweets last week post more election tweets this week. Figure 3 visualizes the marginal prediction, showing the engagement-feedback relationship separately by party. The temporal lag rules out mechanical same-week coupling, and the pattern is consistent with an algorithmic reinforcement interpretation.

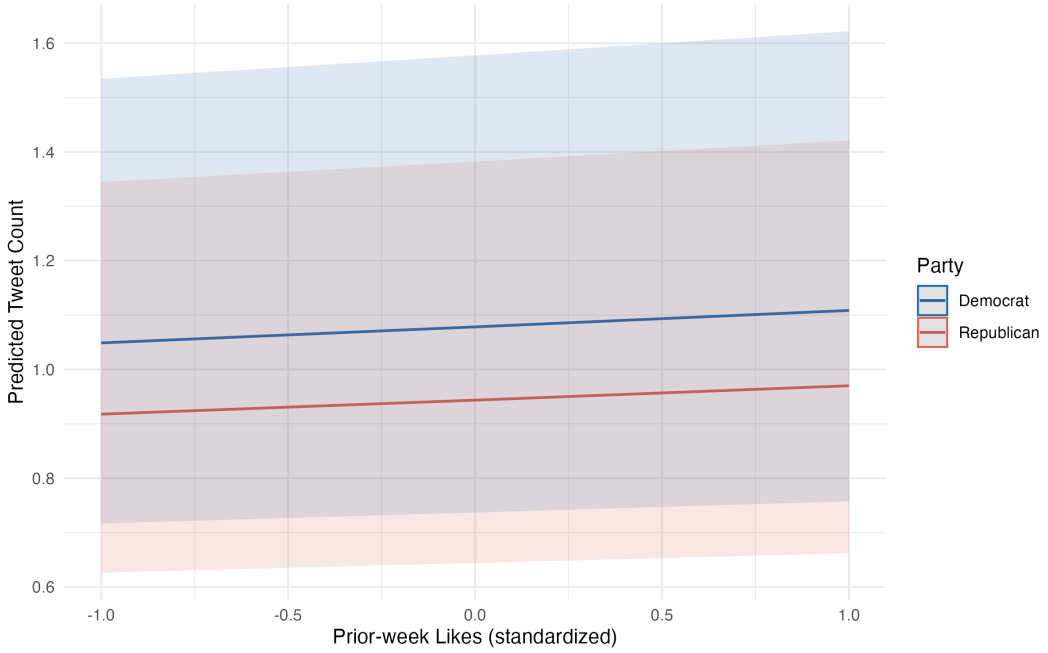

Figure 3: Predicted weekly election-integrity tweet count by party and prior-week likes (standardized), with 95% confidence intervals. Marginal predictions from the hierarchical negative binomial model.

None of the state-level structural predictors reach significance in the hierarchical count model: democratic performance (-0.013, IRR $= 0.987$, $p = 0.529$), legislative professionalism (-0.011, $p = 0.681$), and fraud cases (0.016, $p = 0.494$) are all small and statistically indistinguishable from zero. Vote share is also not significant (-0.001, $p = 0.917$), and House membership is marginal (-0.044, $p = 0.059$). Once legislator-level clustering is absorbed, stable individual differences explain most of the variation in tweet volume; state-level context appears to operate through the selection of legislators into posting rather than as a contemporaneous driver.

4.2.3 H2: Delegitimizing content

Table 4 presents the hierarchical linear probability model predicting whether a legislator ever posted a challenge tweet. The LPM column reports probability-point coefficients; the adjacent Firth OR column reports the same specification estimated via a separation-robust Firth penalized logistic regression, for readers who prefer an odds-ratio-scale interpretation. The Republican indicator is by far the dominant predictor: on the probability scale, being Republican is associated with a 15.6-percentage-point higher probability of having posted at least one challenge tweet (0.156, $p < .001$). On the Firth logit scale, Republicans are roughly 1.79 times as likely as otherwise-similar Democrats to have posted at least one challenge tweet. These are two views of the same finding: the LPM is a linear approximation to the conditional probability; the Firth logit is a penalized multiplicative-odds model on the same data.

Engagement variables remain in the challenge model. In the LPM, the coefficient on mean-per-active-week challenge-tweet likes is 0.003 ($p = 0.712$) and the corresponding challenge-tweet retweets coefficient is 0.057 ($p < .001$). (A Firth logit OR for the engagement variables is not available because challenge-tweet engagement is structurally collinear with the outcome at the logit scale—legislators who never posted challenge tweets have zero challenge engagement by definition—and the Firth sensitivity falls back to a no-engagement specification for a stable party OR.) These LPM coefficients are interpreted cautiously. They describe a cross-sectional association—legislators whose challenge tweets received more engagement are more likely to have ever posted challenge content—but they cannot separate feedback from reverse causation. Our cleaner test of engagement feedback is the hierarchical count model in Table 3, where the lag structure is weekly, within-legislator, and measured on all election tweets.

Beyond party and engagement, the only additional predictor that reaches significance at conventional thresholds is the White control: 0.048 on the probability scale (OR ≈ 1.20, $p = 0.027$). We flag this as real but do not emphasize it. The race control is a coarse binary

Table 4: Hierarchical Linear Probability Model of Whether a Legislator Posted Any Challenge Tweet. Random intercept for state. $N = 2{,}652$ legislators. Coefficients are percentage-point changes in the probability of ever challenging; the second column reports the OR from a separation-robust Firth penalized logistic regression on the same specification.

	LPM (probability)	Firth OR
Variable	Coefficient	OR
Challenge likes (mean/wk)	0.003 (0.009)	—
Challenge retweets (mean/wk)	0.057^{***} (0.009)	—
Republican	0.156^{***} (0.020)	1.79^{***}
House Member	0.018 (0.019)	1.05
White	0.048^{*} (0.022)	$1.20^{\dagger}$
Female	−0.004 (0.019)	1.03
Vote Share (scaled)	0.007 (0.010)	0.96
Democratic Performance (scaled)	−0.025 (0.016)	0.89^{**}
Leg. Professionalism (scaled)	−0.003 (0.018)	0.99
Fraud Cases (scaled)	−0.001 (0.017)	1.01
State (σ)	0.083	

LPM column: coefficients are from a hierarchical linear probability model fit via `lmer` with a state random intercept and a continuous outcome (probability of ever posting a challenge tweet). Standard errors in parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. All continuous predictors are standardized ($M = 0, SD = 1$). Engagement enters as the legislator's mean-per-active-week lagged likes and retweets. Firth OR column: odds ratio from a non-hierarchical Firth penalized logistic regression (Firth, 1993) on the same predictor set. Firth's method is explicitly separation-robust, allowing engagement variables to remain in the specification without inflating standard errors on other coefficients. State-level ICC (from LPM) = 0.034.

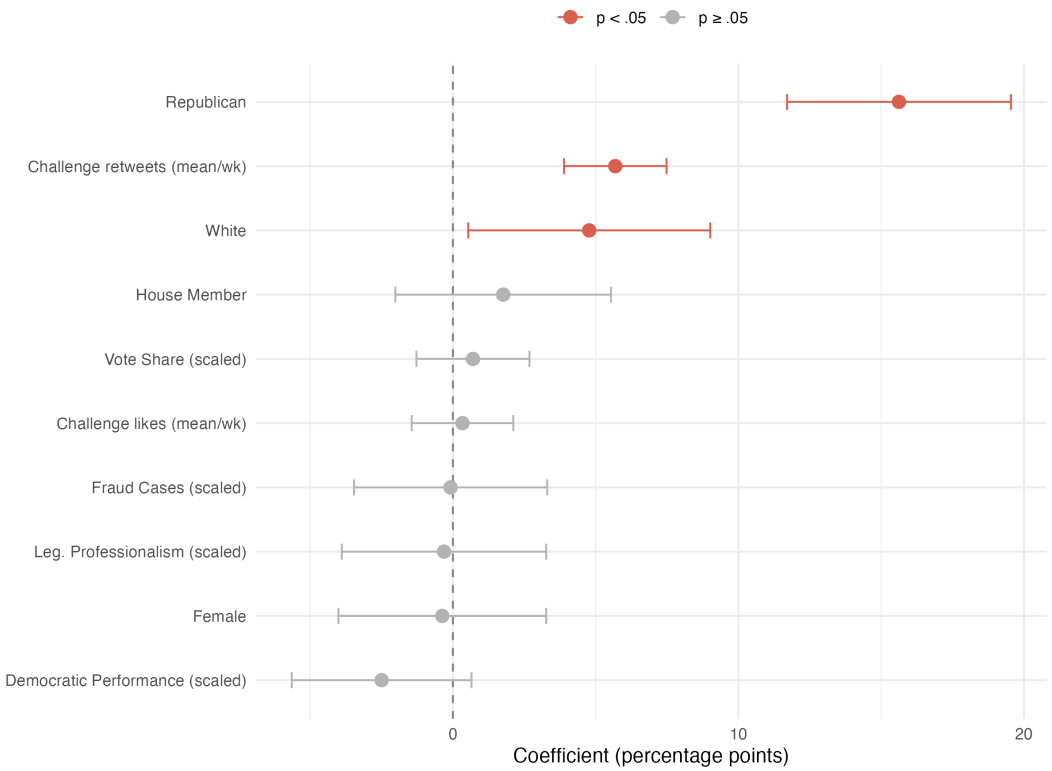

Figure 4: Coefficient plot for the hierarchical LPM challenge model. Coefficients are percentage-point changes in the probability of ever posting a challenge tweet. Points show point estimates with 95% confidence intervals. All continuous predictors standardized.

(white vs. all non-white categories pooled), and almost all Republican challenge-posters are white to begin with, so the coefficient is sensitive to the thin cells it rests on. A finer racial breakdown would be needed before reading a substantive story into this coefficient; we include it to be transparent about what the model says rather than as an interpretive contribution.

No other predictor in the challenge model is significant. State democratic performance (-0.025, $p = 0.117$), vote share (0.007, $p = 0.487$), legislative professionalism (-0.003, $p = 0.860$), and fraud cases (-0.001, $p = 0.961$) are all small and non-significant. Figure 5 visualizes the predicted probability of a challenge tweet by vote share separately by party; the partisan gap is the dominant feature and the vote-share gradient within each party is essentially flat.

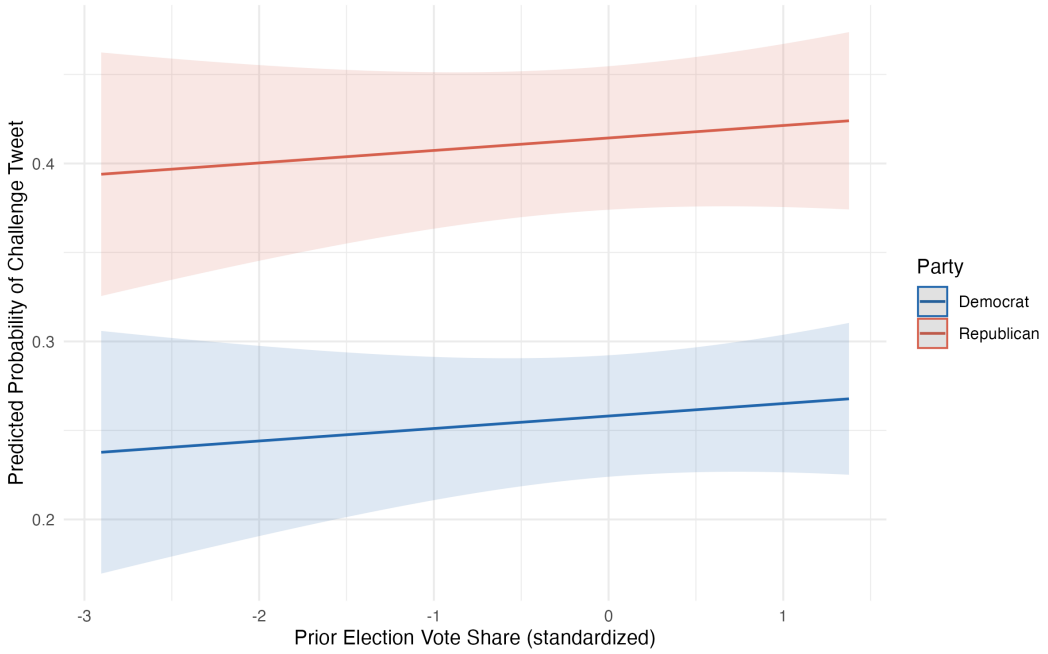

Figure 5: Predicted probability of posting a challenge tweet by prior election vote share (standardized) and party, with 95% confidence intervals. Marginal predictions from the hierarchical linear probability model.

4.3 Initiation and persistence: a joint two-stage hurdle model

The legislator-level LPM in Table 4 treats delegitimization as a single binary outcome—ever posted at least one challenge tweet vs. never—but behaviorally this bundles two distinct decisions: (i) the crossing of the threshold from never-challenged to ever-challenged (*initiation*), and (ii) the ongoing weekly rate of challenge posting among legislators who have crossed that threshold (*persistence*). These decisions have different predictors. Initiation cannot be predicted by engagement on challenge tweets, because a legislator who has never posted a challenge tweet has no challenge-tweet engagement by definition. Persistence, by contrast, is exactly where the challenge-specific engagement signal should matter: does attention on a legislator's own past challenge tweets sustain further challenge posting?

To estimate these two stages together we fit a hurdle negative binomial model (Mullahy, 1986) at the legislator-week level, using `glmmTMB` (Brooks et al., 2017) with `family = truncated_nbinom2` and a separate `ziformula`. A hurdle model decomposes the weekly challenge count into a zero process and a positive-count process, and `glmmTMB` estimates both parts jointly in a single optimization under one likelihood. Each part has its own fixed-effect coefficients and its own state random intercept; the conditional part additionally includes a legislator random intercept. Stage 1 (the zero hurdle) uses engagement on all election tweets as its engagement predictor, because challenge engagement is not defined before initiation. Stage 2 (the truncated NB) uses lagged challenge-tweet engagement specifically, which is where the delegitimization-feedback mechanism is identifiable.

Table 5 reports both stages, with the Stage 1 coefficients sign-flipped so that positive values correspond to higher probability of posting. The partisan gap appears in both stages: in Stage 1, being Republican is associated with a 1.706 higher log-odds of posting any challenge tweet in a given week ($p < .001$); in Stage 2, conditional on posting, Republicans post at a rate 4.375 times that of otherwise-similar Democrats (1.476, $p < .001$). The partisan gap is therefore present at both margins, not just in who initiates.

Engagement feedback operates on different signals in the two stages. In Stage 1

Table 5: Joint Two-Stage Hurdle Model of Weekly Challenge-Tweet Posting. Fit by glmmTMB with `family = truncated_nbinom2` and a `ziformula`. $N = 33{,}288$ legislator-weeks; $N_{>0} = 2{,}631$ weeks with at least one challenge tweet.

	Stage 1: Initiation (logit, $P(\text{post} > 0)$)	Stage 2: Persistence (trunc. NB, log count $\mid > 0$)
Variable	Coefficient	Coefficient
Republican	1.706^{***} (0.048)	1.476^{***} (0.128)
Likes $t-1$ (all election)	0.069^{***} (0.018)	—
Retweets $t-1$ (all election)	0.341^{***} (0.022)	—
Likes $t-1$ (challenge tweets)	—	0.128^{**} (0.041)
Retweets $t-1$ (challenge tweets)	—	0.215^{***} (0.044)
House Member	0.141^{**} (0.049)	-0.284^{*} (0.123)
White	$-0.097^{\dagger}$ (0.058)	-0.165 (0.157)
Female	0.062 (0.047)	0.103 (0.115)
Vote Share (scaled)	0.054^{*} (0.026)	-0.095 (0.067)
Democratic Performance (scaled)	0.013 (0.072)	0.154 (0.136)
Leg. Professionalism (scaled)	-0.022 (0.093)	-0.076 (0.181)
Fraud Cases (scaled)	-0.063 (0.086)	-0.149 (0.167)
N leg.-weeks	33,288	
N leg.-weeks with challenge > 0	2,631	

Joint hurdle negative binomial fit via `glmmTMB` with a single optimization that estimates both stages together. Stage 1 (ziformula, sign flipped) gives the log-odds of posting *at least one* challenge tweet in week t and uses engagement on all election tweets (challenge-tweet engagement is undefined for legislators who have not yet initiated). Stage 2 (truncated NB) gives the log expected count of challenge tweets *given* that at least one challenge was posted that week, and uses challenge-tweet engagement from the prior week. Both stages include week fixed effects (Stage 2) or otherwise control for time; both include state random intercepts; Stage 2 additionally includes legislator random intercepts. Standard errors in parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

(initiation at a weekly cadence), all-election lagged likes are associated with posting (0.069, $p < .001$) and so are lagged retweets (0.341, $p < .001$): more attention to a legislator's broader election-related tweeting last week predicts a higher probability of posting a challenge tweet this week. In Stage 2 (persistence given posting), the relevant signal is engagement on the challenge tweets themselves: 0.128 ($p = 0.002$) for lagged challenge likes and 0.215 ($p < .001$) for lagged challenge retweets, corresponding to IRRs of 1.137 and 1.240 on the rate of challenge posting. The separation between stages illustrates why a single-equation logistic regression that tries to put challenge engagement on the right-hand side of an ever-challenged outcome runs into trouble: Stage 1 and Stage 2 are fundamentally different decisions with different informative signals, and the joint hurdle is the appropriate specification.

4.4 Within-party tweet volume

The large between-party gap naturally invites the question of what drives within-party variation. For the *volume* outcome we estimate within-party negative binomial count models with cluster-robust standard errors by state (Table 6). We do not split the *challenge* outcome by party. Challenge-tweet engagement is structurally sparse—most legislators never post challenge tweets, so their challenge engagement is zero by definition—and splitting by party creates near-perfect collinearity between challenge engagement and the outcome within each subsample. The main hierarchical LPM (Table 4), which includes party as a predictor, is the appropriate specification for the challenge outcome.

Democrats' tweet volume tracks state context to a modest degree: state democratic performance is negatively associated with Democratic tweet count (-0.170, IRR $= 0.844$, $p = 0.021$), meaning Democrats in states with stronger expert-assessed democratic performance post somewhat fewer election-integrity tweets, and documented fraud cases are positively associated with volume (0.117, IRR $= 1.124$, $p = 0.083$). Among Republicans, the analogous structural predictors produce smaller and less precise estimates (vote share 0.021, $p = 0.838$; state democratic performance 0.012, $p = 0.865$), consistent with a picture in which Republican

Table 6: Within-Party Tweet-Volume Models (Negative Binomial). Cluster-robust standard errors by state. Dem. $N = 2{,}025$; Rep. $N = 1{,}212$.

	Democrats	Republicans
Variable	Coefficient	Coefficient
Likes (mean/wk)	1.272^{***} (0.047)	0.112^{***} (0.017)
Retweets (mean/wk)	1.008^{***} (0.032)	0.598^{***} (0.035)
House Member	-0.250^{*} (0.101)	-0.057 (0.154)
White	0.094 (0.079)	0.098 (0.156)
Female	-0.089 (0.072)	0.221 (0.163)
Vote Share (scaled)	-0.040 (0.050)	0.021 (0.104)
Democratic Performance (scaled)	-0.170^{*} (0.074)	0.012 (0.072)
Leg. Professionalism (scaled)	-0.006 (0.071)	0.072 (0.150)
Fraud Cases (scaled)	$0.117^{\dagger}$ (0.068)	0.044 (0.064)
Constant	3.706^{***} (0.114)	2.743^{***} (0.243)
N legislators	1,745	907
θ (NB disp.)	0.79	0.83

Cluster-robust standard errors by state in parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. Continuous predictors standardized ($M = 0, SD = 1$). Engagement variables are the legislator's mean lagged likes / retweets per active week (reviewer comment II.2: prior draft summed engagement, which confounded exposure duration with intensity). Dependent variable is the legislator's total election-integrity tweet count over the study period.

tweet volume does not track the same state-level indicators.

Structural predictors and the challenge null. The structural-predictor null for challenge posting comes from the main hierarchical LPM (Table 4), not from a within-party split. In that model, vote share (0.007, $p = 0.487$), state democratic performance (-0.025, $p = 0.117$), legislative professionalism (-0.003, $p = 0.860$), and fraud cases (-0.001, $p = 0.961$) are all small and non-significant, with 95% confidence intervals on the probability-point scale of roughly -0.013 to 0.027 (vote share), -0.056 to 0.006 (democratic performance), -0.039 to 0.033 (legislative professionalism), and -0.035 to 0.033 (fraud cases). All four intervals straddle zero and are narrow enough to rule out large structural effects on challenge posting. Our companion power simulation (Section 4.6) calibrated to a full-sample Firth logistic confirms that the study has 80% power to detect effects of OR = 1.15 or larger at the observed base rate of challenge posting (27.1%).

4.5 Robustness

Table 7 reports the hierarchical count model across *eight* specifications: the main model (column 1), with winsorized engagement variables at the 99th percentile (column 2; caps at 2,298 likes and 52,111 retweets), with a two-week lag on engagement (column 3), excluding the top 1% most prolific tweeters (column 4; 27 legislators dropped), with state fixed effects instead of state random effects (column 5; state-level predictors absorbed), using original (non-retweet) tweets only (column 6), with event-window time structure that replaces 104 week fixed effects with a linear time trend plus dummies for the 2020 election window (± 2 weeks around November 3) and the January 6, 2021 window (± 1 week) (column 7), and with the lagged engagement variables entered on a $\log(1 + x)$ scale rather than the linear scale used in the main specification (column 8).

The Republican indicator remains negative and statistically significant in every specification (winsorized: -0.135, $p < .001$; two-week lag: -0.107, $p < .001$; trimmed: -0.120,

Table 7: Robustness Checks: Hierarchical Negative Binomial Count Model Across Eight Specifications.

	(1) Main	(2) Wins.	(3) Lag 2	(4) Trim	(5) State FE	(6) Orig. only	(7) Event wind.	(8) Log eng.
Likes $t-1$	0.028^{***}	0.069^{***}	0.017^{***}	0.019^{***}	0.028^{***}	-0.001	0.028^{***}	0.106^{***}
	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.013)	(0.004)	(0.005)
Retweets $t-1$	0.081^{***}	0.114^{***}	0.059^{***}	0.067^{***}	0.081^{***}	0.037^{**}	0.074^{***}	0.144^{***}
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.013)	(0.005)	(0.006)
Republican	-0.133^{***}	-0.135^{***}	-0.107^{***}	-0.120^{***}	-0.133^{***}	-0.089^{***}	-0.141^{***}	-0.092^{***}
	(0.025)	(0.024)	(0.027)	(0.022)	(0.025)	(0.024)	(0.024)	(0.023)
House	$-0.044^{\dagger}$	$-0.043^{\dagger}$	-0.041	-0.033	$-0.043^{\dagger}$	-0.024	$-0.041^{\dagger}$	-0.029
	(0.023)	(0.022)	(0.025)	(0.021)	(0.023)	(0.022)	(0.023)	(0.022)
White	0.023	0.021	0.014	0.019	0.020	0.025	0.027	0.018
	(0.026)	(0.025)	(0.028)	(0.023)	(0.026)	(0.025)	(0.025)	(0.024)
Female	0.008	0.009	0.016	-0.019	0.006	-0.008	0.009	0.013
	(0.022)	(0.021)	(0.024)	(0.020)	(0.022)	(0.021)	(0.022)	(0.021)
Vote share	-0.001	-0.001	-0.001	-0.005	0.002	-0.005	-0.000	0.002
	(0.012)	(0.012)	(0.013)	(0.011)	(0.013)	(0.011)	(0.012)	(0.011)
Dem. performance	-0.013	-0.013	-0.012	-0.019	—	-0.020	-0.011	-0.009
	(0.021)	(0.021)	(0.022)	(0.018)		(0.017)	(0.021)	(0.020)
Leg. prof.	-0.011	-0.010	-0.016	-0.004	—	0.002	-0.008	-0.007
	(0.026)	(0.025)	(0.027)	(0.021)		(0.021)	(0.025)	(0.024)
Fraud cases	0.016	0.016	0.008	0.019	—	0.013	0.017	0.009
	(0.024)	(0.023)	(0.025)	(0.020)		(0.019)	(0.023)	(0.022)
N leg.-weeks	33,288	33,288	30,636	31,306	33,288	22,706	33,288	33,288

All models are hierarchical negative binomial with random intercepts for state and legislator. Columns (1)-(6) and (8) include 104 week fixed effects; column (5) additionally uses state fixed effects in place of the state random intercept. Column (7) replaces week FE with a linear time trend plus event-window dummies for the 2020 election (± 2 weeks around Nov. 3) and January 6, 2021 (± 1 week). All continuous predictors are scaled to mean 0, SD 1. (1) Main. (2) Engagement winsorized at 99th pctile. (3) Two-week lag. (4) Top 1% of prolific legislators excluded. (5) State fixed effects. (6) Original (non-retweet) tweets only. (7) Event-window time structure (II.7). (8) Engagement entered as $\log(1+x)$ instead of raw / 1000 (IV.6). Standard errors in parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

$p < .001$; state FE: -0.133, $p < .001$; originals only: -0.089, $p < .001$; event-window: -0.141, $p < .001$; log-engagement: -0.092, $p < .001$). The lagged-engagement coefficients remain positive and significant across all specifications including the log-transformed version (log likes: 0.106, $p < .001$; log retweets: 0.144, $p < .001$), ruling out the concern that the linear-scale engagement specification is masking or inventing the feedback effect through a small number of viral outliers. The event-window specification (column 7) is particularly informative: the 2020 election window has a large and significant positive coefficient (0.700, $p < .001$), and the January 6 window has a modest positive coefficient (0.740, $p < .001$), but the engagement feedback coefficients persist outside those windows (0.028, $p < .001$; 0.074, $p < .001$), showing that the feedback dynamic is not purely an artifact of the two dominant event spikes. Non-significant structural predictors remain non-significant across all eight specifications. An explicit Poisson vs. NB comparison in Appendix A confirms that negative binomial is strongly preferred on AIC/BIC, and the *hierarchical* Pearson overdispersion (computed from the fitted NB with random effects in the model, not the pooled Poisson) is $\hat{\phi} = 0.947$, indicating that meaningful residual dispersion remains even after the state and legislator random intercepts absorb between-cluster variation.

4.6 Statistical power

Simulation-based power analyses calibrated to each of the three main modeling approaches used in the paper (hierarchical NB count model for H1 and H3; hierarchical LPM and Firth cross-check for H2; second stage of the two-stage hurdle model) are reported in Appendix B. Headline results: the minimum detectable main effect in the count model is 0.03 on the log-count scale, in the Firth challenge logit OR = 1.15, in the challenge LPM 0.03 percentage points, and in the hurdle Stage 2 conditional-count model $\beta = 0.075$. All observed main effects corresponding to the paper's hypothesized estimands clear these floors.

5 Discussion

Our analysis of 90,355 tweets from 3,235 state legislators yields three main findings.

First, election-integrity discourse *volume* is predominantly a Democratic phenomenon at the state legislative level, and it is responsive to engagement feedback from prior weeks. The combination of a negative Republican coefficient in the count model and positive, significant lagged engagement coefficients is consistent with a picture in which Democratic legislators produce the bulk of ongoing election-integrity discourse, with social media feedback amplifying the behavior of already-active posters. This is a different picture than one often implied by the recent public debate: it is not the case that election-integrity tweeting is dominated by Republicans raising fraud concerns. On the contrary, the volume is Democratic.

Second, the binary decision to *delegitimize* is overwhelmingly a partisan phenomenon. Being Republican is associated with a 15.6-percentage-point higher probability of ever posting challenge content (Firth logit OR = 1.79). This gap holds at the legislator level after adjusting for chamber, race, gender, vote share, state democratic performance, legislative professionalism, fraud cases, engagement, and state clustering. The scale of the gap—spanning all fifty states and more than a thousand Republican legislators—shows how deeply delegitimizing rhetoric has penetrated the Republican state legislative caucus, at a level of elected official that retains direct authority over election administration. This extends LaPlant, Lee and LaPlant (2024)'s findings about federal House Republicans to the state legislative context.

Third, the structural and electoral predictors we measured do not clearly predict challenge posting. In the main LPM (Table 4), vote share (0.007, $p = 0.487$), state democratic performance (-0.025, $p = 0.117$), legislative professionalism (-0.003, $p = 0.860$), and fraud cases (-0.001, $p = 0.961$) all produce small, non-significant coefficients on the probability-point scale. The 95% confidence intervals are narrow: vote share -0.013 to 0.027, democratic performance -0.056 to 0.006, legislative professionalism -0.039 to 0.033, fraud cases -0.035 to 0.033. These intervals rule out large structural effects in either direction. Our companion power analysis, calibrated to a full-sample Firth logistic at the observed base rate, confirms

80% power to detect effects of OR = 1.15 and above.

What the data *cannot* rule out are small structural effects in the detection-threshold band, and more importantly, mechanisms we did not measure. Richer measures—especially individual-level ideology scores, primary-challenger exposure, connections to the national conservative media ecosystem, and explicit endorsements of "Stop the Steal" figures—would be needed to pin down within-GOP variation further. Our null is informative (it narrows the space of plausible mechanisms) without being exhaustive (it does not name the mechanism that would take the measured predictors' place). Future work with data on elite conservative networks and primary-challenger dynamics would be well positioned to make progress here.

5.1 Implications and limitations

When elected officials with institutional authority challenge the legitimacy of elections, they contribute to an information environment of democratic distrust. Our findings suggest that, at the state legislative level in 2020–2021, this supply is primarily a partisan phenomenon rather than a response to local election conditions. Reform efforts narrowly focused on improving election administration—though independently valuable—are unlikely to reduce the supply of delegitimizing elite rhetoric, because within-Republican variation in that supply does not appear to track administrative conditions.

Several limitations warrant acknowledgment. First, we study a single platform (Twitter/X) during an exceptional political period; the patterns we document may not generalize to other platforms or other cycles. Second, our classifier is imperfect, though the held-out accuracy and inter-coder reliability metrics are strong. Third, the within-Republican null is conditional on the specific structural predictors we measure, and richer mechanisms we did not measure may be important. Fourth, we make no causal claim: the engagement-feedback pattern is consistent with a reinforcement interpretation, but we cannot rule out common-cause stories (e.g., a week-specific event drives both engagement and subsequent posting). Fifth, our study speaks to the supply of delegitimizing rhetoric, not its downstream effects

on constituent attitudes or behavior; linking elite rhetoric to constituent outcomes is an important next step.

6 Conclusion

Using a dataset of 90,355 election-related tweets from 3,235 state legislators over 2020–2021, we document a clear divergence between the two outcomes of interest. The *volume* of election-integrity discourse on state-legislative social media is predominantly Democratic and responds to engagement feedback. The binary *decision to delegitimize*—to post even one tweet challenging electoral legitimacy—is overwhelmingly partisan: being Republican is associated with a 15.6-percentage-point higher probability of ever posting challenge content (Firth logit OR = 1.79), after conditioning on chamber, race, gender, structural predictors, engagement, and state clustering. None of the measured structural or electoral predictors significantly predict challenge posting, an honest null that narrows but does not resolve the question of mechanism. Our findings document how delegitimizing rhetoric has diffused to state legislative actors who retain direct authority over election administration, and they clarify what is and is not a plausible driver of that diffusion.

References

Bates, Douglas, Martin Mächler, Benjamin Bolker and Steven Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67(1):1–48.

Berlinski, Nicolas, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan and Jason Reifler. 2021. "The Effects of Unverified Claims of Voter Fraud on Confidence in Elections." *Journal of Experimental Political Science* 8(2):129–143.

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker and Jay J. Van Bavel.

2017. "Emotion Shapes the Diffusion of Moralized Content in Social Networks." *Proceedings of the National Academy of Sciences* 114(28):7313–7318.

Brooks, Mollie E., Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Mächler and Benjamin M. Bolker. 2017. "glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Models." *The R Journal* 9(2):378–400.

Cameron, A. Colin, Jonah B. Gelbach and Douglas L. Miller. 2011. "Robust Inference with Multiway Clustering." *Journal of Business & Economic Statistics* 29(2):238–249.

Cameron, A. Colin and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data*. 2 ed. Cambridge, UK: Cambridge University Press.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20(1):37–46.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. pp. 4171–4186.

Donno, Daniela, Nasos Roussias and Adam Scharpf. 2022. "Electoral Malpractice and Post-Election Protest." *Journal of Politics* 84(1):1–15.

Edelson, Jack, Alexander Alduncin, Christopher Krewson, James A. Sieja and Joseph E. Uscinski. 2021. "The Effect of Conspiratorial Thinking and Motivated Reasoning on Belief in Election Fraud." *Political Research Quarterly* 74(4):902–917.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1):27–38.

Heinze, Georg and Michael Schemper. 2002. "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21(16):2409–2419.

Kawintiranon, Kornraphop and Lisa Singh. 2022. PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*. pp. 7360–7367.

LaPlant, Kristina M., Keith E. Lee and James T. LaPlant. 2024. "Stopping the Steal and Selling the Big Lie: An Analysis of Tweets and Certification Votes among House Republicans in the Wake of the 2020 Presidential Election." *American Politics Research* 52(2):141–156.

Mullahy, John. 1986. "Specification and Testing of Some Modified Count Data Models." *Journal of Econometrics* 33(3):341–365.

Nakagawa, Shinichi, Paul C. D. Johnson and Holger Schielzeth. 2017. "The Coefficient of Determination R^2 and Intra-class Correlation Coefficient from Generalized Linear Mixed-Effects Models Revisited and Expanded." *Journal of the Royal Society Interface* 14(134):20170213.

Norris, Pippa. 2019. *Electoral Integrity in America: Securing Democracy*. New York: Oxford University Press.

Przeworski, Adam. 2018. *Why Bother with Elections?* Cambridge, UK: Polity Press.

Schnaudt, Christian. 2025. "Electoral Legitimacy and Compliance with Public Regulations." *Political Behavior* .

Settles, Burr. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648 University of Wisconsin–Madison.

Squire, Peverill. 2017. "A Squire Index Update." *State Politics & Policy Quarterly* 17(4):361–371.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pp. 38–45.

A Poisson vs. Negative Binomial

Our count outcomes are heavily overdispersed at the legislator-week level. The raw variance-to-mean ratio on the active-week panel is 7.74; this estimate is conservative for the true dispersion because the legislator-week panel includes only weeks in which the legislator tweeted at least once about elections, so zero-count weeks are not represented in the denominator. To get a dispersion estimate that accounts for the hierarchical structure, we also compute the Pearson overdispersion from the fitted hierarchical NB model, conditional on both the state and legislator random intercepts: $\hat{\phi} = 0.947$. Both diagnostics indicate substantial residual dispersion even after the hierarchical variance components absorb between-cluster variation, supporting the NB over Poisson.

Table 8 presents a side-by-side comparison of pooled Poisson and pooled negative binomial specifications on the same legislator-week data, without random effects, to keep the family comparison clean. The NB is overwhelmingly preferred on both AIC and BIC, and the Pearson overdispersion ratio of the pooled Poisson model is $\hat{\phi} = 5.92$ (by construction larger than the hierarchical counterpart because the pooled specification cannot use the random effects to absorb between-cluster variation). Coefficient signs and substantive interpretations are consistent across the two specifications, and the main results in the paper come from the hierarchical negative binomial model rather than either pooled fit shown here.

Table 8: Pooled Poisson vs. Pooled Negative Binomial Coefficient Comparison. Used to motivate the negative binomial family in the main hierarchical model.

	Poisson	Negative Binomial
Likes $t-1$ (scaled)	0.032^{***} (0.001)	0.089^{***} (0.004)
Retweets $t-1$ (scaled)	0.096^{***} (0.001)	0.331^{***} (0.004)
Republican	-0.103^{***} (0.009)	-0.148^{***} (0.013)
House Member	-0.083^{***} (0.007)	-0.051^{***} (0.011)
White	0.089^{***} (0.008)	0.059^{***} (0.013)
Female	0.135^{***} (0.007)	0.097^{***} (0.011)
Vote Share (scaled)	-0.011^{**} (0.004)	-0.006 (0.006)
Democratic Performance (scaled)	0.049^{***} (0.004)	0.024^{***} (0.006)
Leg. Professionalism (scaled)	-0.032^{***} (0.004)	-0.012^{*} (0.006)
Fraud Cases (scaled)	-0.031^{***} (0.004)	-0.014^{**} (0.005)
AIC	173,146	133,961
BIC	173,239	134,062
Poisson overdispersion $\hat{\phi}$	5.92	

Pooled (non-hierarchical) fits on the same legislator-week data as the main hierarchical model. This table motivates the choice of family; the main model of the paper is hierarchical negative binomial. Standard errors in parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. Poisson overdispersion ratio $\hat{\phi} = \sum r_p^2/(n-k)$, where r_p are Pearson residuals.

B Statistical Power

We report simulation-based power analyses for the main-effect estimands that the theory actually calls for, spanning each of the three modeling approaches used in the paper: the hierarchical NB count model (H1 and H3), the hierarchical LPM and Firth cross-check for challenge posting (H2), and the second stage of the two-stage hurdle model (conditional persistence in challenge posting). No interaction between party and engagement is theorized in Section 2, and so none is powered. The calibrations differ because the estimands sit on different outcome scales, use different sample sizes, and call for different power-assessment machinery; each sweep is centered around the coefficients actually observed in the corresponding fit.

B.1 Power for main effects in the hierarchical NB count model (H1 and H3)

The first simulation calibrates the data-generating process to the fitted hierarchical NB model ($\theta = 4.72$, $\sigma_{\text{state}} = 0.130$, $\sigma_{\text{legislator}} = 0.387$) and the observed sample structure (3,237 legislators, 50 states, 104 weeks; 33,288 legislator-weeks after filtering). We sweep the true value of the lagged-likes main-effect coefficient from 0 to 0.20 on the log-count scale in increments of 0.01, fixing all other coefficients at their fitted values, and run 200 simulations per effect size. The range is centered on the observed main effects ($|b| = 0.028$ for likes, $|b| = 0.081$ for retweets, $|b| = 0.133$ for the Republican indicator), which fall in the middle third of the sweep. Estimation in the simulations uses pooled NB (no random effects) as a conservative lower bound; the full hierarchical model is more efficient and therefore at least as well powered. We target the lagged-engagement slope rather than the Republican main effect because H3 is the tightest estimand: the observed engagement coefficients are much smaller than the Republican main effect, so a power floor that is adequate for H3 is *a fortiori* adequate for H1.

Figure 6 presents the power curve. The minimum detectable effect size (MDES) at 80% power is 0.03 on the log-count scale. The observed Republican main effect is well above this threshold; the observed H3 coefficients sit at or above the MDES boundary, consistent with the moderate but reliable engagement-feedback effects reported in Table 3.

B.2 Power for main effects in the Firth challenge logit (H2, OR scale)

The second simulation assesses power for a main-effect coefficient in the challenge Firth logit, which is the nonlinear cross-check of the LPM estimate of H2. We calibrate to the full legislator sample ($N = 3{,}237$, overall base rate of any challenge posting = 27.1%) and sweep true odds ratios from 1.05 to 2.00 for a single main-effect predictor. We ran 200 simulations

Table 9: Statistical Power Curve for a Main-Effect Coefficient in the Hierarchical NB Count Model. 200 simulations per effect size, calibrated to the fitted hierarchical NB model (state and legislator random effects, NB dispersion $\theta = 4.72$).

Main-Effect Size	Power	Valid Simulations
0.00	0.060	200
0.01	0.220	200
0.02	0.515	200
0.03	**0.890**	**200**
0.04	0.995	200
0.05	1.000	200
0.06	1.000	200
0.07	1.000	200
0.08	1.000	200
0.09	1.000	200
0.10	1.000	200
0.11	1.000	200
0.12	1.000	200
0.13	1.000	200
0.14	1.000	200
0.15	1.000	200
0.16	1.000	200
0.17	1.000	200
0.18	1.000	200
0.19	1.000	200
0.20	1.000	200

Bold row marks the minimum detectable effect size (MDES) at 80% power (0.03 on the log-count scale). The swept coefficient is the lagged-likes main effect (H3); the data-generating process fixes all other coefficients at their fitted values. Estimation uses pooled NB (conservative; no random effects in the fit) to avoid convergence failures during simulation. Because the fitted model has random effects, actual study power is equal to or greater than the reported values.

per effect size; for each simulated dataset, a standardized predictor is drawn $\mathrm{N}(0, 1)$, the outcome is generated from a logit model with the calibrated intercept and the target OR, and a Firth penalized logit is fit. The minimum detectable OR at 80% power is 1.15. The observed Republican main effect on challenge posting is orders of magnitude above this floor, so H2 is far from a power-limited test.

B.3 Power for main effects in the challenge LPM (H2, probability-point scale)

The hierarchical LPM is the primary model used to test H2 in the paper (Table 4). The Firth simulation above establishes the detection floor on the odds-ratio scale; this third simulation establishes the floor on the probability-point scale that the LPM coefficients are expressed in. We simulate a binary challenge outcome at the legislator level as $y \sim \text{Bernoulli}(p_0 + \beta x)$ with $p_0 = 27.1\%$ and $x \sim \mathrm{N}(0, 1)$, sweep β from 0 to 0.20 in increments of 0.01, and fit $\text{lm}(y \sim x)$ to each simulated sample ($N = 3{,}237$; 200 simulations per effect size). The range is centered on the observed LPM main effects: the Republican indicator (0.156), lagged retweets (0.057), the white-legislator indicator (0.048), and the four non-significant structural predictors, whose coefficients all lie below ~ 0.03. The minimum detectable LPM coefficient at 80% power is 0.03 on the probability-point scale. The observed Republican main effect clears this floor comfortably; the structural-predictor nulls (vote share, democratic performance, legislative professionalism, fraud cases) are all below the floor, which is informative: an effect of 0.03 probability points per one-SD predictor is the smallest the design could have reliably detected.

B.4 Power for the second-stage conditional count in the hurdle model

The two-stage hurdle model (Section 4.3) has a conditional-count stage that governs how intensely a legislator challenges once she has crossed the initiation threshold. Stage 2 is

a truncated NB on challenge-tweet counts conditional on count > 0 and uses challenge-tweet engagement rather than all-election engagement. This simulation powers a main-effect coefficient in Stage 2 on the log-count scale. We calibrate to the observed sample of challenge-posting legislator-weeks ($N_{\text{cond}} = 2{,}631$, conditional mean count $\bar{\mu}_0 = 2.02$, NB dispersion $\theta = 1.78$), sweep β from 0 to 0.30 in increments of 0.015 (centered on the observed Stage 2 engagement coefficients of 0.128 for likes and 0.215 for retweets), and fit pooled NB on the positive subset of each simulated sample (200 simulations per effect size). Fitting on the positive subset without an explicit truncation correction mildly attenuates the estimated coefficient, so the reported power is a conservative lower bound on the actual Stage 2 power. The minimum detectable Stage 2 coefficient at 80% power is $\beta = 0.075$ on the log-count scale. Both observed Stage 2 engagement coefficients (0.128 for likes, 0.215 for retweets) are above this threshold, and the observed Republican Stage 2 coefficient (1.476) is far above, so every Stage 2 finding in Section 4.3 is well clear of the detection floor.

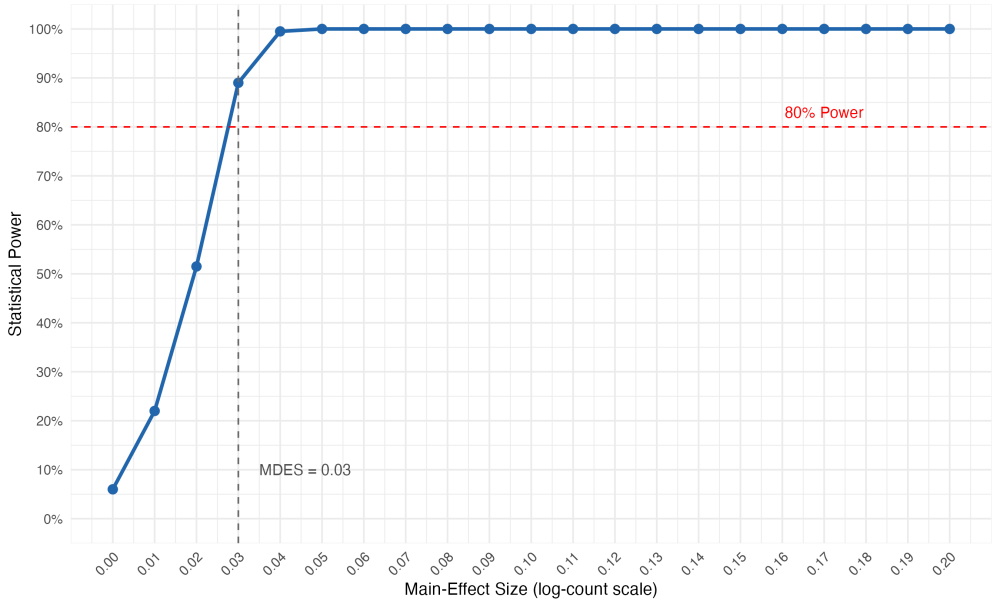

Figure 6: Statistical power curve for a main-effect coefficient in the hierarchical NB count model. Dashed horizontal line marks 80% power; the minimum detectable effect size (MDES) at 80% power is 0.03 on the log-count scale. Each point is based on 200 simulations.

C Exploratory Text Analysis

This appendix provides descriptive characterization of the linguistic features of election-related tweets from state legislators, using exploratory text analysis methods. These analyses informed our two-stage classification approach and provide additional descriptive context.

C.1 Word frequency

Figures 7 and 8 show the most frequent words in challenge and non-challenge election tweets, respectively, after stopword removal. Challenge tweets are characterized by terms such as "fraud," "steal," "rigged," and "illegal"—the lexical signature of the "big lie" narrative. Non-challenge tweets employ a broader vocabulary centered on "vote," "election," and process-oriented terms.

C.2 Collocation analysis

Figure 9 shows the top four-word collocations in challenge tweets, ranked by lambda collocation score. The most distinctive phrases reflect multi-word constructions referencing stolen elections, voter fraud, and Stop the Steal.

C.3 Monthly composition

Figure 10 shows the monthly share of challenge vs. non-challenge election-related tweets from December 2019 through December 2021. The figure illustrates the surge in challenge content beginning in November 2020 and peaking in December 2020, coinciding with the post-election certification dispute.

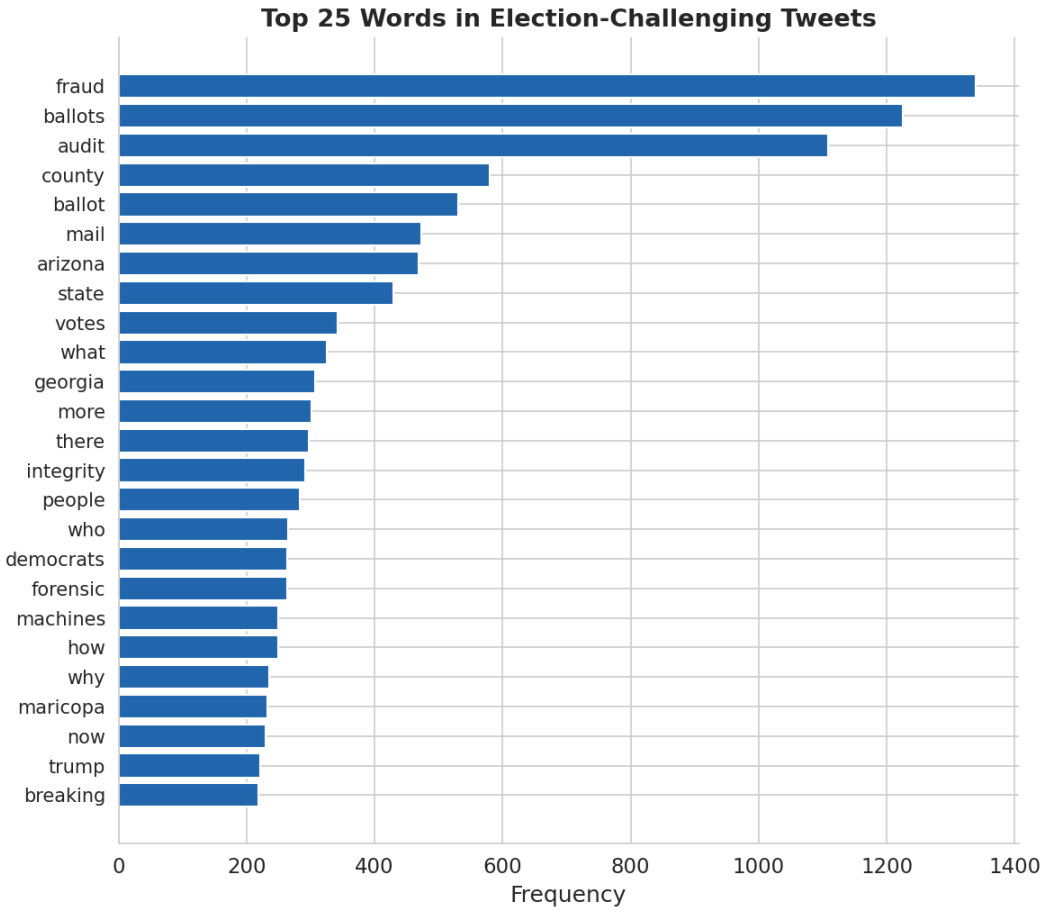

Figure 7: Most frequent words in challenge tweets (after stopword removal).

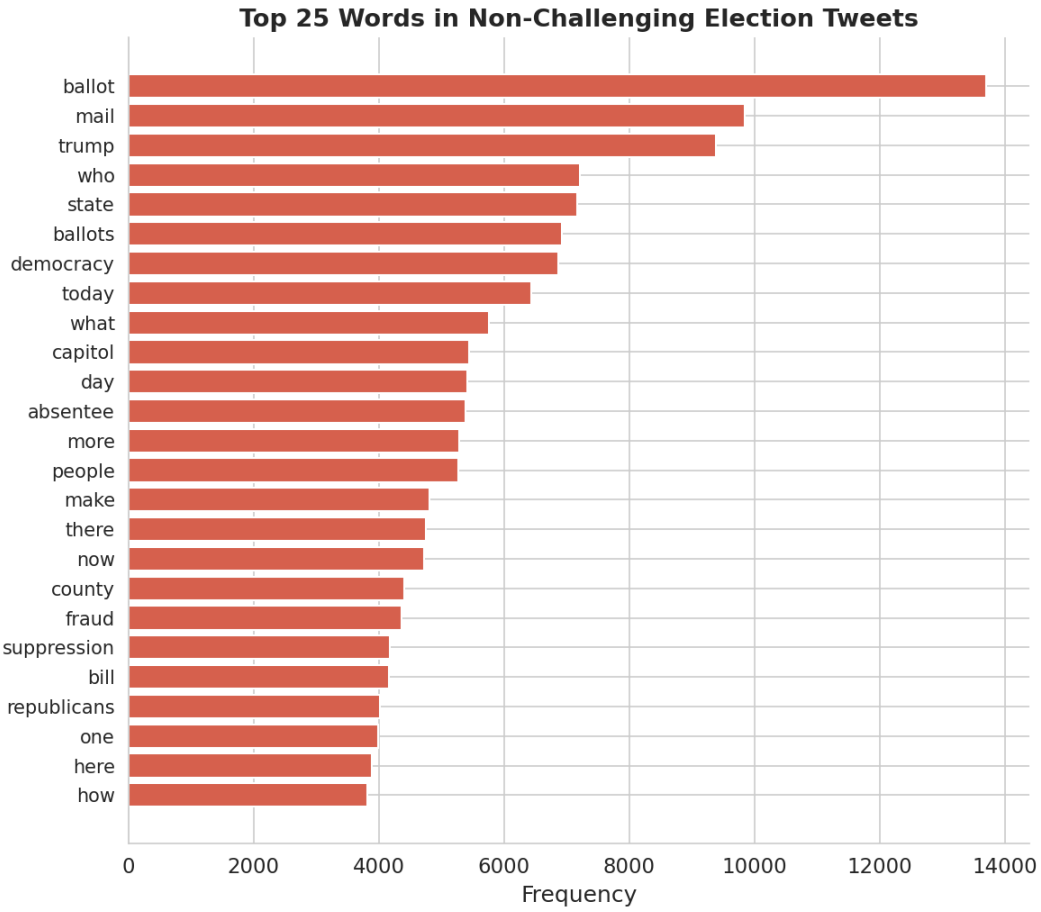

Figure 8: Most frequent words in non-challenge tweets (after stopword removal).

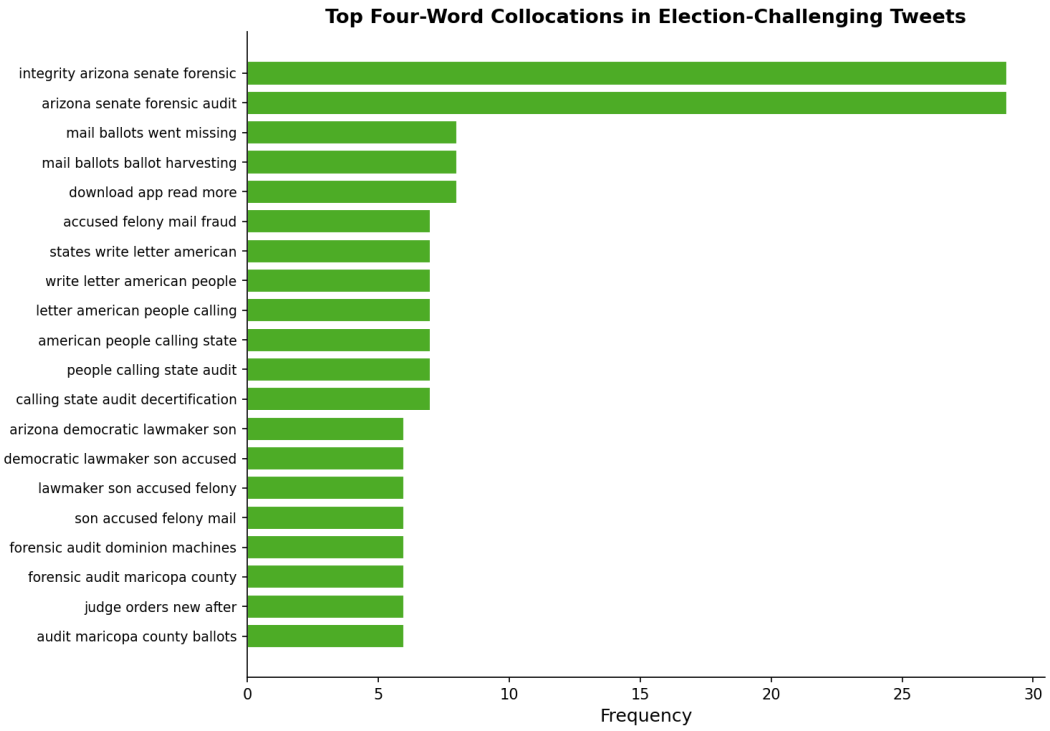

Figure 9: Top four-word collocations in challenge tweets (lambda score).

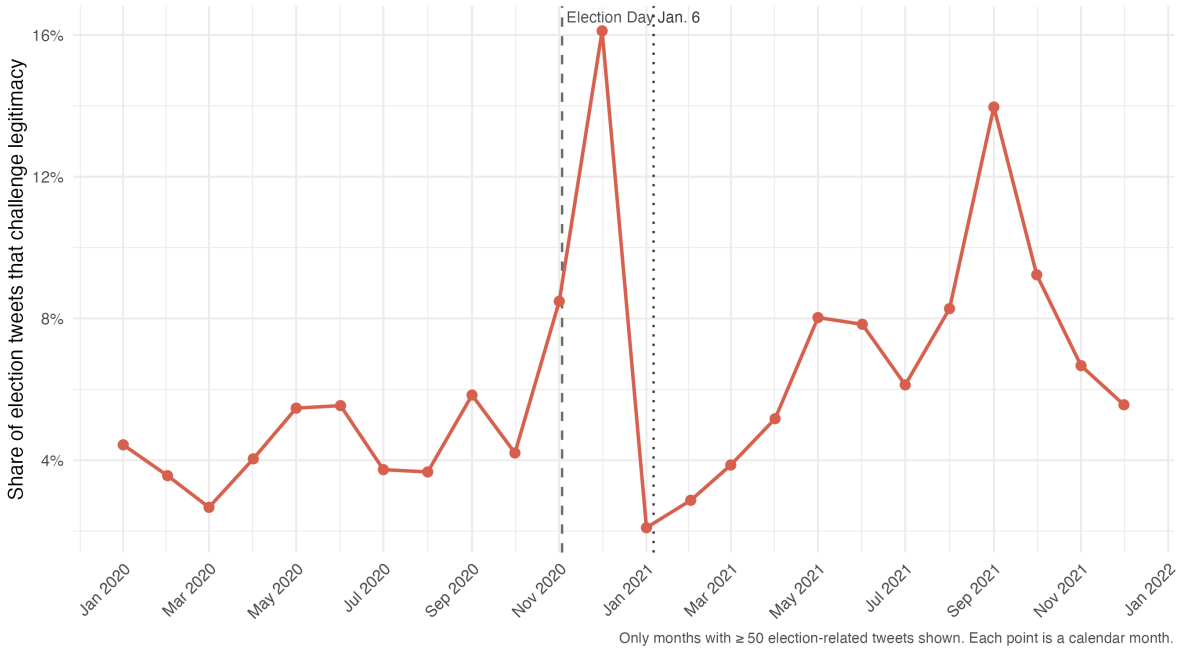

Figure 10: Monthly composition of election-related legislator tweets, 2019–2021. Proportion of challenge vs. non-challenge tweets by month.