
Introducing AI to an online petition platform changed outputs but not outcomes

In the format provided by the
authors and unedited

List of Figures

1	Empirical CDF of Outcomes	3
2	AI Detection Time Series	4
3	AI Detection Labeling Task	5
4	Writing Quality and Persuasiveness Boxplots	6
5	Writing Quality and Persuasiveness Marginal Effect on 1+ Comment Outcome	7
6	Writing Quality and Persuasiveness Marginal Effect on 10+ Signature Outcome	8
7	Correlation Between Lexical Features and Labeled Writing Quality/Persuasiveness	9
8	Sentiment KDE	10
9	Title concreteness KDE	11
10	Parallel Trends	12
11	Difference-in-differences for Alternative Outcome Metrics	13
12	Difference-in-differences for Alternative Lexical Features	14
13	Difference-in-differences for Hypothetical Narrow Pre-Treatment Period	15
14	Difference-in-differences for Hypothetical Longer Post-Treatment Period	16
15	Difference-in-differences with Short Petitions Included	17
16	Difference-in-differences with Covariates	18
17	Difference-in-differences with SE Clustered By Country	18
18	Difference-in-differences for Share of Petitions about Government Topics	19
19	Difference-in-differences for Government Topic Subset	20
20	Placebo Difference-in-differences	20
21	Time Series of Petition Title Verb Usage	21
22	Time Series of “Implement” Usage in Petition Titles	22
23	Time Series of main Lexical and Outcome Metrics	23
24	Time Series of English Petitions Classified by English Variety	24
25	Difference-in-differences with only U.S. as Treated Group	25
26	Difference-in-differences with Each Country Compared Alone to Australia	26
27	Outcomes by English Variety by Country	27
28	Distribution of Petition Topics	28
29	Incident of Same Topic Among Repeat Petition Writers	29
30	Difference-in-differences on Civic Engagement Topic Subset	29
31	Difference-in-differences on Non-Civic-Engagement Topic Subset	30
32	Marginal Estimates of Lexical Features Regressed on 1+ Comment Outcome	31
33	Repeat Petition Writer Outcomes and Lexical Features	32
34	Synthetic Control Time Series: US Treated Unit	33
35	Synthetic Control Time Series: CA Treated Unit	34
36	Synthetic Control Time Series: GB Treated Unit	35
37	BSTS: Time Series of User Participation	36
38	BSTS: Time Series of Petition Production	37
39	Interrupted Time Series of Lexical and Outcome Metrics	38

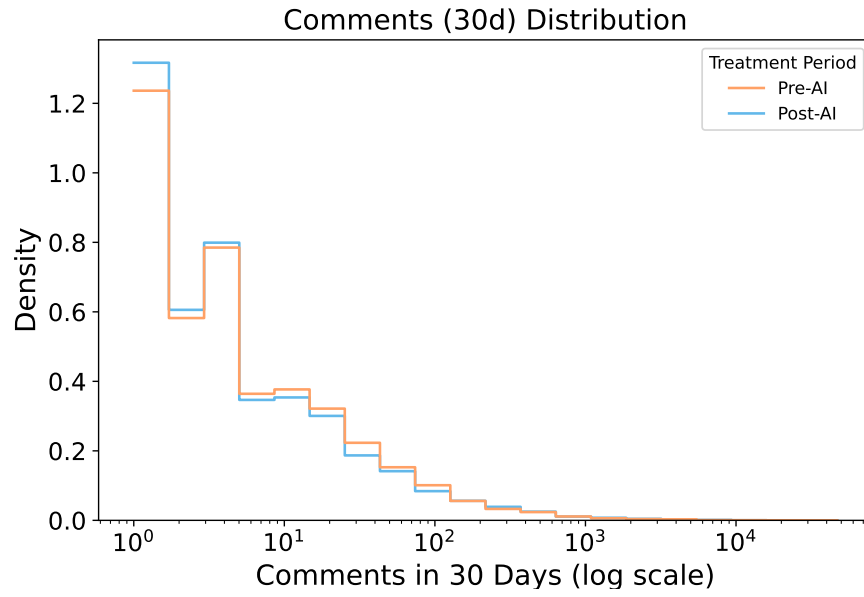
List of Tables

1	Balance table	38
2	Pre-treatment (January 3, 2022 - April 2, 2023) descriptive statistics of main text metrics . .	39
3	Post-treatment (October 2, 2023 - December 15, 2023) descriptive statistics of main text metrics	39
4	Labeling Task Measures	39
5	GLM Estimates: Regressing Outcome (1+ Comment 30days) on AI Period and Persuasiveness	39
6	GLM Estimates: Regressing Outcome (1+ Comment 30days) on AI Period and Writing Quality	39
7	GLM Estimates: Regressing Outcome (10+ Signature) on AI Period and Persuasiveness . . .	40
8	GLM Estimates: Regressing Outcome (10+ Signature) on AI Period and Writing Quality . .	40
9	Static Difference-in-differences: Percent 1+ Comment (30 Days)	40

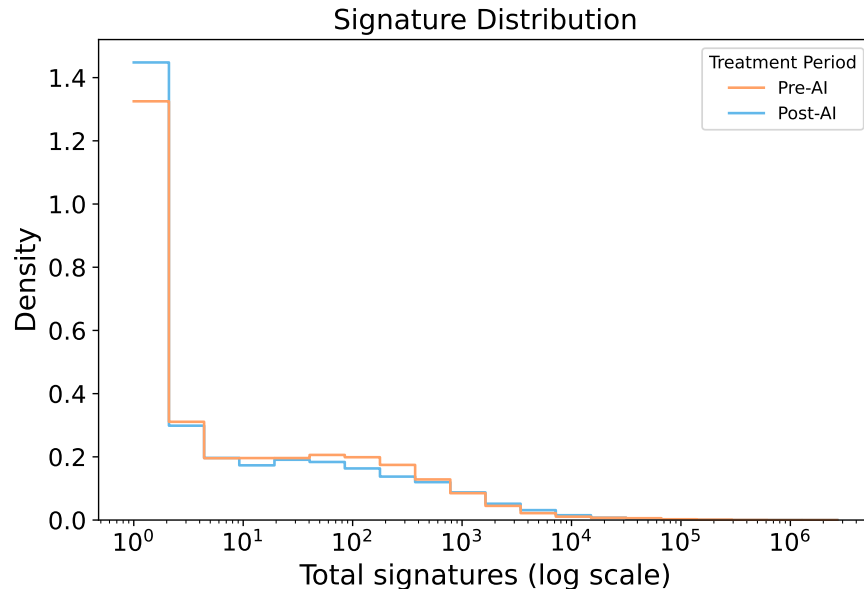
51	10	Static Difference-in-differences: Percent with 10+ Signatures	40
52	11	Static Difference-in-differences: MATTR	40
53	12	Static difference-in-differences: Word Count	41
54	13	Static Difference-in-differences: Flesch-Kincaid Grade Level	41
55	14	Difference-in-differences for Mean Pairwise Similarity	41
56	15	Static Difference-in-differences with Country-Clustered Errors: Percent with 1+ Comment (30d)	41
57	16	Static Difference-in-differences with Country-Clustered Errors: Percent with 10+ signatures	41
58	17	Static Difference-in-Differences with Country-Clustered Errors: MATTR	42
59	18	Static Difference-in-Differences with Country-Clustered Errors: Word Count	42
60	19	Static Difference-in-Differences with Country-Clustered Errors: Flesch-Kincaid Grade Level	42
61	20	Mixed Effects Model for Outcome Metrics	43
62	21	Mixed Effects Model for Lexical Metrics	44

63 Contents

64	Change.org and the Write with AI Tool	45
65	Data	45
66	AI Detection	46
67	Petition Writing Quality and Persuasiveness	47
68	Sentiment Analysis	48
69	Title Concreteness	48
70	Difference-in-Differences	49
71	Difference-in-Differences Robustness Checks	49
72	Verb Frequency	51
73	Petition Language Variant	51
74	Petition Topic	52
75	Predictive Strength of Lexical Features	54
76	Repeat Petition Writers	54
77	Synthetic Control	55
78	Bayesian Structural Time Series	56
79	Interrupted Time Series	56



(a)



(b)

Figure 1: Distribution of outcome measures (a) total comments in 30 days and (b) total signatures in pre- and post-AI periods (excluding A/B test period) during the period of study for the main analysis, January 2022 - December 2023, using 237,020 petitions written during the pre-AI period and 42,111 petitions written during the post-AI period. Post AI-the density of low-outcome petitions increases.

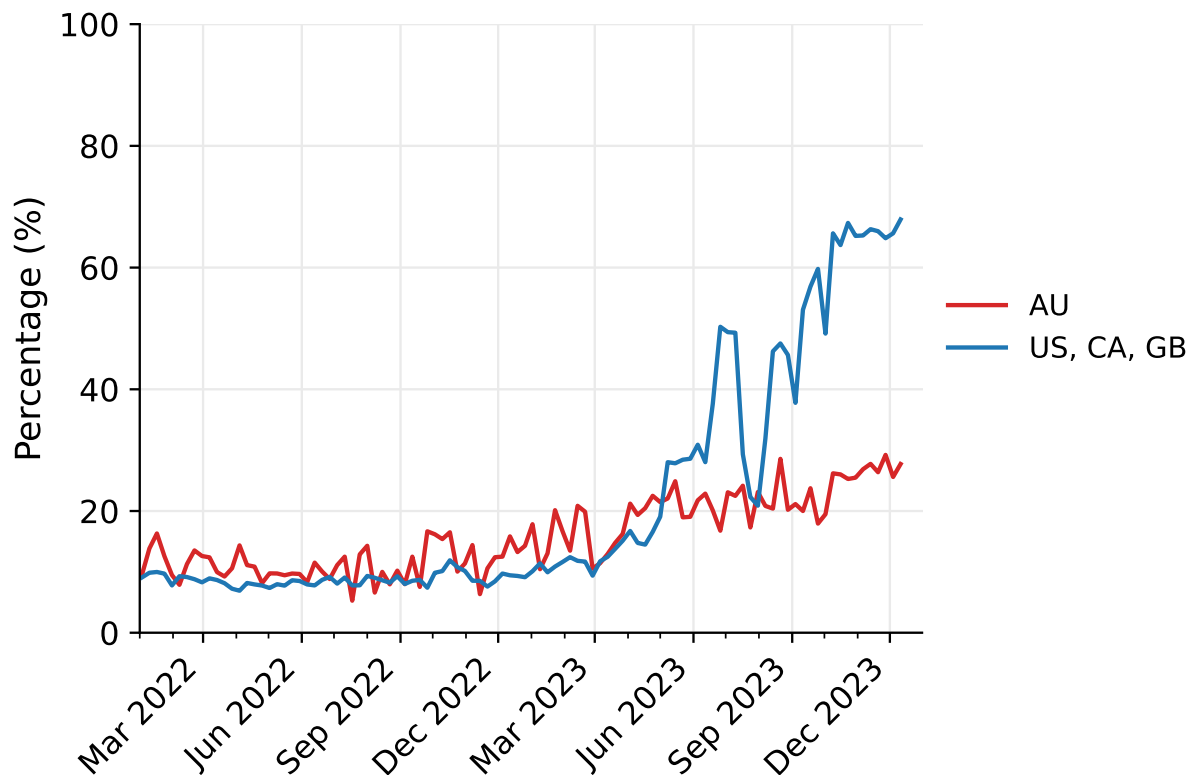


Figure 2: Time series showing the share of petitions, by week, that our ensemble model classifies as AI-generated in English petitions from treated and control countries ($N = 361,635$ petitions). The share of AI-generated petitions increases sharply after the introduction of the AI-generated draft feature in countries with access to the tool (US, GB, CA). As of the week of December 11, 2023 (the last week prior to AU gaining access to the in-platform AI tool), most (67.9%) of petitions written with access to the AI-generated draft feature on Change.org were generated with AI, while only 27.7% of petitions written in AU were labeled as AI-generated. Positive AI detection prior to April, 2023 reflects false positives or off-platform AI usage. Our ensemble classifier achieved 99% accuracy and 99% precision on a test set of 674 annotated petitions.

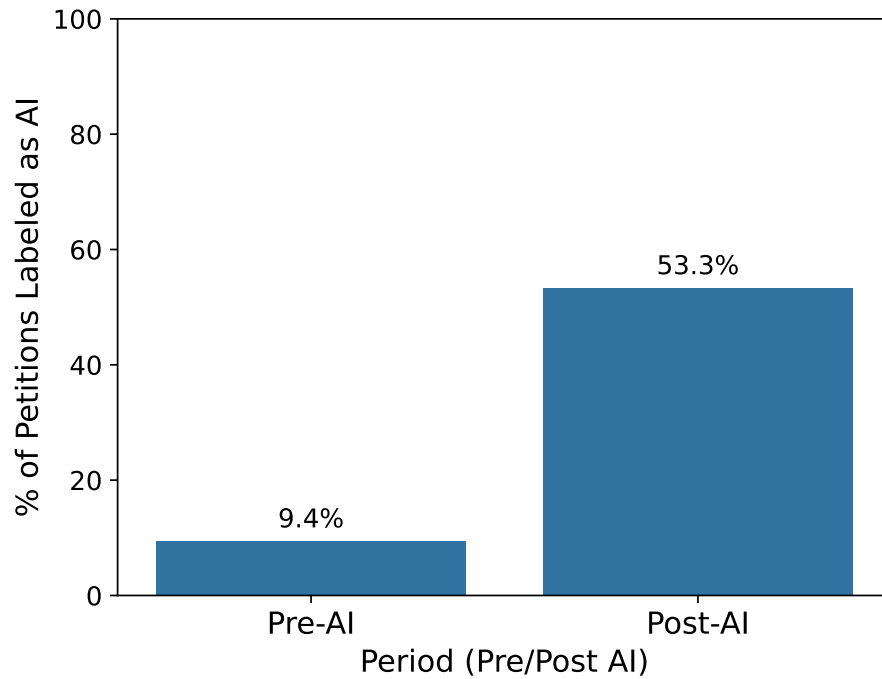
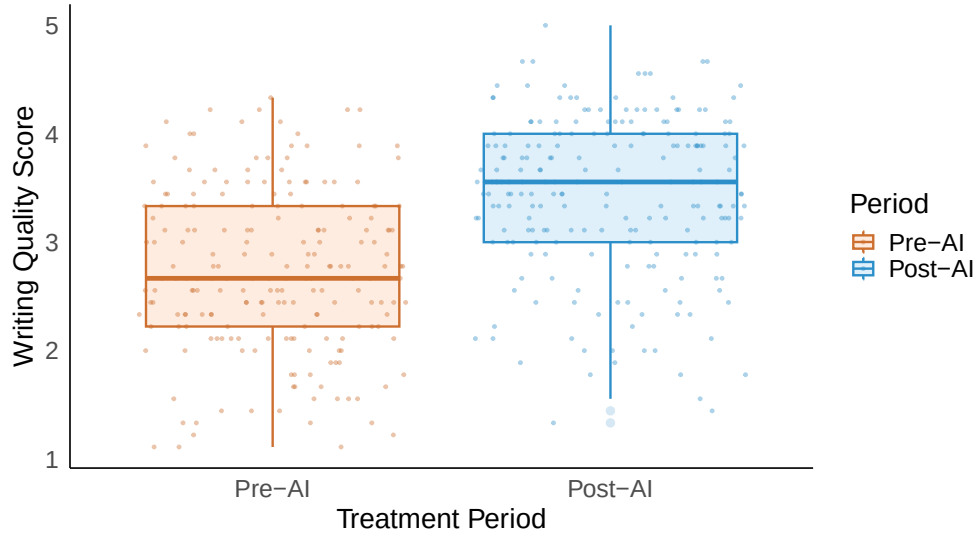
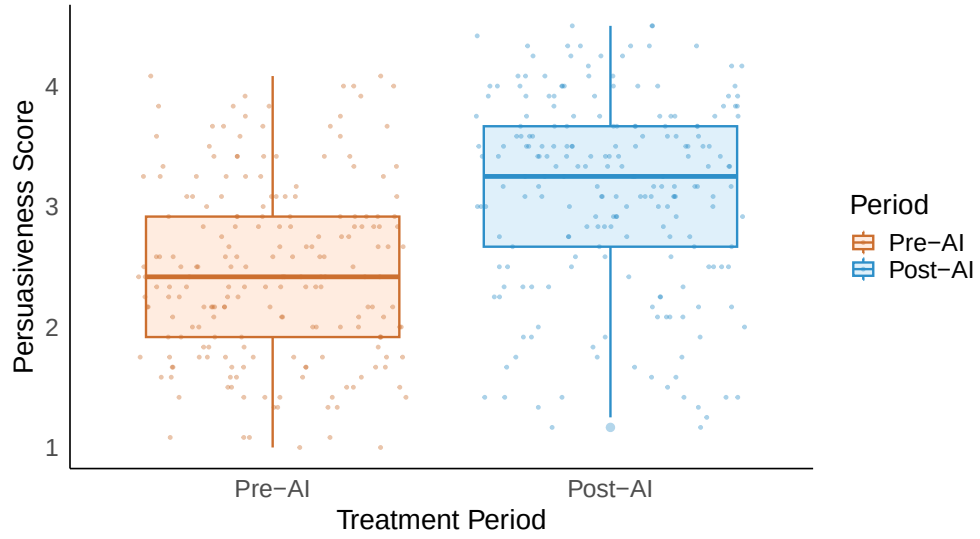


Figure 3: We use our ensemble method for AI classification to predict if petitions in the sample used for labeling writing quality ($N = 417$ petitions) are written by the AI tool. 9.4% of petitions in the pre-AI sample were flagged as AI-written, while 53.3% of the post-AI petitions were flagged as AI-written.

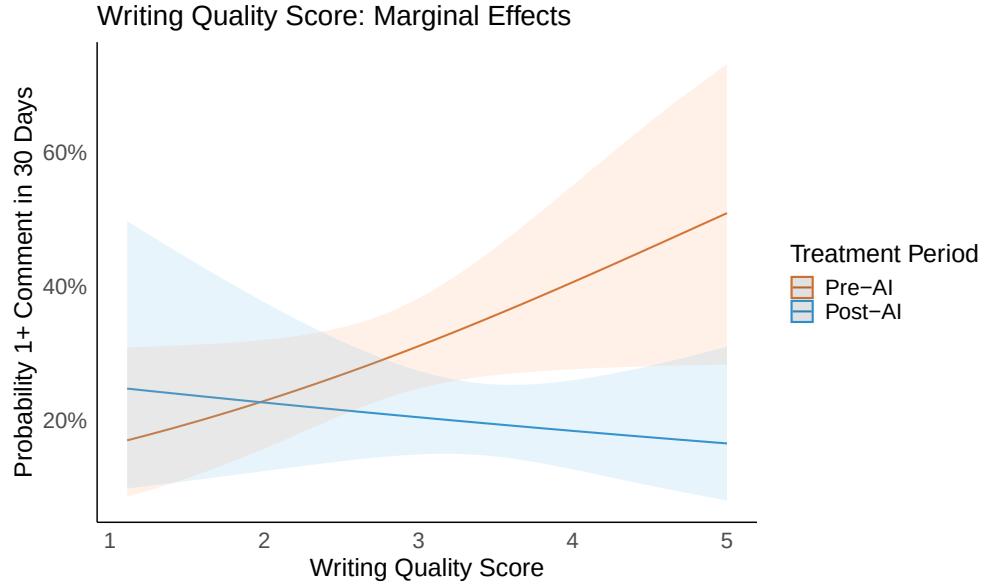


(a)

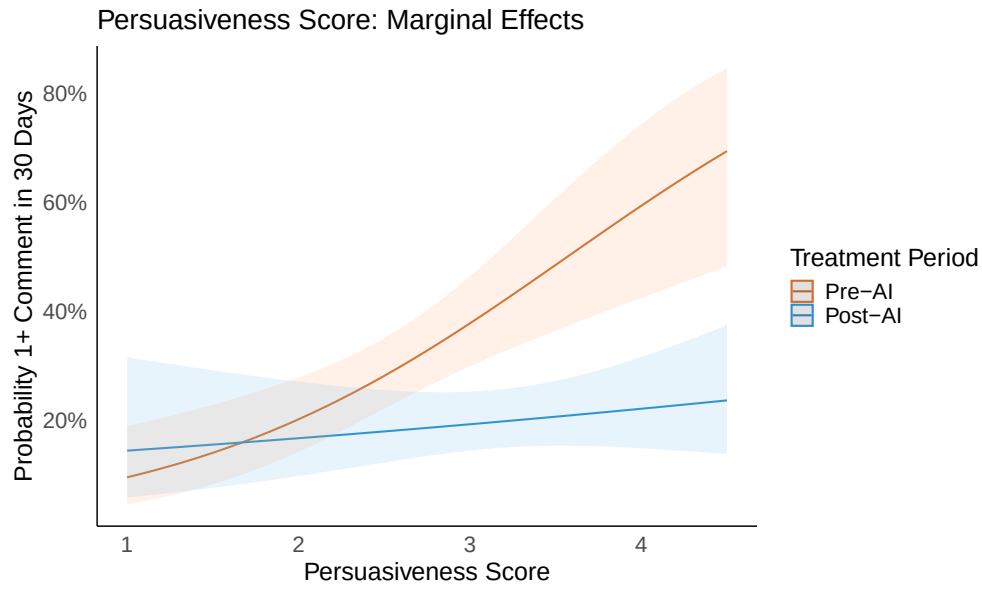


(b)

Figure 4: Boxplots showing the distribution of (a) writing quality score and (b) persuasiveness scores of 203 petitions written Pre-AI (before in-platform AI access) and 214 petitions written Post-AI (after in-platform AI access). The distribution of writing quality and persuasiveness scores shifts to be higher among petitions written with AI access.



(a)

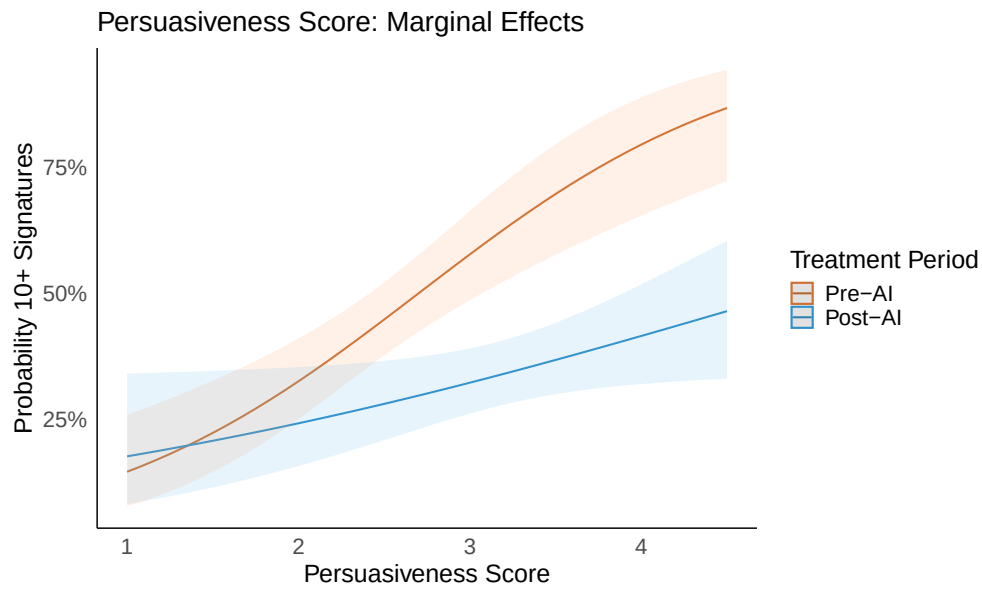


(b)

Figure 5: Marginal effect of (a) writing quality score and (b) persuasiveness score regressed on outcome metric (1+ comment in 30 days), by treatment period (Pre-AI, Post-AI). Model uses a sample of 417 petitions labeled for writing quality and persuasiveness. Pre-AI, increasing writing quality and persuasiveness scores were predictive of an increased probability of receiving a successful outcome (i.e., reaching 1+ comment in 30 days). Post-AI, increasing persuasiveness and writing quality scores were not predictive of improved probability of receiving a successful outcome. In the case of writing quality, higher writing quality scores were weakly negatively predictive of the outcome metric. Figure shows 95% CI.



(a)



(b)

Figure 6: Marginal effects of (a) writing quality score and (b) persuasiveness score regressed on petition outcome (10+ signatures), by treatment period (Pre-AI, Post-AI). Model uses a sample of 417 petitions labeled for writing quality and persuasiveness. Pre-AI, higher writing quality and persuasiveness scores were predictive of a higher probability of a successful outcome. Post-AI, higher persuasiveness and writing quality scores were attributed with a reduced increase in probability of a successful outcome. Figure shows 95% CI.

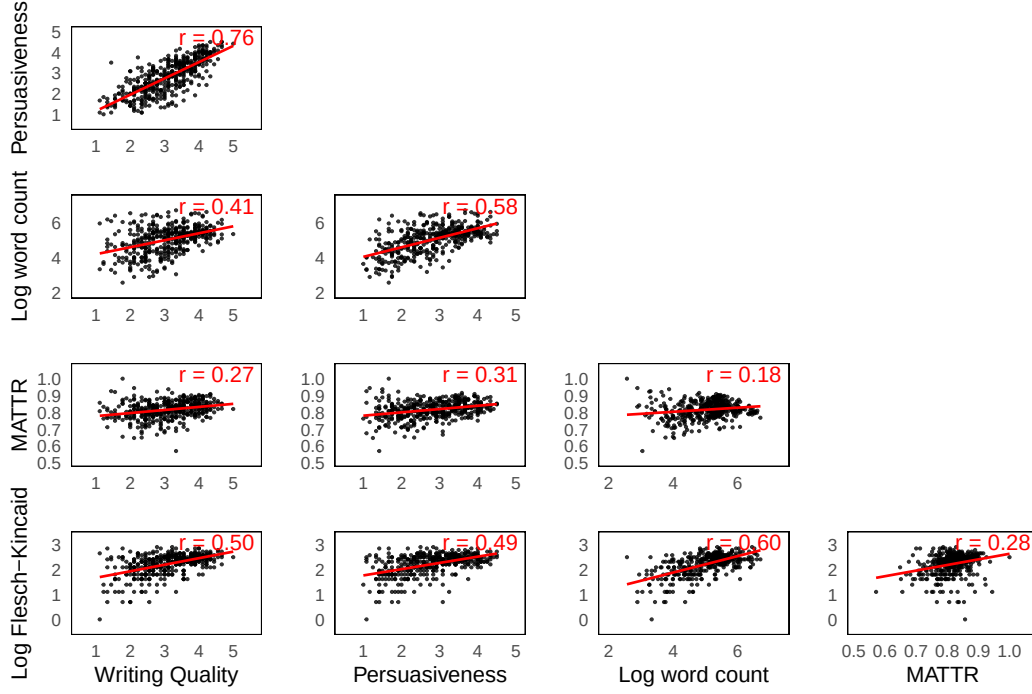


Figure 7: Correlation between lexical measures (log-transformed word count, MATTR, and log-transformed Flesch-Kincaid grade level) and measures of petition text quality (writing quality, persuasiveness) for a random sample of pre- and post-AI annotated petitions. We then drop four petitions for which Flesch-Kincaid score is negative to permit log transformation ($N = 413$ petitions). We observe positive correlations between lexical features and quality measures. Scatterplots include a linear fit line and the Pearson correlation coefficient in red text.

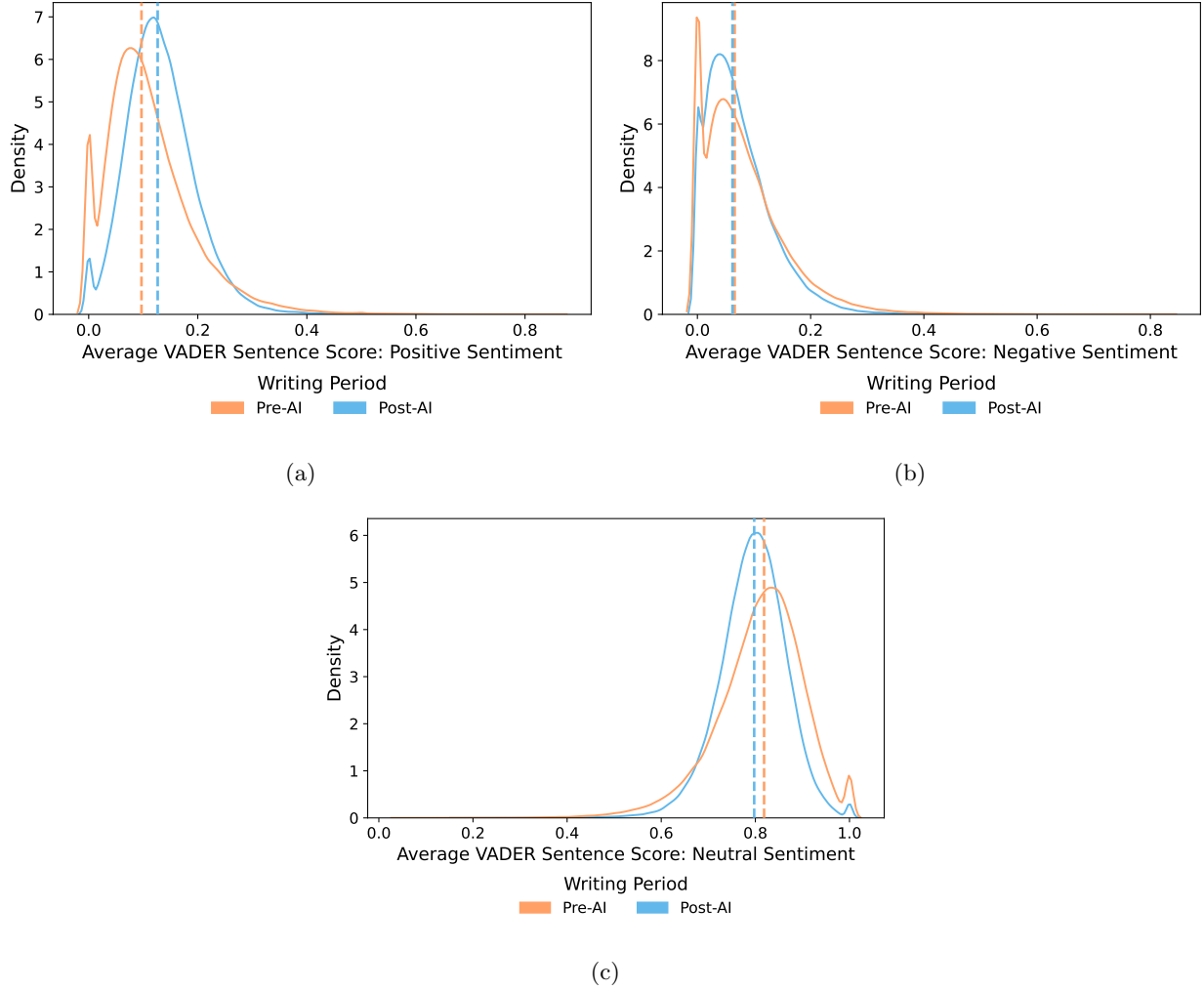


Figure 8: Distribution of average sentence VADER sentiment scores for English language petitions (237,020 petitions pre-AI, 182,229 petitions post-AI) written in the US, GB, CA, AU between January 2022-December 2024 (excluding A/B test period and period of differential access) by (a) positive sentiment, (b) negative sentiment, and (c) neutral sentiment. Vertical lines reflect median values. We observe that post-AI, the distribution of positive sentiment shifts to the right, suggesting petitions written with access to AI tend to be more positive in valence. The distribution of negative sentiment also changed post-AI, with a greater density of petitions at the mode of 0.1 negative sentiment. Finally, we see a decrease in neutral sentiment, with the distribution shifting to the left.

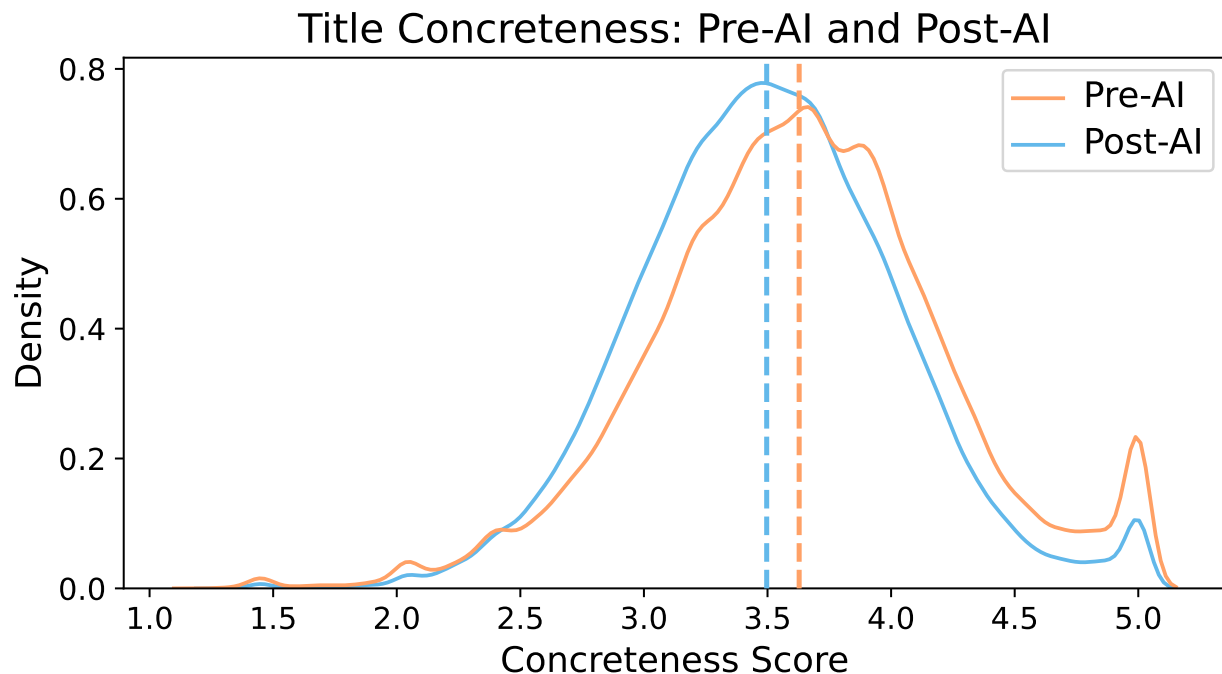


Figure 9: Distribution of title concreteness for English language petitions (237,020 petitions pre-AI and 182,229 petitions post-AI) written in the US, GB, CA, AU between January 2022-December 2024 (excluding A/B test period and period of differential access). Vertical lines display median title concreteness scores pre- and post- AI. We observe that the distribution of title concreteness decreases with in-platform AI access, with the median value decreasing from 3.63 to 3.50 (Kruskal-Wallis test $H(1)=5341.433$, $p < 0.001$, $\eta^2 = 0.011$).

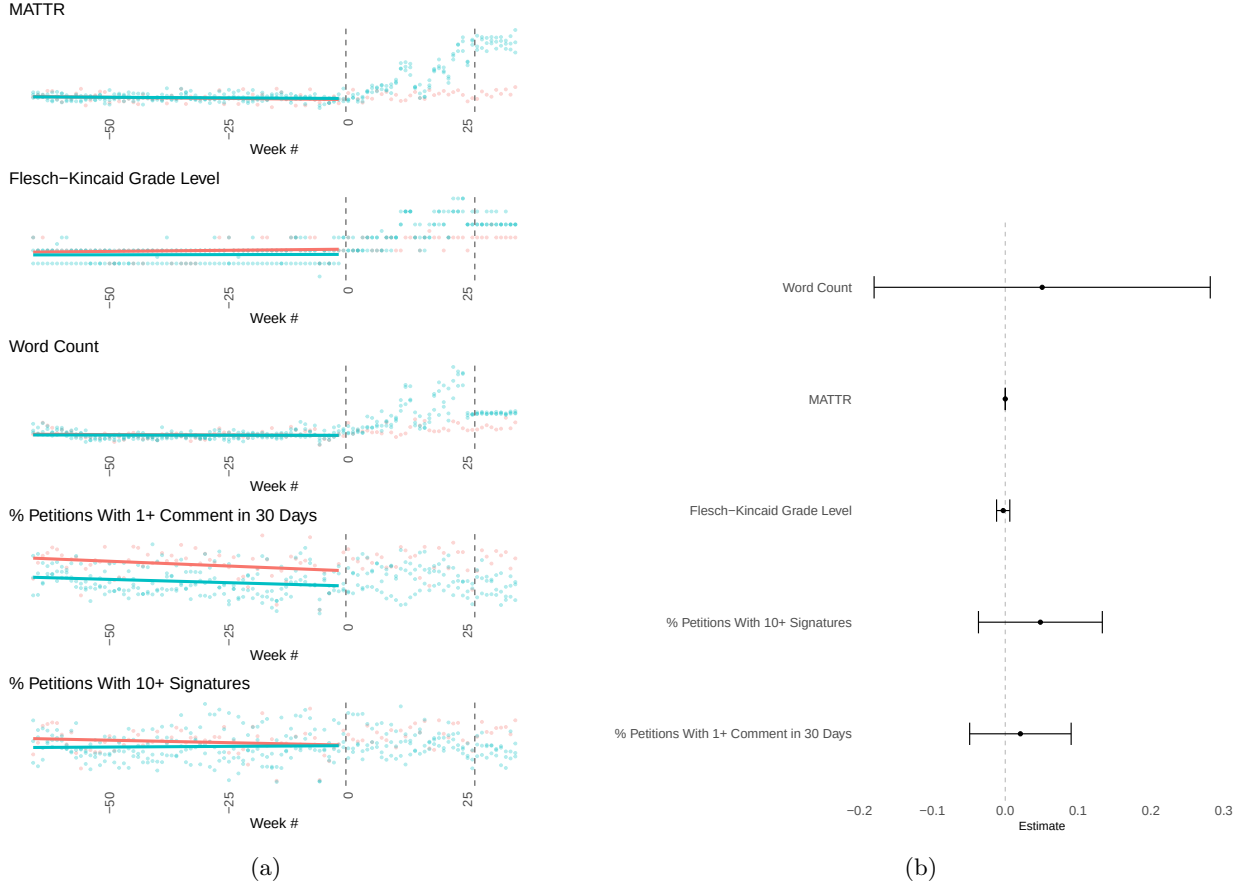


Figure 10: To assess if our core lexical features and petition outcomes satisfy the assumption of parallel trends pre-intervention, we fit simple linear models to predict the metric by week for both treated and untreated groups. We used $N = 237,020$ petitions from the pre-treatment period. Panel (a) shows the fit regression line to control and treated units in the pre-AI period. Panel (b) shows 95% CI on the coefficient of the interaction term by week and treatment group. We find for each metric that the interaction is insignificant, as the confidence intervals all include 0.

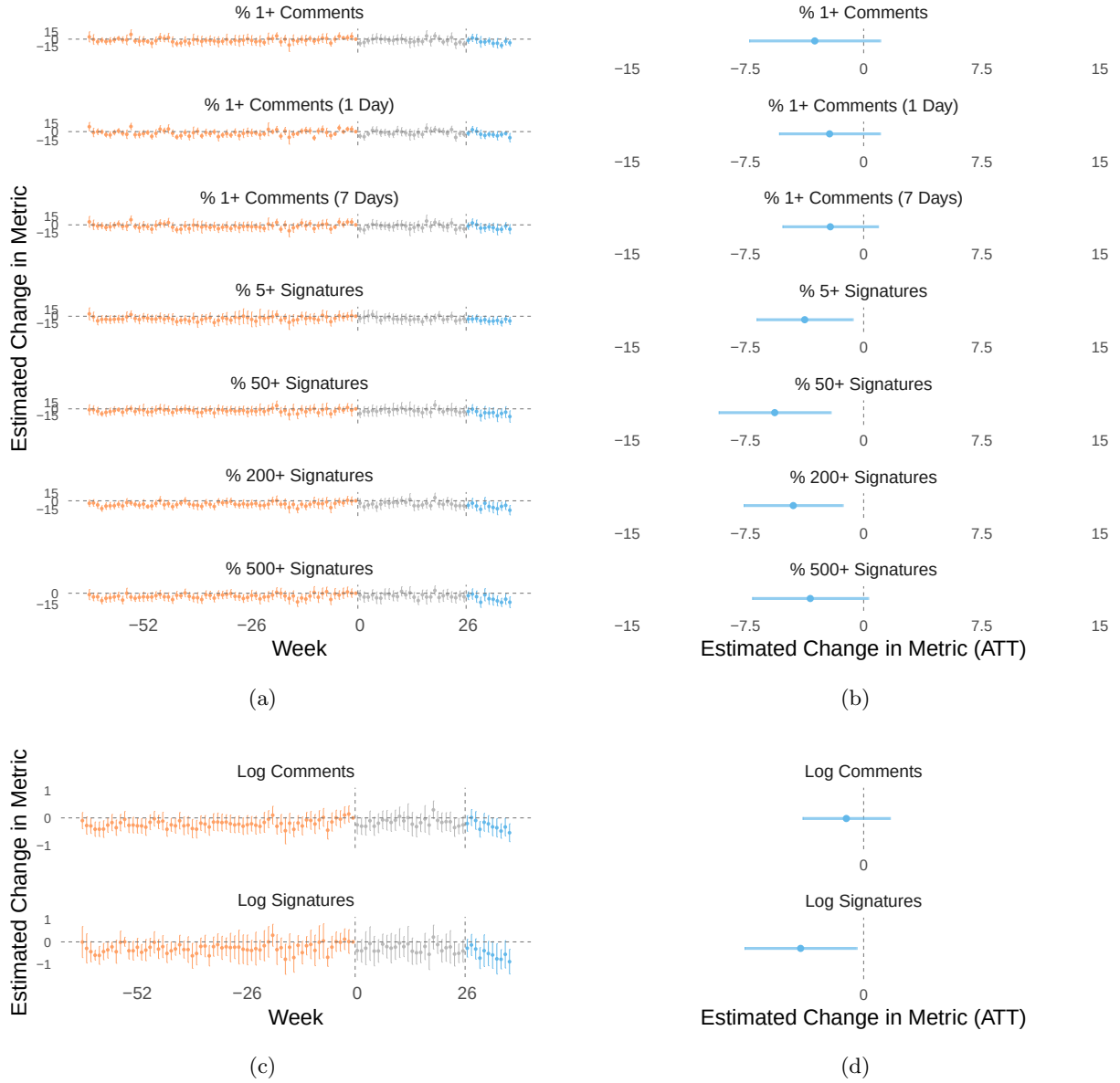


Figure 11: Static and dynamic difference-in-differences estimates for additional outcome measures. In the main text we report one set of thresholds for petitions: 1 comment in 30 days and 10 signatures. Here we show that the results observed in the main text (insignificant changes, or significant negative effects) are consistent for a variety of alternative thresholds. Panels (a) and (b) show the estimated change in percentage points for additional thresholds of comments and signatures. Panels (c) and (d) show the estimated change in log-transformed comment and signature counts, normalized by the number of petitions posted during the time period. We log-transform comments and signatures due to heavy positive skew. We report 95% CI bands and use the same model specification as outcomes in main analysis. Pre-AI period was 65 weeks ($N = 237,020$ petitions), A/B test period was 26 weeks ($N = 82,504$ petitions), and the Post-AI period was 11 weeks ($N = 42,111$ petitions).

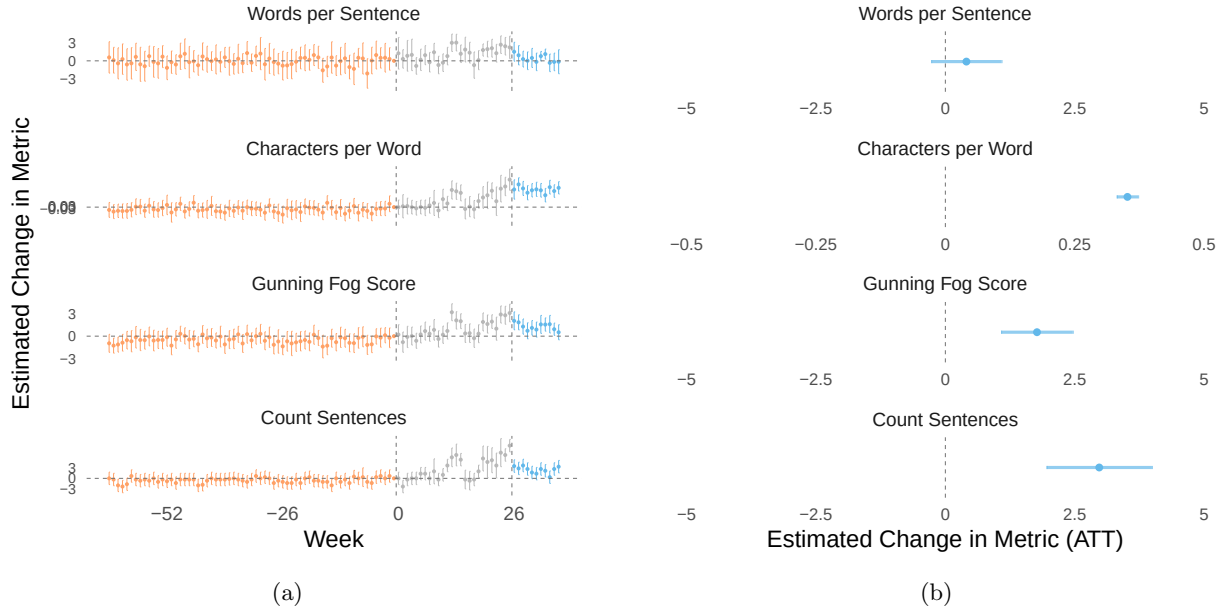


Figure 12: Panel (a) shows dynamic difference-in-differences and Panel (b) shows static difference-in-differences for additional measures of lexical style. We report 95% CI bands and use the same model specification as thresholds in main text. Estimates show expected change in median lexical feature for treated countries with access to AI. Pre-AI period was 65 weeks ($N = 237,020$ petitions), A/B test period was 26 weeks ($N = 82,504$ petitions), and the Post-AI period was 11 weeks ($N = 42,111$ petitions).

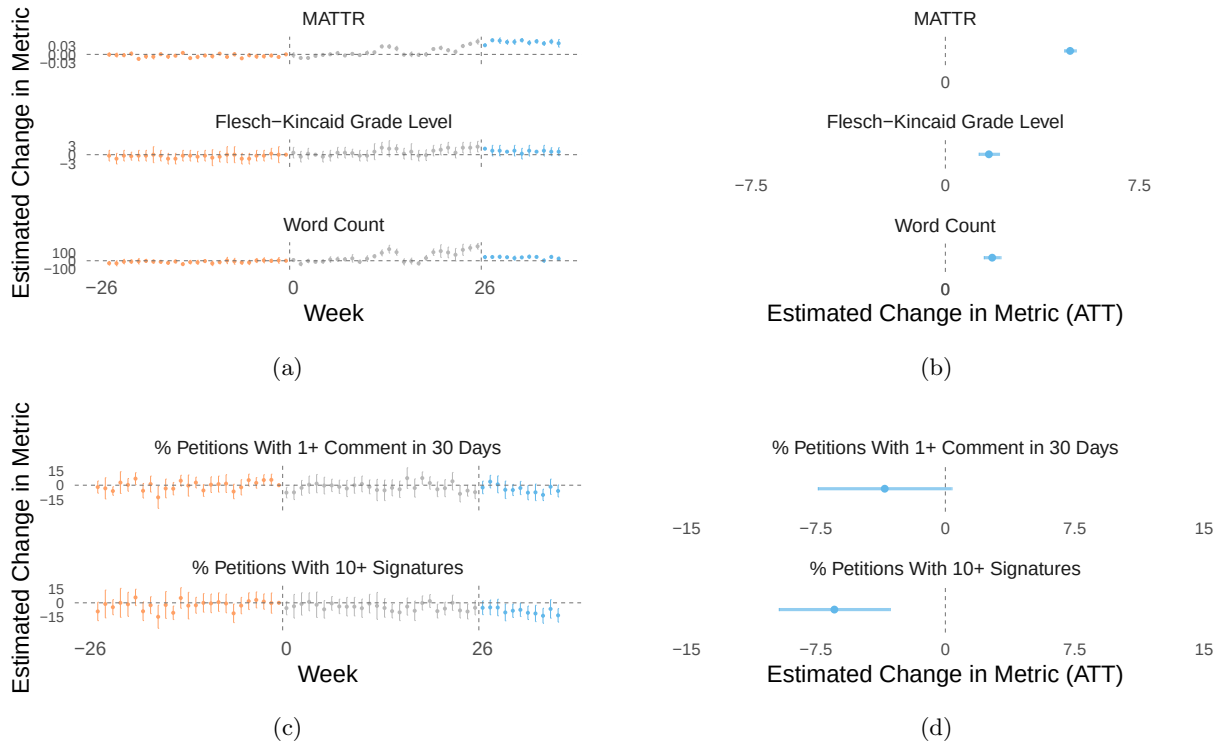


Figure 13: Panels (a) and (c) show dynamic difference-in-differences analysis and Panels (b) and (d) show static difference-in-differences estimates for lexical features and outcomes. Here, we report the same metrics captured in the main text, with a shorter time frame (25 weeks pre-AI, rather than 65 weeks). We use the shorter time frame as a robustness check to demonstrate results are not driven by trends in early 2022. Our results are consistent with the shortened pre-AI period. This data includes 83,363 petitions posted during the pre-AI period, 82,504 petitions posted during the A/B Test period, and 42,111 petitions posted in the post-AI period.

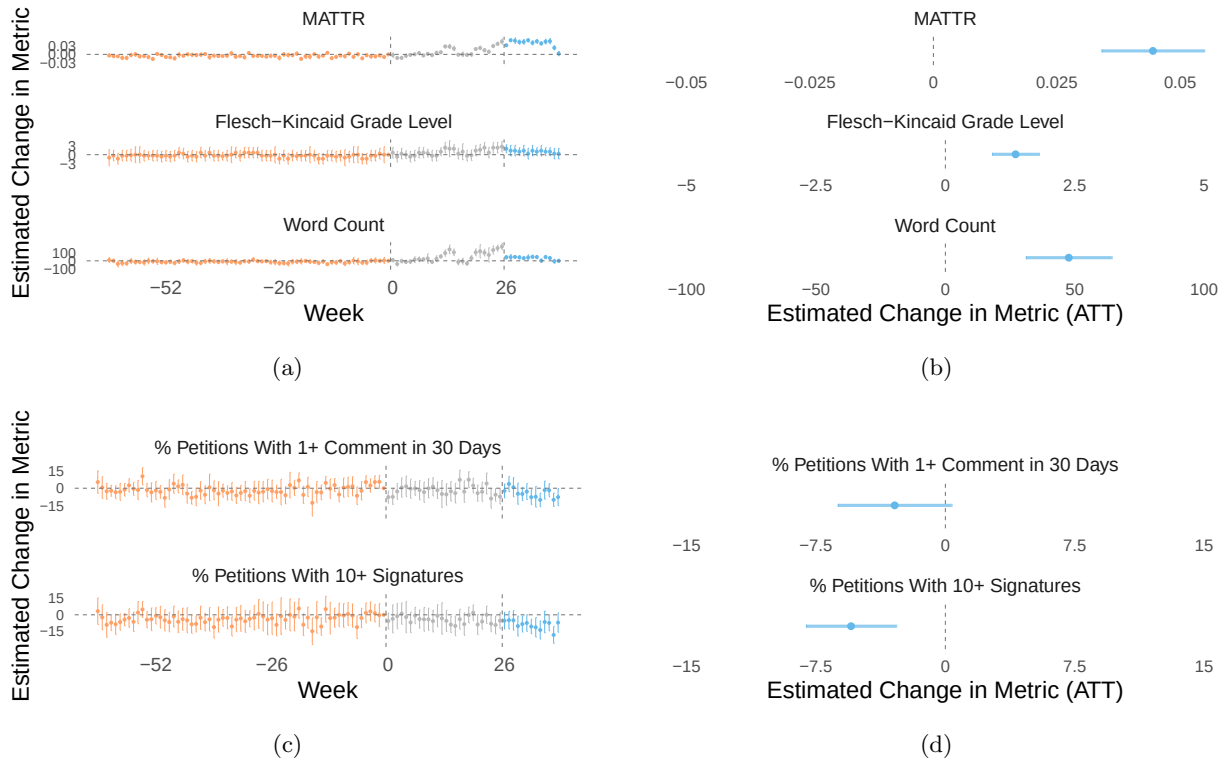
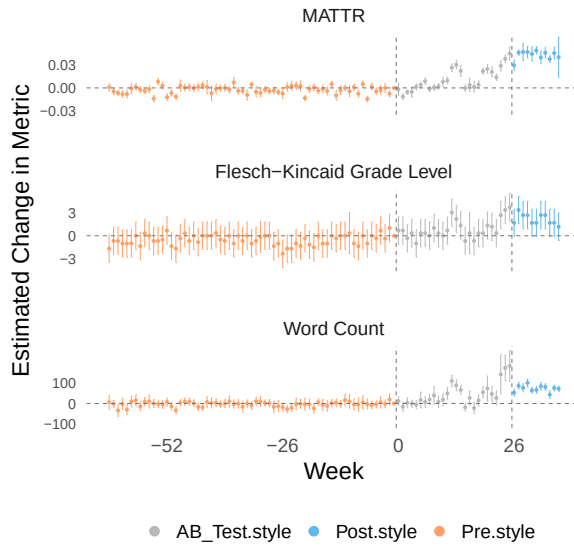
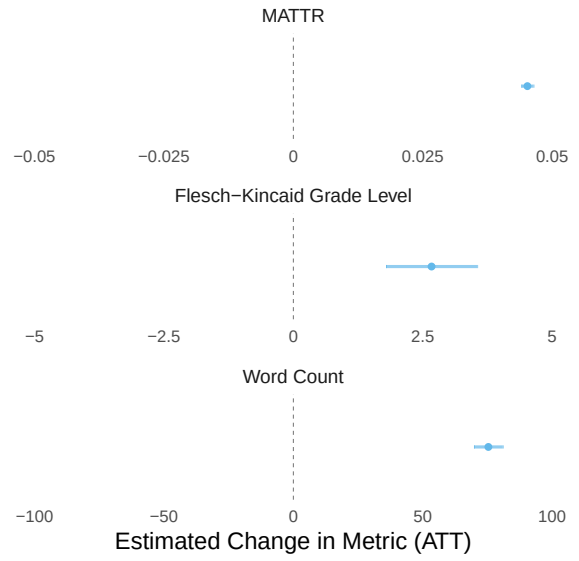


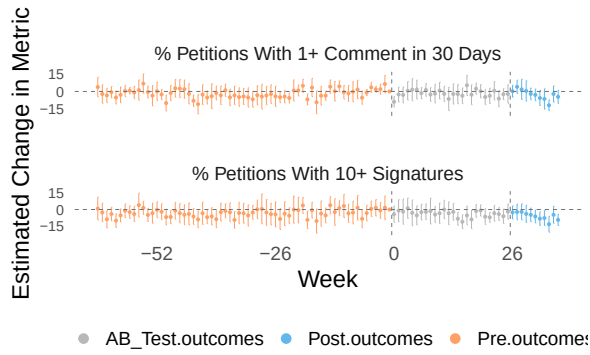
Figure 14: Difference-in-differences with the full month of December included as the post-AI period of differential access. Analysis includes 237,020 petitions during the Pre-AI period, 82,504 petitions during the A/B test period, and 47,539 petitions during the post-AI period. The post-AI period is extended from 11 weeks to 13 weeks (spanning October 2, 2023-December 31, 2023). Results are similar to those calculated using the eleven-week period in our main-text; however, the reduced gap between treated and control country style metrics is apparent in the final two weeks, as we expect control users did have access to in-platform AI during this period.



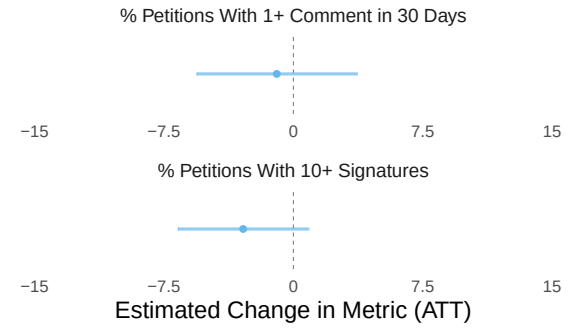
(a)



(b)



(c)



(d)

Figure 15: Dynamic and static difference-in-difference estimates for ATT estimate for lexical features (a,b) and outcomes (c,d) using a full dataset that did not exclude short text (fewer than 5 words; 2 sentences). This analysis includes 322,217 petitions posted during the pre-AI period, 104,843 petitions posted during the A/B test period, and 48,803 petitions posted during the post-AI period.

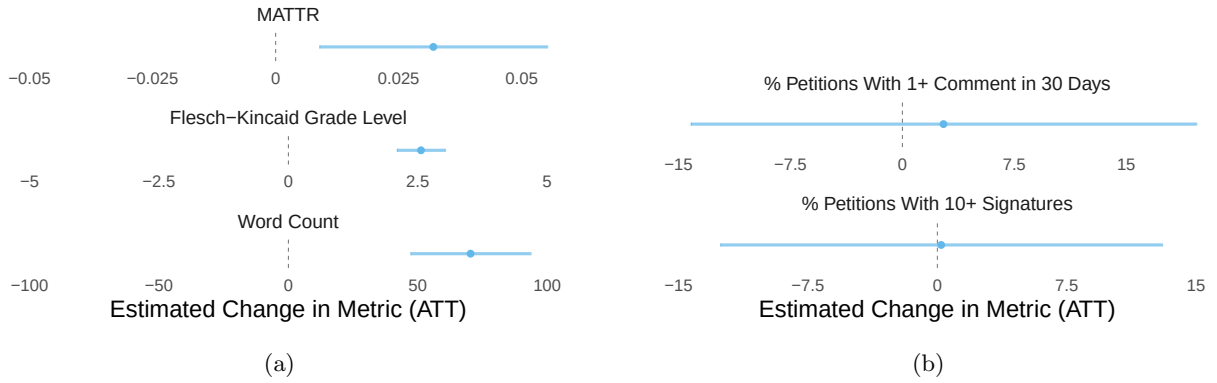


Figure 16: Static difference-in-differences with covariates (the share of petition-writers that had commented previously, the share of petition-writers that were new to the platform, the share of petitions with decision makers labeled, and the share of petitions that had been posted over the weekend). Panel (a) shows average treatment effect on treated units (ATT) for lexical features, (b) shows average treatment effect on treated units (ATT) for outcomes. Estimates are doubly robust. Figure shows 95% CI. Pre-AI period was 65 weeks ($N = 237,020$ petitions), A/B test period was 26 weeks ($N = 82,504$ petitions), and the Post-AI period was 11 weeks ($N = 42,111$ petitions).

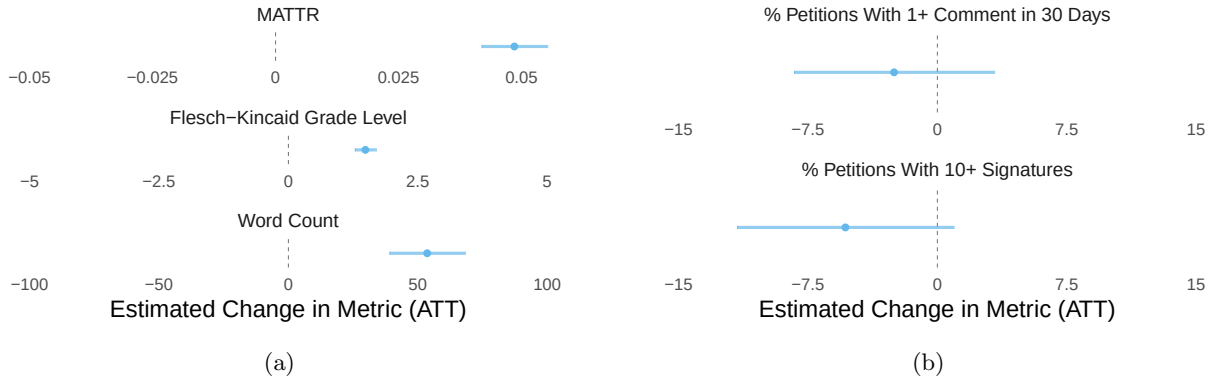


Figure 17: Static difference-in-differences with standard errors clustered by country. Panel (a) shows average treatment effect on treated units (ATT) for lexical features, (b) shows average treatment effect on treated units (ATT) for outcomes. The figure shows 95% CI. Pre-AI period was 65 weeks ($N = 237,020$ petitions), A/B test period was 26 weeks ($N = 82,504$ petitions), and the Post-AI period was 11 weeks ($N = 42,111$ petitions).

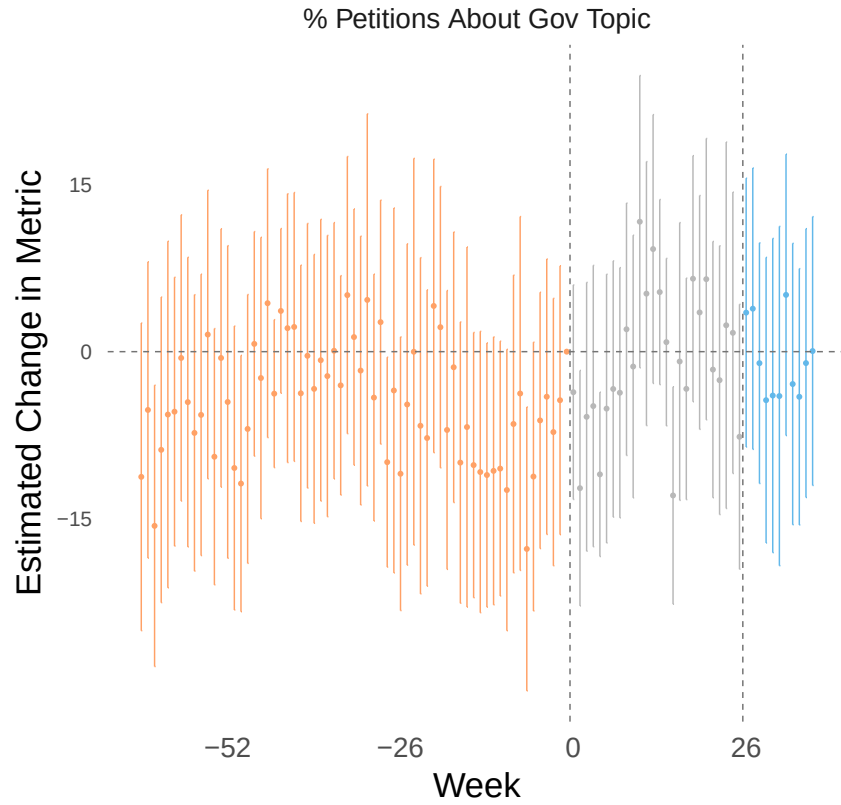


Figure 18: Dynamic difference-in-differences for the percent of petitions that are about a government topic. The y-axis reflects the percent of petitions published that week that we labeled as related to a government or political topic. The figure shows that post-AI weeks do not differ from pre-AI baseline (Week = -1). Pre-AI period was 65 weeks ($N = 237,020$ petitions), A/B test period was 26 weeks ($N = 82,504$ petitions), and the Post-AI period was 11 weeks ($N = 42,111$ petitions). The figure shows 95% CI.

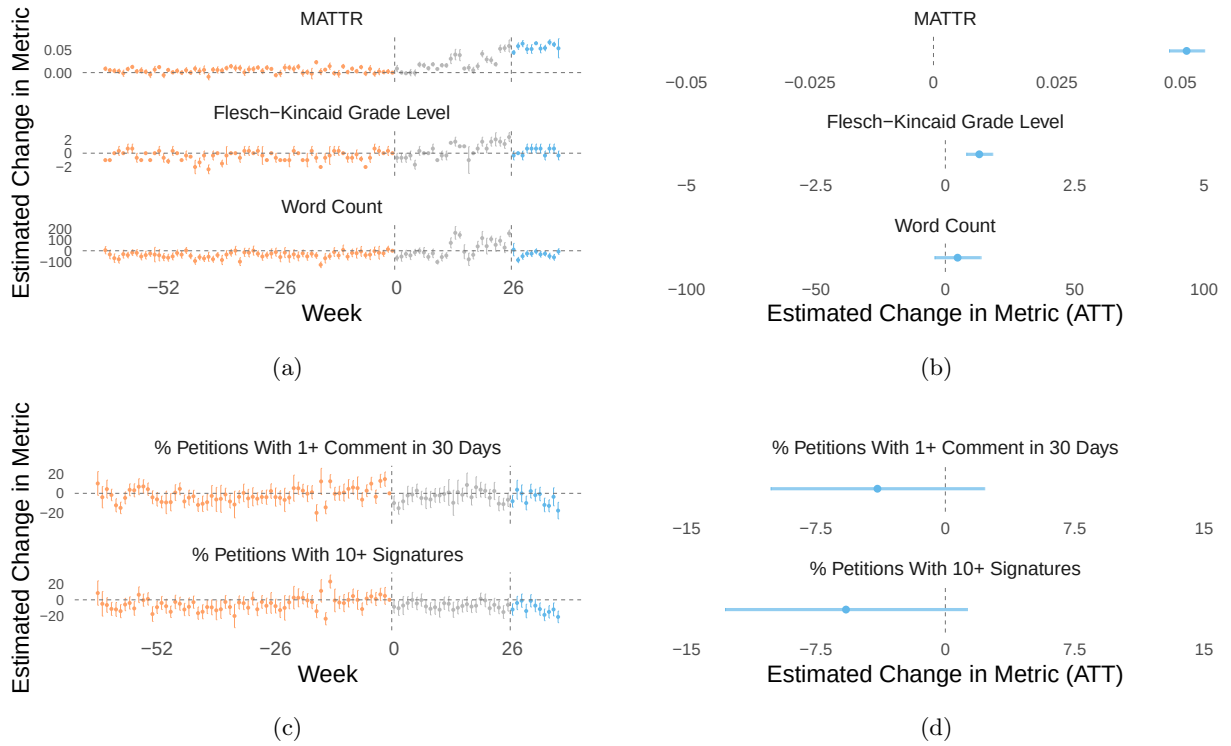


Figure 19: Difference-in-differences on a subset of petitions labeled as pertaining to a government topic, using 86,638 petitions written pre-AI, 33,957 petitions written during the A/B test period, and 17,051 petitions written post-AI. (a) shows dynamic difference-in-differences of lexical features, (b) static difference-in-differences of lexical features, (c) dynamic difference-in-differences of outcomes, (d) static difference-in-differences of outcomes. The figure shows 95% point-wise CI.

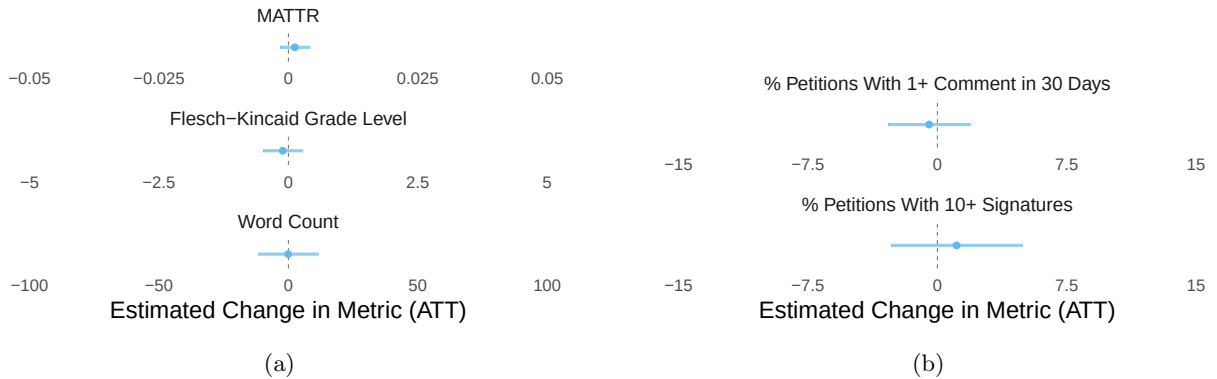
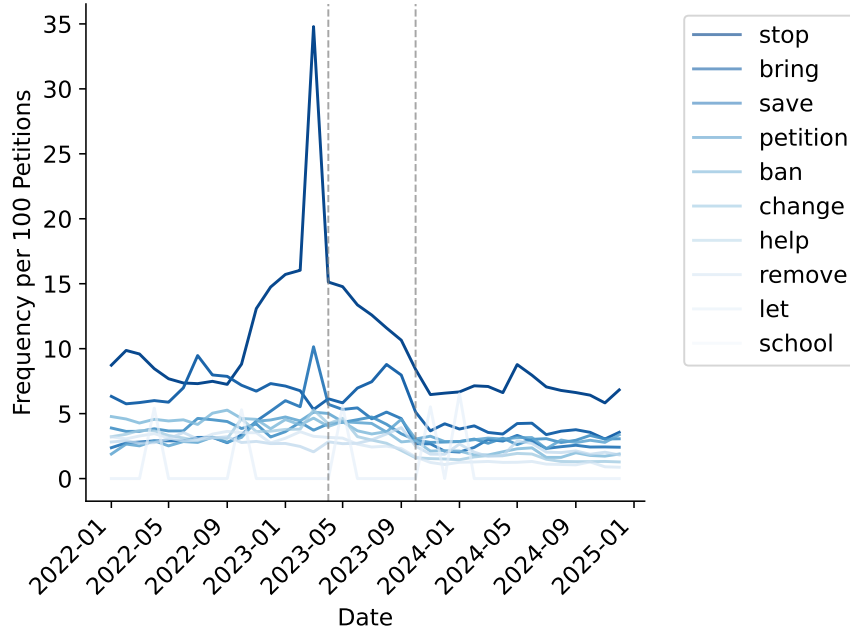
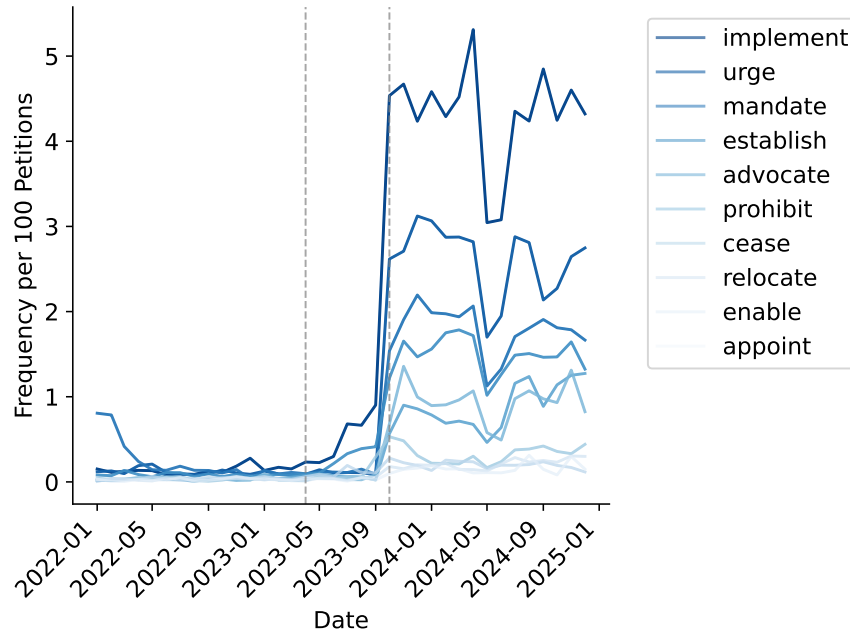


Figure 20: Static difference-in-differences estimates using placebo intervention date and A/B test period. Data is from January 3, 2022 to March 31, 2023. The placebo A/B test period is April 1, 2022 -October 1, 2022 and the placebo intervention date is October 1, 2022. The placebo data includes 13 weeks pre-intervention (59,186 petitions) and 26 weeks post-intervention (87,066 petitions). Panel (a) shows lexical features, Panel (b) shows outcomes. The figure shows 95% CI.



(a)



(b)

Figure 21: Time series showing most common verbs in petition titles written by users in the United States, Great Britain, and Canada (517,658 petitions). Both plots show the frequency per 100 words of verbs in petition titles. Panel (a) shows the 10 verbs that were most common in petition titles before November 2022. We select November 2022 as a threshold to identify terms that were popular before the release of on- or off-platform AI tools (November 2022: ChatGPT release). While some verbs spiked with relevant causes (e.g., “stop” during the viral #StoptheWillowProject movement in March, 2023), many of the verbs that were popular in 2022 titles retained consistent use after the introduction of AI. Panel (b) shows the 10 verbs that had the greatest ratio of post-ChatGPT frequency to pre-ChatGPT frequency. In both figures, the first dashed line shows the start of A/B testing of the AI feature, the second dashed line shows the start of full AI feature launch. The increase in verbs like “implement” with access to both general and in-platform AI tools suggests AI tools may change the language used in petitions.

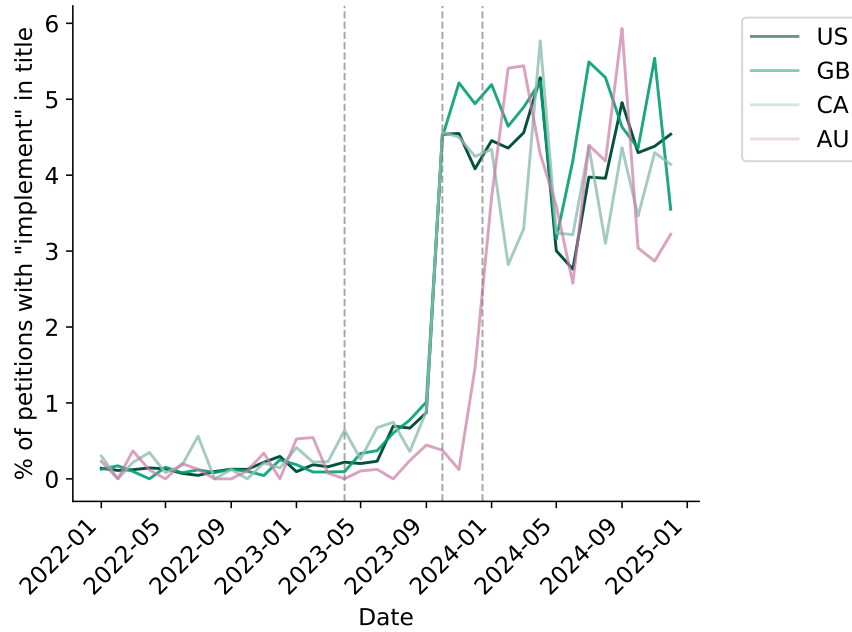


Figure 22: Time series of the percent of petitions per month, by country, that had the word “implement” in the title, among petitions written in the United States, Great Britain, Canada, and Australia. The word “implement” was the verb with the greatest ratio of post-AI frequency to pre-AI frequency among these countries (549,292 petitions). The figure demonstrates that the spike in the verb “implement” follows the timeline by which users received access to the in-platform AI tool.

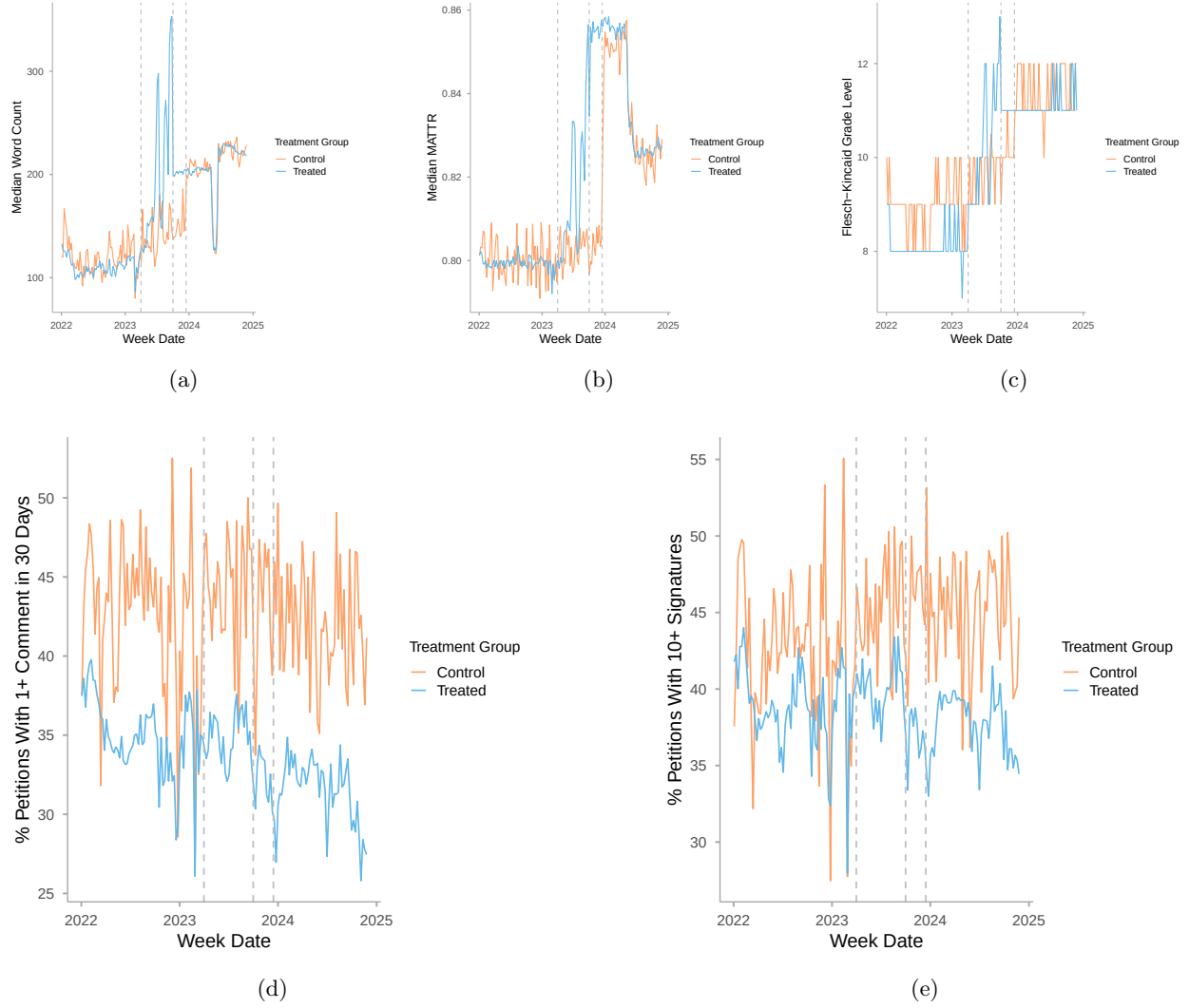


Figure 23: Time series of main lexical features and outcome metrics, by week. Treated countries include the United States, Canada, and Great Britain. The control country is Australia. Data is shown for weeks between January 3, 2022 and November 30, 2024 ($N = 536,598$ petitions). We exclude the final month of data collection to have a full month of data for comment and signature outcome metrics. Each figure has three dashed vertical lines, from left to right these lines have the intercepts of April 2, 2023; October 2, 2023; December 15, 2023. These lines correspond to the start of A/B test period in the treated countries; the full launch of the AI feature in treated countries; the full launch of the AI feature in Australia.



(a) Percent of petitions labeled American or British English



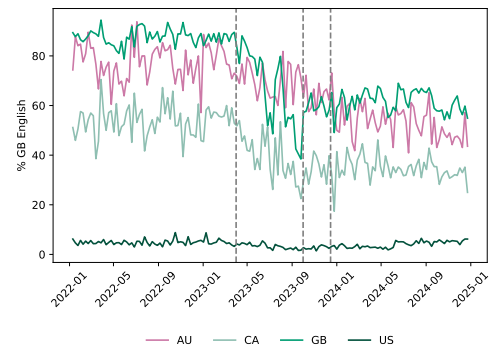
(b) Percent of petitions labeled American English, overall



(c) Percent of petitions labeled British English, overall



(d) Percent of petitions labeled American English, conditional on receiving a label



(e) Percent of petitions labeled British English, conditional on receiving a label

Figure 24: Percent of English language petitions by week (x-axis), and country (color saturation), that are classified as (a) either British or American English (b) American English (c) British English, (d) American English, conditional on being assigned a single language variety label, (e) British English, conditional on being assigned a single language variety label. 549,292 petitions were classified as either British English, American English, Unknown English, or Mix of English varieties. Of these, 171,875 petitions were assigned a single language variety label (British or American English). A petition is labeled as either American English or British English using the `abclf` Python package, which relies on vocabulary and spelling to assign a label. Overall, the share of petitions that receive a label of either American or British English increases with access to the AI tool: this increase is driven by an increase in petitions labeled as American English.

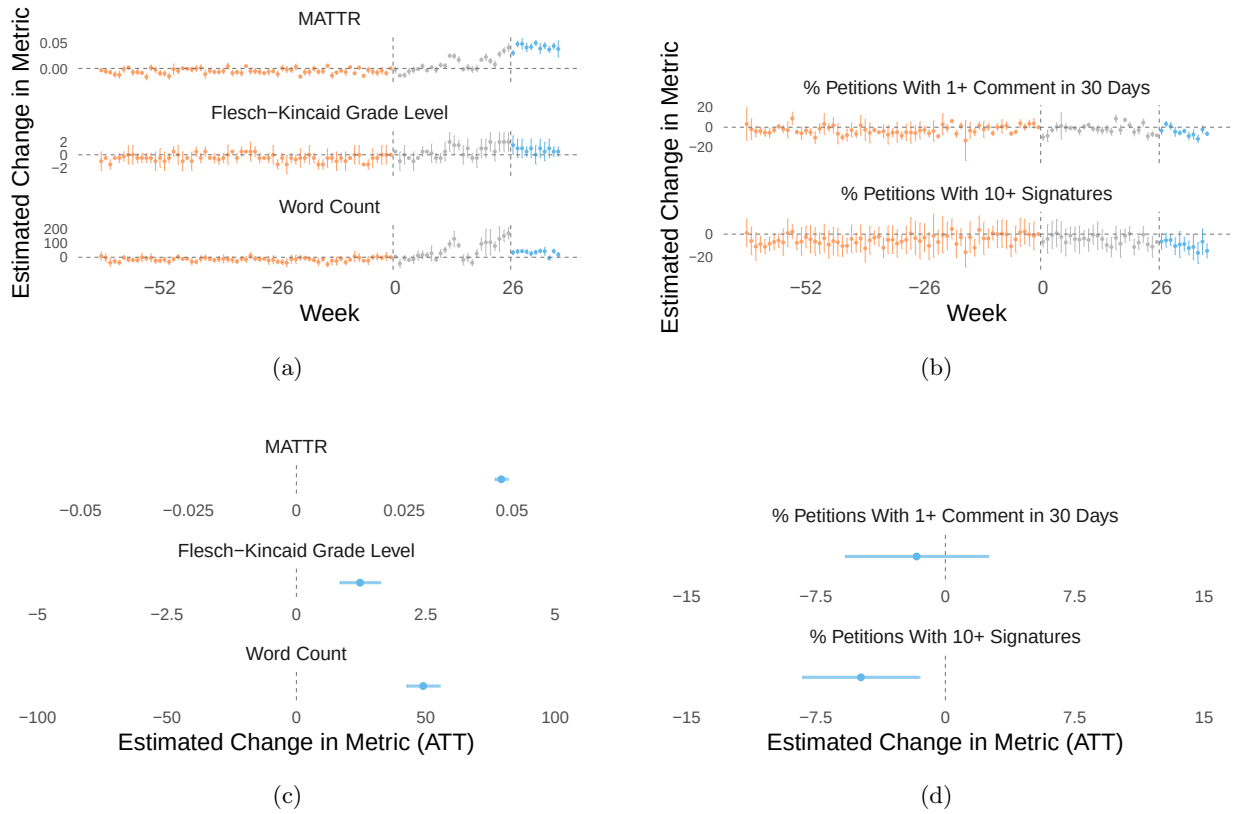


Figure 25: Static and dynamic difference-in-differences estimates without the United States as a treated group. Treated group includes Great Britain and Canada; control group contains Australia. Analysis includes 68,264 pre-AI petitions, 24,401 petitions written during the A/B period, and 12,479 petitions written post-AI. (a) Dynamic difference-in-differences with lexical features, (b) dynamic difference-in-differences with outcomes, (c) static difference-in-differences with lexical features, (d) static difference-in-differences with outcomes. The figure shows 95% CI. Results show significant changes to lexical features and a lack of improved outcomes, which are consistent with difference-in-difference results that include the United States as a treated country. We show point-wise, rather than simultaneous, 95% CI bands to account for weeks lacking sufficient variation to estimate week-group ATT.

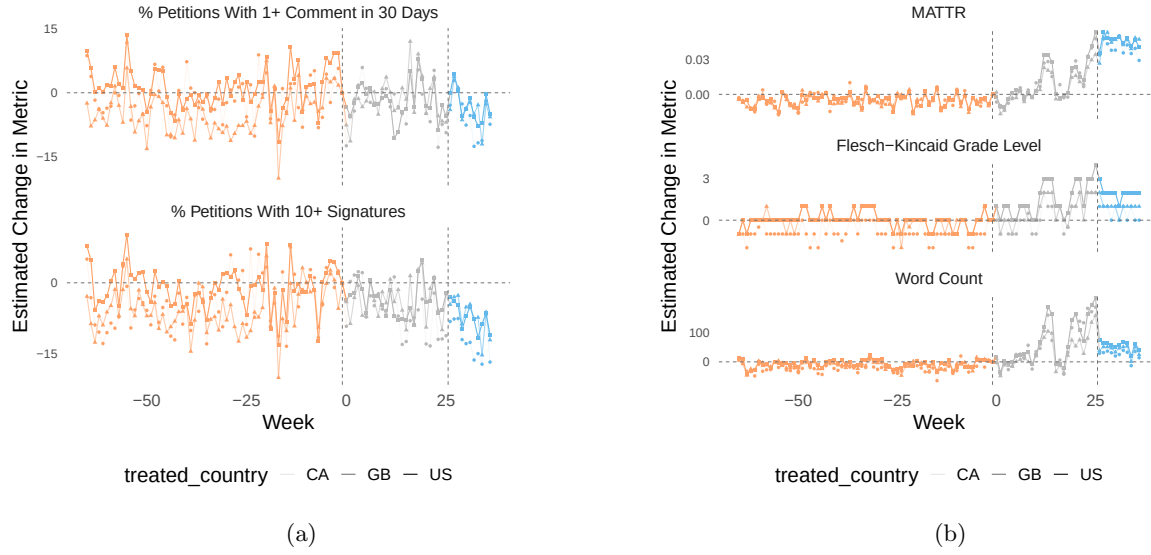


Figure 26: Dynamic difference-in-difference estimates comparing each treated country (United States: $N = 256,491$ petitions; Canada: $N = 25,250$ petitions; Great Britain: $N = 59,651$ petitions) to Australia ($N = 20,243$ petitions). Each point reflects the difference between the treated country's weekly metric value and Australia's weekly metric value; relative to the baseline week. The figure demonstrates that shifts in (a) outcomes and (b) lexical features are consistent across countries.

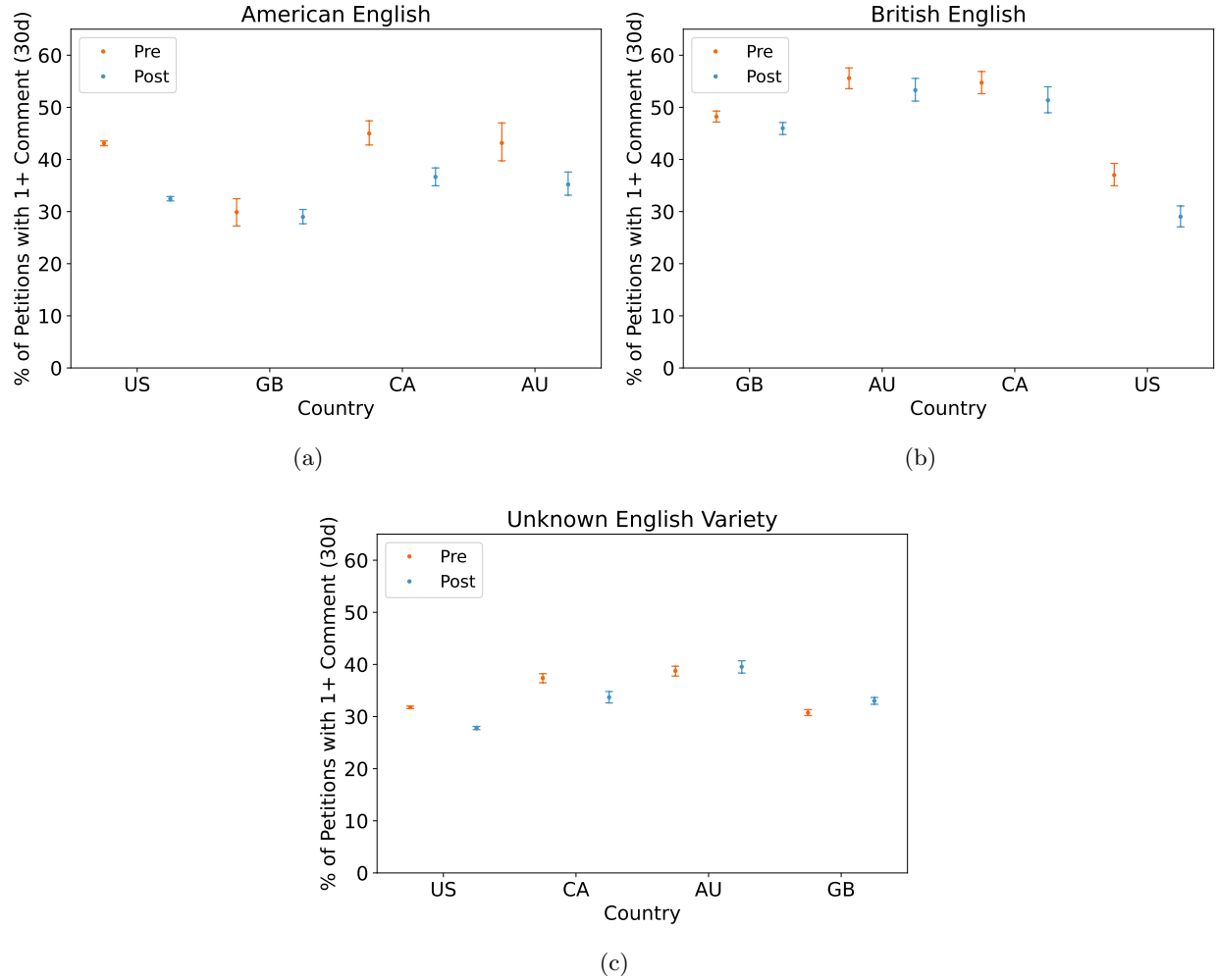


Figure 27: The share of petitions that reach the outcome threshold of 1+ comment in 30 days when written in (a) American English, (b) British English, or (c) Unknown English variety. Data is drawn from January 2022 - December 2024, excluding the period of differential access and A/B testing and includes petitions with five or more words and 2 sentences. During this period, 100,199 petitions were labeled American English, 27,243 petitions were labeled as British English, 4,619 labeled a mix of American and British, and 287,188 labeled as Unknown English variety. We observe that the share of American English variety petitions that reach 1 comment in 30 days decreases post-AI in the United States, Canada, and Australia, with no observable difference in Great Britain. The share of British English petitions that reach 1+ comment in 30 days has the greatest decrease post-AI in the United States, with slight decreases in Canada, Great Britain, and Australia. The share of unknown English variety petitions that reach 1+ comment in 30 days decreases post-AI in the United States, Canada, and Australia, and increases modestly in Great Britain. The figure shows 95% CI.

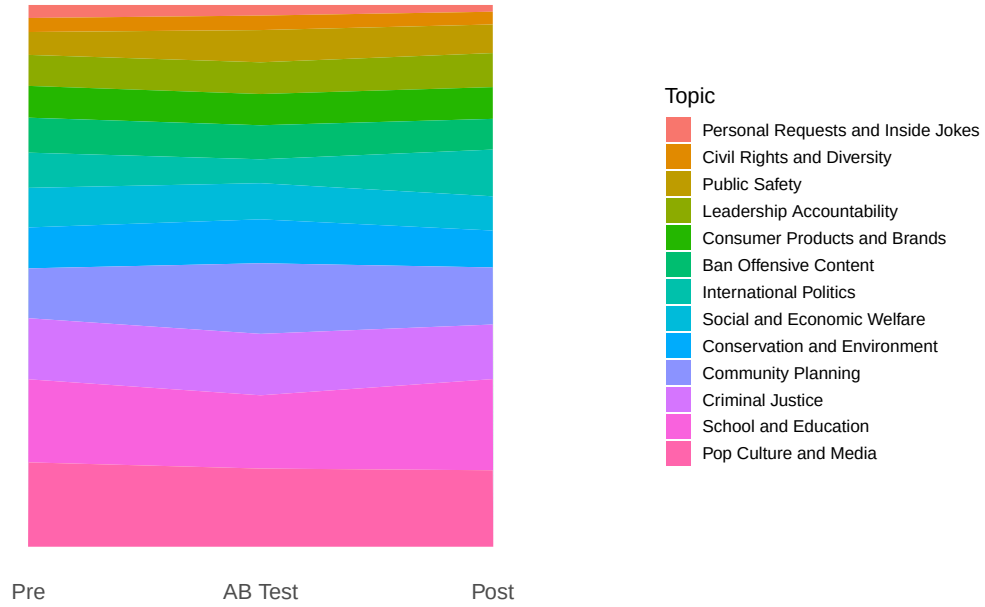


Figure 28: The distribution of petition topics between January 2022 and December 2023 over pre-AI ($N = 237,020$ petitions), A/B period (82,504 petitions), and post-AI period ($N = 42,111$ petitions) in the United States, Great Britain, Australia, and Canada. We observe a consistent distribution of petitions by topic across the three periods. A 13x3 chi-square test comparing the distribution of topics over the three periods (pre-AI, A/B period, post-AI) showed period to have a negligible effect on topic distribution ($\chi^2(24) = 2950.5$, $p < 0.0001$, Cramer's $V=0.064$), suggesting little to no relationship between topic and time period, despite statistical significance (likely due to the large sample size).

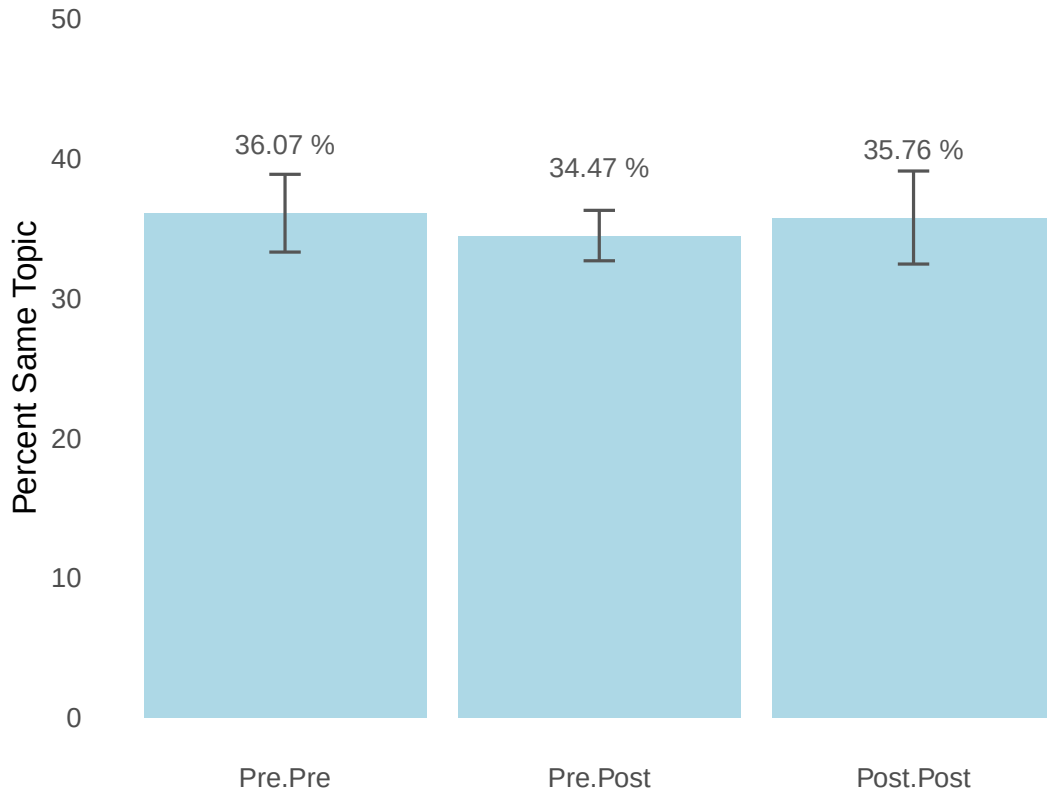


Figure 29: Among repeat petition-writers included in the main analysis (N=4,611 users), between 34-36% of users in the Pre/Pre, Pre/Post, and Post/Post cohorts wrote both their first and second petitions about the same topic. The bars represent the proportion of second petitions that had the same topic as the first for the repeat writers in each cohort, with 95% CI.

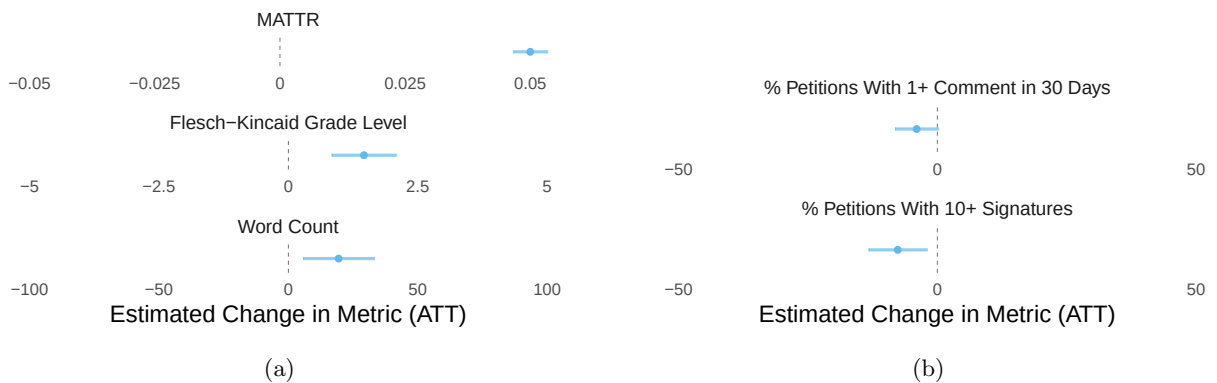


Figure 30: Static difference-in-differences on a subset of petitions (115,284 petitions pre-AI, 43,061 petitions during the A/B period, 21,055 petitions post-AI) about civic engagement topics (Conservation and Environment, Community Planning, International Politics, Social and Economic Welfare, Civil Rights and Diversity, Criminal Justice, Public Safety). Panel (a) shows civic engagement petitions lexical features shift significantly: petitions have greater lexical diversity, readability, and are longer. Panel (b) shows outcomes for civic engagement petitions do not appear to improve, with significant negative ATT estimates for the percent of petitions that reach minimum comment and signature thresholds.

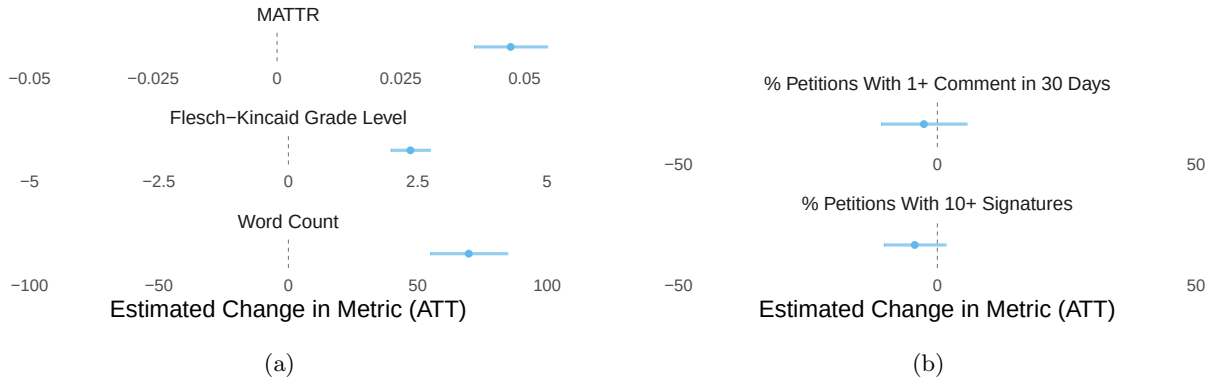


Figure 31: Static difference-in-differences on subset of petitions about “other” topics (121,736 petitions pre-AI, 39,443 petitions during the A/B test period, 21,056 petitions post-AI), which do not pertain to civic engagement (Pop Culture and Media, Consumer Products and Brands, Leadership Accountability, School and Education, Personal Requests and Inside Jokes, Ban Offensive Content). Panel (a) shows lexical features shift significantly: petitions have greater lexical diversity, readability, and are longer. Panel (b) shows outcomes do not appear to improve, with negative ATT estimates for the percent of petitions that reach minimum comment and signature thresholds.

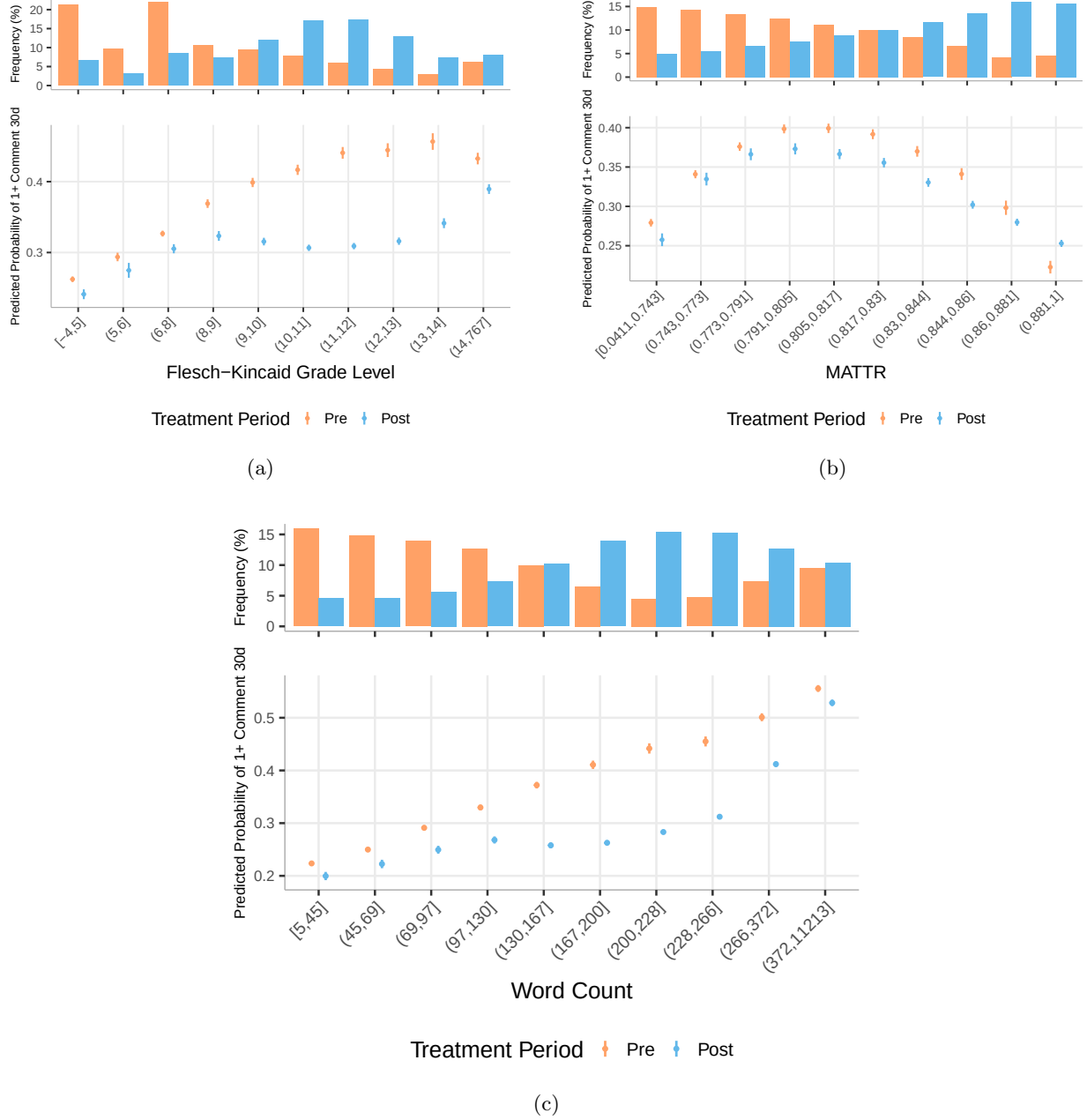
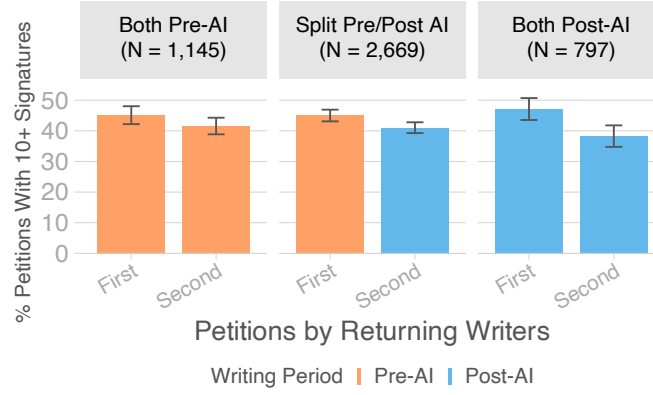
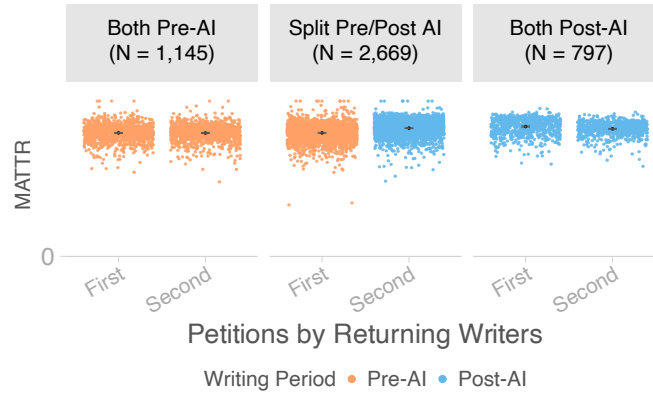


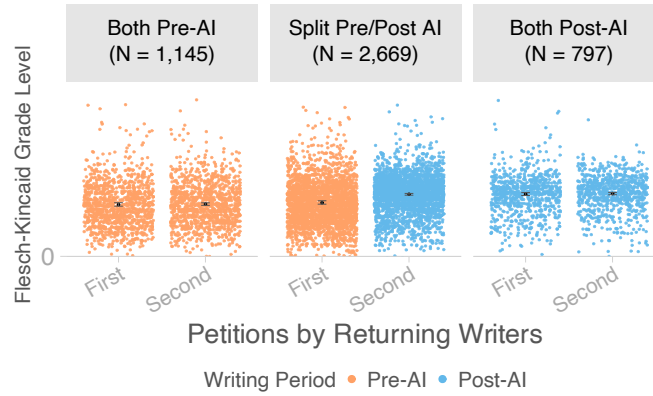
Figure 32: Marginal estimates for lexical features regressed on outcome metric (1+ comment in 30 days), using petitions written between 2022 and 2024 in the US, GB, and CA (257,590 petitions pre-AI, 244,521 petitions post-AI). (a) Flesch-Kincaid grade level, (b) MATTR, and (c) word count are less predictive of petition probability of receiving 1+ comment in 30 days when petitions are written with AI access. With AI access, word count deciles in the 130-372 range are more common, but have significantly lower marginal predicted probability of 1+ comment in 30 days, as predicted by a logistic regression model. The figure shows 95% CIs with HC3 heteroskedasticity-consistent standard errors.



(a) 10+ Signature Count



(b) MATTR



(c) Flesch-Kincaid Score

Figure 33: Average petition outcomes and lexical features for petitions written by repeat-petition-writers ($N = 4,611$ users). (a) shows the share of petitions in each cohort that reach 10+ signatures, (b) shows average MATTR (c) shows average Flesch-Kincaid grade level. Points show petition metric value. Outliers (beyond ± 3 standard deviation) are excluded from the visualization. The figure shows 95% CI, bootstrapped with 1,000 iterations.

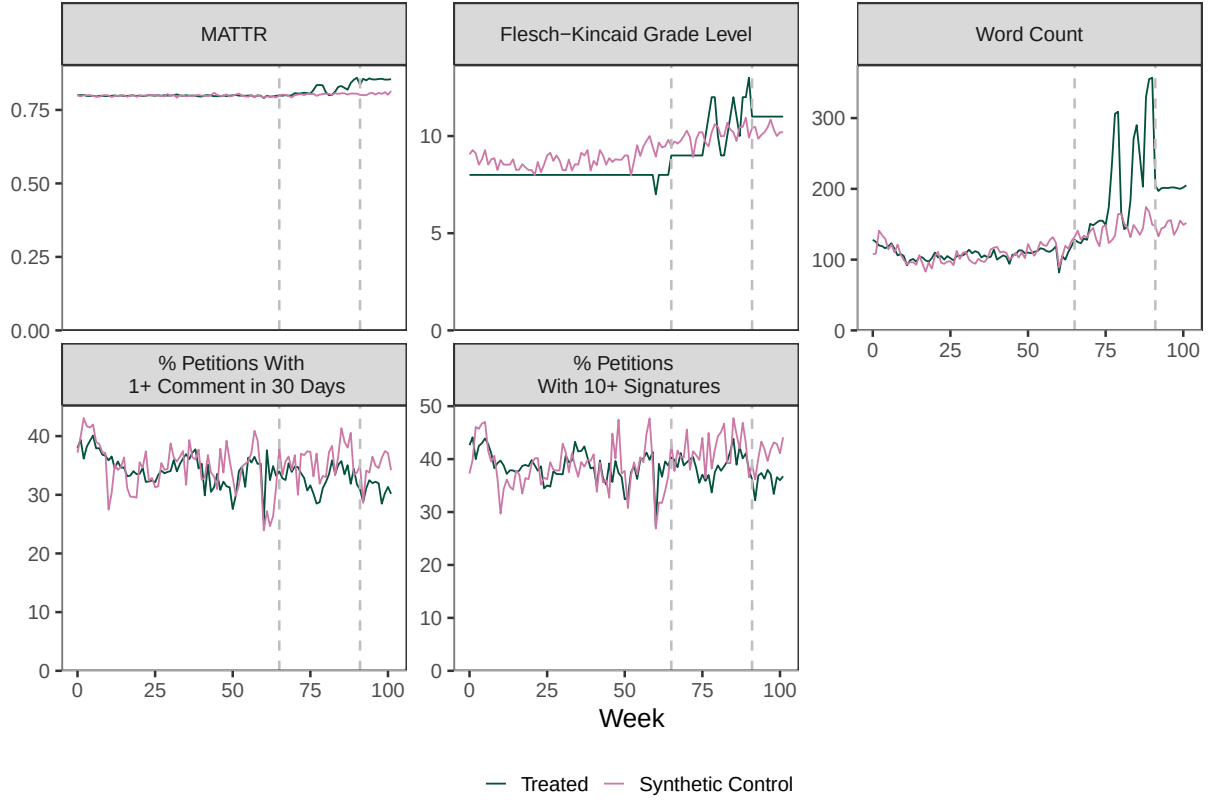


Figure 34: Synthetic control analysis comparing a treated time series against a weighted combination of control time series. This figure compares the United States as a treated unit against a weighted combination of Australia, New Zealand, and India as control units ($N = 301,876$ petitions). The y-axis for each facet is defined in the title of each facet plot. Because New Zealand and India fail to demonstrate parallel trends with our treated countries we exclude them from the difference-in-differences analysis. However, because the synthetic control method creates a balanced control time series to compare with treated units, we can include these countries in the synthetic control analysis. We find that the estimated effects derived from synthetic controls are directionally consistent and similar in scale to the estimates derived from our difference-in-differences analysis.

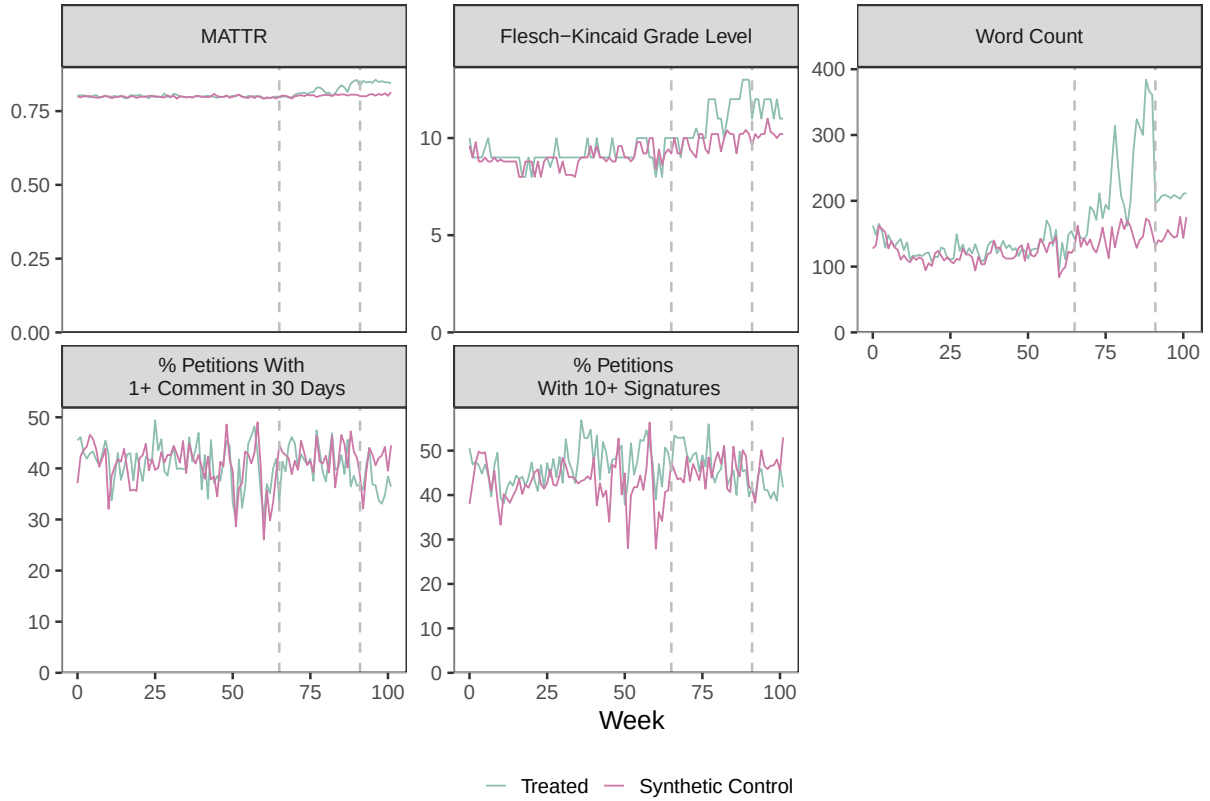


Figure 35: This figure shows a synthetic control analysis comparing Canada as a treated unit against a weighted combination of Australia, New Zealand, and India as control units ($N = 70,635$ petitions). We find that the estimated effects derived from synthetic controls are directionally consistent and similar in scale to the estimates derived from our difference-in-differences analysis.

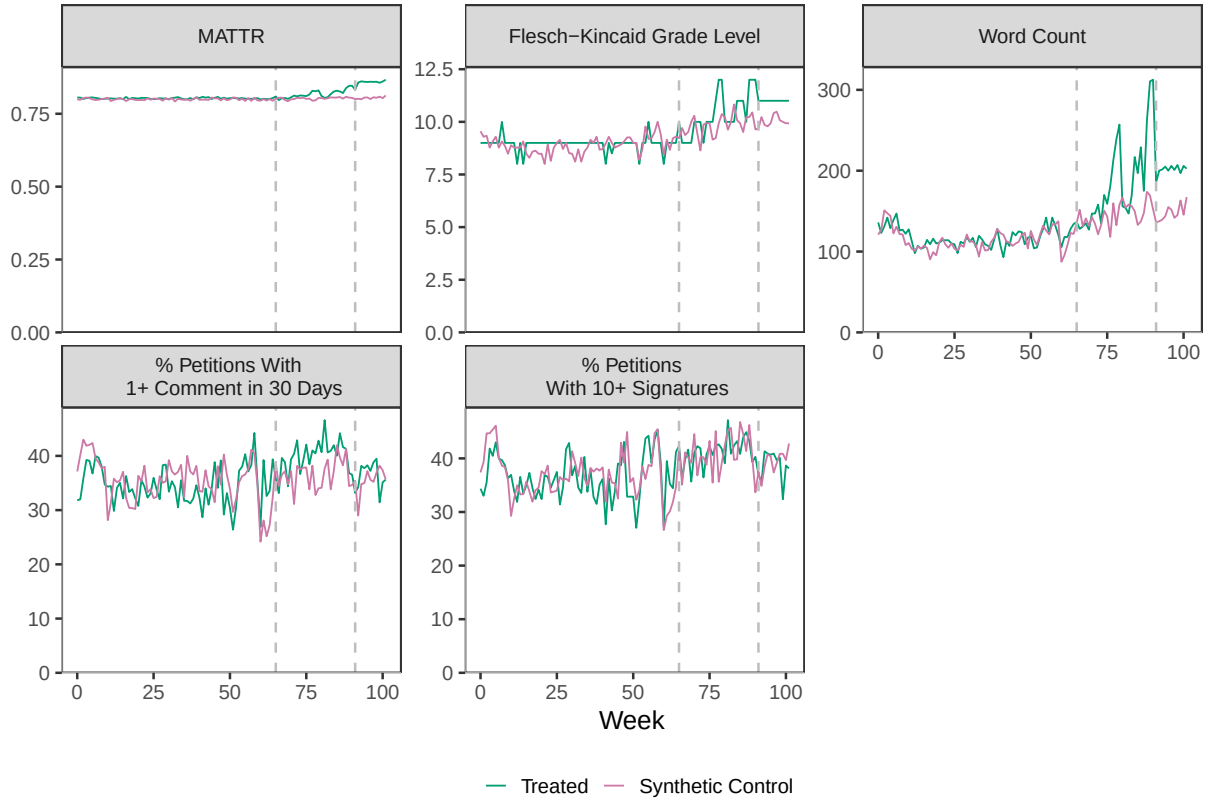


Figure 36: This figure shows a synthetic control analysis comparing Great Britain as a treated unit against a weighted combination of Australia, New Zealand, and India as control units ($N = 105,036$ petitions). We find that the estimated effects derived from synthetic controls are directionally consistent and similar in scale to the estimates derived from our difference-in-differences analysis.

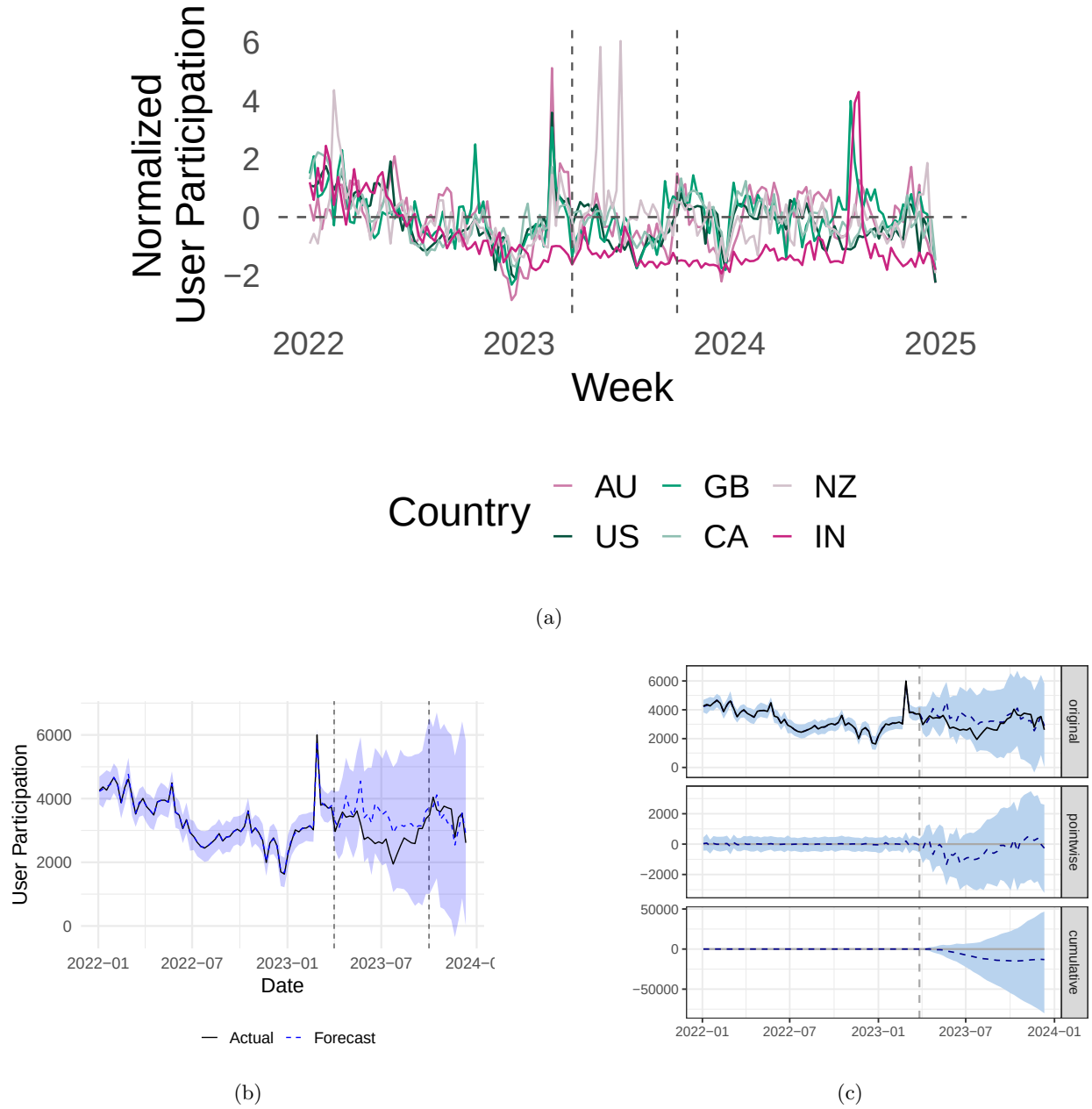


Figure 37: Time series of platform participation, as measured by normalized (by pre-AI distribution) count of petition-writers among countries that received access in October 2023 (US, GB, CA) and countries that had known delayed release dates (AU, NZ, IN). NZ and IN are excluded from main analyses due to their lack of parallel trends with treated countries. Panel (a) shows normalized counts of petition-writers by country (AU, IN, NZ, US, GB, CA) through the full data collection period ($N = 497,426$ users). Vertical lines reflect the start of the A/B testing began in the US, GB, and CA and the start of full access in the US, GB, and CA. Panels (b) and (c) show our Bayesian Structural Time Series (BSTS) analysis, which we train on pre-AI data to predict platform participation through the period of differential AI access (December 2023), using counts of petition-writers ($N = 356,745$ users). In the BSTS analysis control countries are AU, IN, NZ and treated countries are the US, GB, and CA. This analysis creates a synthetic counterfactual for platform participation in the absence of an intervention. Panel (b) shows our forecast vs the actual user participation. Panel (c) shows 95% credible intervals for the estimated difference between the synthetic counterfactual and actual user participation. As bands include 0 throughout post-intervention period, we do not find evidence of participation changing due to the shock of in-platform AI access.

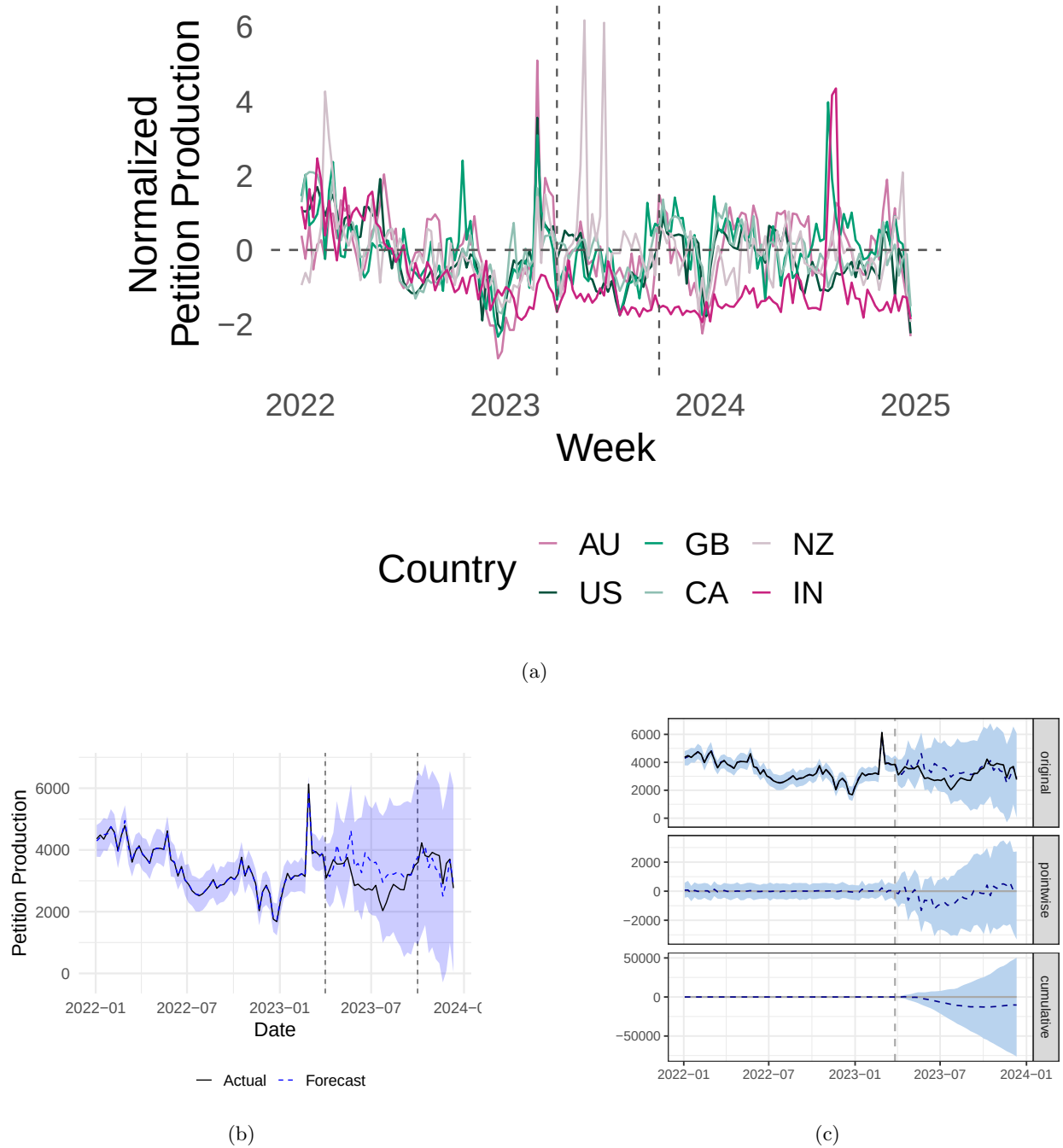
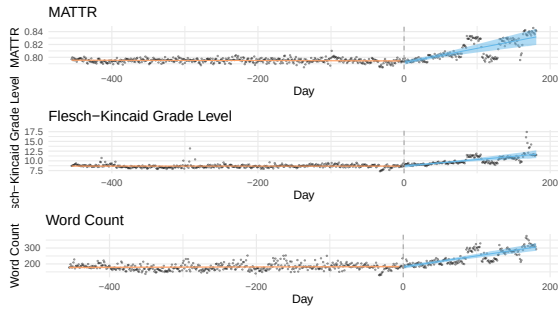


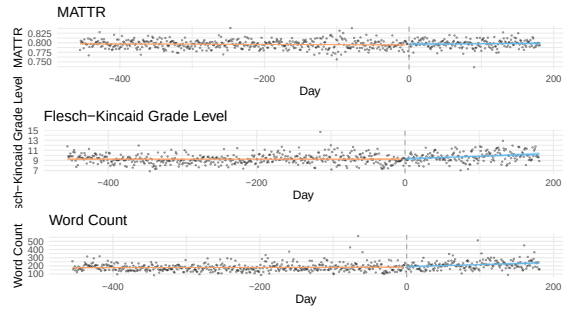
Figure 38: Time series of normalized (by pre-AI distribution) count of published petitions. NZ and IN are excluded from main analyses due to their lack of parallel trends with treated countries. Panel (a) shows normalized petition counts by country (AU, IN, NZ, US, GB, CA) across the full period of data collection ($N = 548,656$ petitions). Panels (b) and (c) show our Bayesian Structural Time Series (BSTS) analysis, which we train on pre-AI data to predict petition counts, using published petition counts in that period ($N = 386,777$ petitions). In the BSTS analysis control countries are AU, IN, NZ and treated countries are the US, GB, and CA. This analysis creates a synthetic counterfactual for petition counts in the absence of an intervention. Panel (b) shows our forecast vs the actual petition count. Panel (c) shows 95% credible intervals for the estimated difference between the synthetic counterfactual and actual petition count. As bands include 0 throughout post-intervention period, we do not find evidence of petition counts changing due to the shock of AI access.

Table 1: Balance table of pre/post covariates for main difference-in-differences, thresholded to the first week of each period, for consistency with difference-in-difference time periods. Pre-AI period is from January 3, 2022 to April 2, 2023. Post-AI period is from October 2, 2023 to December 15, 2023. Difference in means is result of difference in means t-test. Significance reported at $\alpha = 0.05$.

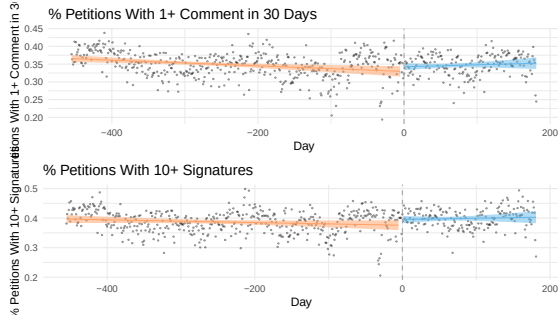
	Pre (N=237020)		Post (N=42111)		Diff. in Means	Std. Error
	Mean	Std. Dev.	Mean	Std. Dev.		
Has Previous Comments	0.04	0.2	0.074	0.26	0.034***	0.0013
Has Decision Makers	0.21	0.41	0.11	0.31	-0.11***	0.0017
New User	0.93	0.26	0.88	0.32	-0.044***	0.0017
Government Topic	0.37	0.48	0.4	0.49	0.039***	0.0026



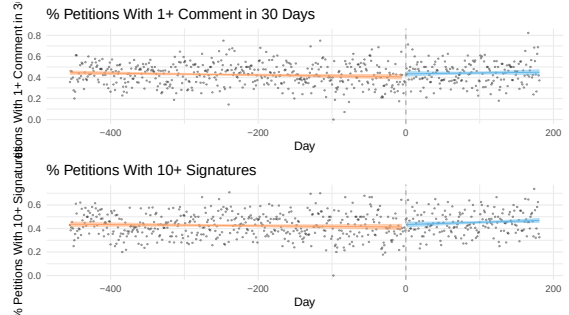
(a) Treated Lexical Features



(b) Control Lexical Features



(c) Treated Outcome



(d) Control Outcome

Figure 39: Interrupted time series (ITS) showing regression discontinuity in time fitted models, where the date of intervention aligns with the start of the A/B testing period April 3, 2023; the start of A/B testing period in treated countries (N=237,020 petitions posted pre-intervention, N=82,504 petitions posted post-intervention, during the A/B test period). Here, we measure if there is a substantial change in the slope of regression lines fit to the pre-A/B testing and post-A/B testing periods. Panel (a) shows lexical features (top to bottom: MATTR, Flesch-Kincaid grade level, word count) in treated countries (United States, Great Britain, Canada). Panel (b) shows lexical features in control country (Australia). We observe that while there is a marked increase in the slope of the regression line fit during the A/B testing period for treated countries, and less so in the control countries. Panel (c) shows outcomes (top to bottom: % of petitions that reach 1 comment (30d), % of petitions that reach 10 signatures) in treated countries. (d) Shows outcomes in control countries (Australia). We observe little difference in the change in outcomes. In both conditions, the slope of outcomes is slightly increased in the A/B testing period. Panels show 95% CI with Newey-West standard errors.

Table 2: Pre-treatment (January 3, 2022 - April 2, 2023) descriptive statistics of main text metrics

Variable	Unique	MissingPct	Mean	SD	Min	Median	Max
mattr_title_text	126784	0	0.79	0.06	0.04	0.8	1
flesch_kincaid_grade_level	134	0	8.65	5.57	-4.00	8.0	767
word_count_title_text	2262	0	179.12	270.09	6.00	110.0	10778
has_comment_30d_1plus	2	0	0.35	0.48	0.00	0.0	1
has_signature_count_10plus	2	0	0.39	0.49	0.00	0.0	1

Table 3: Post-treatment (October 2, 2023 - December 15, 2023) descriptive statistics of main text metrics

Variable	Unique	MissingPct	Mean	SD	Min	Median	Max
mattr_title_text	36360	0	0.84	0.06	0.04	0.85	1
flesch_kincaid_grade_level	63	0	10.69	4.69	-4.00	11.00	257
word_count_title_text	1313	0	228.67	225.36	7.00	201.00	10571
has_comment_30d_1plus	2	0	0.33	0.47	0.00	0.00	1
has_signature_count_10plus	2	0	0.37	0.48	0.00	0.00	1

Category	Item (1–5 Likert scale)
Writing Quality	The text is well-structured and easy to follow.
	The paragraphs and sentences are logically organized.
	The language is clear and easy to understand.
	The text is free of errors in grammar, punctuation, and spelling.
Persuasiveness	The petition provides reasons for its cause that are believable.
	The petition provides reasons for its cause that are convincing.
	This petition has the potential to change reader behavior.
	This petition has the potential to influence decision-makers.

Table 4: Labeling task measures to evaluate petition writing quality and persuasiveness. Each item is assessed in a 1-5 Likert scale from 1 = strongly disagree to 5 = strongly agree.

Table 5: GLM estimates (exp. coef.): Logistic regression model regressing outcome (has 1+ comment in 30 days) on persuasiveness score and AI period

term	estimate	std.error	statistic	p.value
(Intercept)	0.043	0.621	-5.047	0.000
persuasive_avg	2.405	0.227	3.859	0.000
treatment_periodpost_ai	3.239	0.956	1.229	0.219
persuasive_avg:treatment_periodpost_ai	0.495	0.317	-2.216	0.027

Table 6: GLM estimates (exp. coef.): Logistic regression model regressing comment outcome (1+ comment in 30days) on writing quality and AI period

term	estimate	std.error	statistic	p.value
(Intercept)	0.129	0.622	-3.288	0.001
writing_quality_avg	1.516	0.214	1.947	0.052
treatment_periodpost_ai	2.934	1.023	1.052	0.293
writing_quality_avg:treatment_periodpost_ai	0.580	0.317	-1.719	0.086

Table 7: GLM estimates (exp. coef.): Logistic regression model regressing outcome (has 10+ signatures) on persuasiveness and AI period

term	estimate	std.error	statistic	p.value
(Intercept)	0.060	0.574	-4.897	0.000
persuasive_avg	2.829	0.222	4.691	0.000
treatment_periodpost_ai	2.382	0.855	1.015	0.310
persuasive_avg:treatment_periodpost_ai	0.527	0.293	-2.183	0.029

Table 8: GLM estimates (exp. coef.): Logistic regression model regressing outcome (has 10+ signatures) on writing quality and AI period

term	estimate	std.error	statistic	p.value
(Intercept)	0.146	0.577	-3.335	0.001
writing_quality_avg	1.855	0.202	3.057	0.002
treatment_periodpost_ai	1.737	0.924	0.598	0.550
writing_quality_avg:treatment_periodpost_ai	0.659	0.287	-1.452	0.146

Table 9: Static difference-in-differences estimate for outcome metric reported in main text: percentage of petitions with 1+ comment in 30 days. 95% CIs are shown.

VARIABLES	(1)				
	% Petitions With 1+ Comment in 30 Days				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI	-2.508	1.826	[-6.360, 1.343]	-1.374	0.187
Observations	304				

Table 10: Static difference-in-differences estimate for outcome metrics reported in main text: percentage of petitions with 10+ signatures in 30 days. 95% CIs are shown.

VARIABLES	(1)				
	% Petitions With 10+ Signatures				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI	-5.329	1.363	[-8.205, -2.454]	-3.911	0.001
Observations	304				

Table 11: Static difference-in-differences estimate for lexical feature metric MATTR, as reported in main text. 95% CIs are shown.

VARIABLES	(1)				
	MATTR				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI	0.049	0.001	[0.047, 0.050]	61.482	0.000
Observations	304				

Table 12: Static difference-in-differences estimate for lexical feature metric word count, as reported in main text. 95% CIs are shown.

VARIABLES	(1)				
	Word Count				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI	53.628	3.105	[47.077, 60.178]	17.273	0.000
Observations	304				

Table 13: Static difference-in-differences estimate for lexical feature metric Flesch-Kincaid grade level, as reported in main text. 95% CIs are shown.

VARIABLES	(1)				
	Flesch-Kincaid Grade Level				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI	1.487	0.162	[1.144, 1.830]	9.157	0.000
Observations	304				

Table 14: Static difference-in-differences for mean pairwise similarity, as reported in the main text. 95% CIs are shown.

VARIABLES	(1)				
	Mean Pairwise Similarity				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI	0.056	0.003	[0.049, 0.063]	17.825	0.000
Observations	304				

Table 15: Static difference-in-differences estimate for outcome metric: percent of petitions with 1+ comment in 30 days, with standard errors clustered by country. 95% CIs are shown.

VARIABLES	(1)				
	% Petitions With 1+ Comment in 30 Days				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI	-2.508	1.822	[-8.308, 3.291]	-1.376	0.262
Observations	304				

Table 16: Static difference-in-differences estimate for outcome metric: percent of petitions with 10+ signatures, with standard errors clustered by country. 95% CIs are shown.

VARIABLES	(1)				
	% Petitions With 10+ Signatures				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI	-5.329	1.973	[-11.609, 0.950]	-2.701	0.074
Observations	304				

Table 17: Static difference-in-differences estimate for lexical feature metric MATTR, with standard errors clustered by country. 95% CIs are shown.

VARIABLES	(1)				
	MATTR				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI Observations	0.049 304	0.002	[0.042, 0.055]	23.150	0.000

Table 18: Static difference-in-differences estimate for lexical feature metric word count, with standard errors clustered by country. 95% CIs are shown.

VARIABLES	(1)				
	Word Count				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI Observations	53.628 304	4.584	[39.040, 68.215]	11.699	0.001

Table 19: Static difference-in-differences estimate for lexical feature metric Flesch-Kincaid Grade Level, with standard errors clustered by country. 95% CIs are shown.

VARIABLES	(1)				
	Flesch-Kincaid Grade Level				
	Estimate	Std. Error	CIs	t	p
Treated Country x Post-AI Observations	1.487 304	0.064	[1.284, 1.690]	23.299	0.000

Table 20: Mixed effects generalized logistic regression model predicting outcome metrics. Metric is a binary value: petitions either meet this threshold or fail to meet the threshold. We include the interaction between petition order and repeat-petition writer cohort, to understand the predictive relationship between user cohort and the outcome metric, relative to the user’s experience and past performance.

	1 + Comment (30d)	10+ Signatures
Intercept	-0.403 (0.082)	-0.291 (0.085)
Second Petition	-0.193 (0.101)	-0.212 (0.100)
Cohort: Pre/Post	-0.136 (0.098)	-0.004 (0.101)
Cohort: Post/Post	0.002 (0.128)	0.123 (0.132)
Interaction: Second Petition $\times$ Cohort (Pre/Post)	-0.090 (0.121)	-0.033 (0.120)
Interaction: Second Petition $\times$ Cohort (Post/Post)	-0.220 (0.158)	-0.328 (0.157)
Num.Obs.	9222	9222
R2 Marg.	0.005	0.004
R2 Cond.	0.358	0.418
AIC	11978.9	12126.9
BIC	12028.8	12176.8
ICC	0.4	0.4
RMSE	0.36	0.34

Table 21: Mixed effects linear regression model for lexical style metrics. While word count and Flesch-Kincaid scores are log transformed, raw MATTR does not show skew and is therefore not log transformed. We include the interaction between petition order and repeat-petition writer cohort, to understand the predictive relationship between user cohort and the lexical feature metric, relative to the user’s experience and past performance. Note 72 observations are removed from the FKGL model due to missingness (negative FKGL, invalid for log transformation).

	MATTR	Log(Flesch-Kincaid Grade Level)	Log(Word Count)
Intercept	0.795 (0.002)	2.096 (0.013)	4.890 (0.025)
Second Petition	-0.001 (0.002)	0.021 (0.015)	0.059 (0.025)
Cohort: Pre/Post	-0.000 (0.002)	0.027 (0.016)	0.031 (0.029)
Cohort: Post/Post	0.041 (0.002)	0.222 (0.020)	0.329 (0.038)
Interaction: Second Petition $\times$ Cohort (Pre/Post)	0.031 (0.002)	0.176 (0.017)	0.273 (0.030)
Interaction: Second Petition $\times$ Cohort (Post/Post)	-0.014 (0.003)	-0.001 (0.023)	0.005 (0.039)
Num.Obs.	9222	9150	9222
R2 Marg.	0.087	0.052	0.038
R2 Cond.	0.186	0.410	0.508
AIC	-28101.9	10322.3	115177.6
BIC	-28044.8	10379.3	115234.6
ICC	0.1	0.4	0.5
RMSE	0.05	0.30	0.49

Change.org and the Write with AI Tool

Change.org launched their in-platform AI tool to select countries, including the United States, Canada, Great Britain, in October 2023. Change.org then launched the AI tool to Australia in late December 2023. We received confirmation of the delayed release in Australia through personal communications with two Change.org employees. The delay was due to internal logistics, and was not an intentional experiment (as far as the authors are aware). While a staggered roll-out is typical practice for tech companies, including Change.org, this delay may also be attributable to the new and quickly-shifting landscape of generative AI legislation. At the time, users learned about the existence of the AI tool only upon entering the AI-assisted writing workflow after starting a petition [1]. During the staggered roll out period that we study in the main text difference-in-differences analysis (October-December 2023), the home page did not advertise the existence of an AI tool. Furthermore, users routed directly to petition pages did not see any information about whether or not the petition had been written with the support of the in-platform AI tool. However, by 2025, Change.org had updated the platform home page to advertise their AI tool [2]. On one Change.org home page, visible on February 14, 2025, there was a text box that advertised the AI tool as providing help for users’ productivity and the quality of their petitions [2]. The home page read, “Create a compelling petition in minutes!” It is possible that this messaging could increase AI suspicion for petition readers who visit the home page before reading petitions. Furthermore, this message suggests that the platform’s intent in releasing the AI tool was to enable petition-writers to create high quality petitions with reduced effort. By lowering the barrier to petition-creation, we might speculate that the platform sought to increase the volume of petitions produced on the platform or increase petition completion rates. In 2026, Change.org launched an AI-first home page, in which visitors to the home page could only start a petition with the AI tool [3]. While it is unclear how broadly this home page design was rolled out, it is clear that the platform has increased their commitment to supporting petition writing with AI.

We note that we conducted our study without formal data-sharing collaboration from Change.org. Without access to internal data, our focus is limited to platform-level trends rather than the impact of AI on individual petitions or petition-writers. For example, while we suspect that the AI tool might have increased petition completion rates, we are unable to study this with publicly available data. Prior work has discussed the challenges of studying online platforms without internal data access [4, 5]. We use this opportunity to echo previous calls for open collaboration between platforms and researchers. With appropriate transparency [6], such collaborations could improve public understanding of the trade-offs, harms, and opportunities of in-platform AI-integration.

Data

We collected 1,517,433 petitions for our dataset, and used a subset of 361,635 for the difference-in-differences analysis in the main text. We constructed this subset by including only English language petitions that were posted by users in the United States, Canada, Great Britain, and Australia. We derive user language by selecting petitions that have an English language tag on the petition’s web page and for which the `nltk` library recognized at least 50% of the words as English [7]. We also exclude petitions that have fewer than 5 words or 2 sentences. In Table 1, we show the distribution of petition and user attributes for this difference-in-differences data subset before and after the release of AI feature. We also show the distribution of our primary outcome and lexical feature metrics pre-AI access in Table 2, and the distribution of these metrics post-AI access in Table 3. We also show the distribution of density of our outcome metrics, comments and signatures, on a log scale in SI Appendix Figure 1. Note that references to treatment refer to the full launch of the in-platform AI tool. We define the AI periods using the first complete week after each threshold date (January 1, 2022 for the start of pre-AI period; April 1, 2023 for start of A/B test period; October 1, 2023 for the start of the post-AI period). We use the week-level dates for consistency across analyses with our main difference-in-difference analyses, which operate using week-level time series. We mapped dates to weeks through the `pandas` Python package, version 2.2.2. We use the default week mapping, by which `pandas` anchors weeks to end on Sunday (so a week runs Monday through Sunday). With the weekly aggregation, our pre-AI period includes 65 weeks, from the week of January 3, 2022 to the week of March 27, 2023 ($N = 237,020$ petitions); the A/B period includes 26 weeks, from the week of April 3, 2023 to the week of September 25, 2023 ($N = 82,504$ petitions); and the post-AI period of differential access includes 11 weeks, from the week of October 2, 2023 to the week of December 11, 2023 ($N = 42,111$ petitions).

Preregistration

We separately preregistered our main causal analysis and our supplemental petition labeling task. The preregistration for the causal analysis is at: <https://aspredicted.org/xtcd-tvqk.pdf>. The preregistration for the labeling task is at: <https://aspredicted.org/z6g8xe.pdf>.

We did not deviate from the preregistration for our labeling task. We note three deviations from the preregistration for the causal analysis. First, in our preregistration we defined English-language petitions as those with a “25% match to the NLTK English corpus”. However, upon inspecting our data, we found that this threshold was far too low, with many non-English petitions passing the threshold. We then increased the minimum threshold for English language detection to 50%, which is the threshold we used for all analyses in the main and SI text. To allow for consistent interpretation of lexical features, it would have been inappropriate to include non-English petitions despite their happening to fall above the threshold of 25%. Second, we defined our exclusion criteria as “anomalies, which we define as petitions without any alphanumeric characters and those with a negative number of signatures.” However, in our main analysis we exclude petitions with fewer than 5 words or 2 sentences. We removed these short petitions because they are inappropriate for our analyses of lexical features or text homogeneity, which require at least 2 sentences for reasonable interpretation. Furthermore, these short petitions are often meaningless. For example, Change.org does not allow users to delete their petitions directly, and as a result, some users simply edit their petition text to contain a single word or punctuation in place of full removal. We replicated our main analysis and observe consistent results when we use the initial exclusion criteria, which includes petitions with any word count or sentence count (SI Appendix Figure 15). Third, our preregistration also noted that we would use inverse propensity weighting. Our main analysis does not use weighting due to a lack of reliable pre-treatment covariates (described further under “Difference-in-Differences Robustness Checks” below). We include a doubly robust difference-in-differences model, which includes inverse propensity weighting, in SI Appendix Figure 16. This analysis supports our main finding that access to in-platform AI changes lexical features but does not improve platform-level petition outcomes.

AI Detection

We created a Change.org AI tool classifier, and applied it to the sample of English petitions used in the main analysis ($N = 361,635$). We assigned this label with an ensemble method that combined an open-source AI-detection model’s predictions with two heuristics for in-platform AI usage [8]. We identified two key heuristics for the Change.org AI in-platform tool. The first heuristic was word count, as petitions created with the AI feature tended to have at least 100 words. The second heuristic was the presence of even text breaks, as petitions written with the AI feature tended to have multiple paragraphs that were separated by an even number of text break html tags. We then used an open-source AI detection model created by Desklib and trained for general use AI detection in education contexts [8]. As follows, we created three labels: word count over 100 words, an even number of text breaks (i.e., 2,4,6,8,10), and a Desklib AI prediction probability over 0.9. We then compiled these labels through a decision tree classifier predicting if text was produced with the in-platform AI feature or not.

In order to train and test the ensemble classifier we generated 1,125 petitions using the Change.org in-platform AI feature. To generate these petitions, we prompted the Change.org AI tool to produce petitions using a random sample of petitions that had been posted in 2022, before general-purpose AI tools were released. We did not post any of the petitions that were generated, we only saved the text for use in AI detection. We then created a new random sample of 1,125 petitions from the pre-AI period to use for training. With this set of 2,250 annotated petitions, we used a random sample of 70% ($n = 1,576$) of the annotated petitions for training and the remaining 30% ($n = 674$) for testing. We did not train the model using any of the human-written petitions that we prompted the AI tool with. Our ensemble classifier achieved 99% accuracy, 99% precision, and 100% recall on the test set of 674 annotated human and AI petitions. SI Appendix Figure 2 shows the percentage of petitions that we classify as AI-generated per week, in English-market countries with access to the AI feature (US, CA, GB) and without access to the feature (AU). Prior to Australian users receiving access, during the week of December 11, 67.9% of petitions written in countries with access to the AI feature were labeled as AI-generated, while only 27.7% of petitions created in Australia were labeled as AI-generated. We can also assess false positive rates through AI detection scores prior to the release of general purpose off-platform AI tools: in early 2022 (pre-ChatGPT release in November 2022), our

AI detection ensemble method reported 8.8% of petitions across countries as AI generated. Prior to April 1, 2023, we see an uptick to 11.0% of petitions flagged as AI generated. As such, we see that our approach identifies at least some off-platform AI assisted petitions, which explains the 27.7% AI usage rate that we estimate in Australia during the period for which users did not have access to the in-platform tool. This figure demonstrates high utilization of the AI tool feature; just weeks after its full release most petitions with access to the feature were AI-generated.

Petition Writing Quality and Persuasiveness

To better understand petition quality before and after the introduction of the AI tool, we labeled petition writing quality and persuasiveness in a sample of petitions across the two time periods. Our additional data collection shows that petitions written with access to in-platform AI have higher writing quality and persuasiveness, on average, than petitions written without access to AI. Furthermore, writing quality and persuasiveness, while positively predictive of petition outcomes in the pre-AI period, are weaker, or negative, predictors of petition outcomes in the post-AI period.

We study petition writing quality and persuasiveness through a labeling task in which human labelers assigned writing quality and persuasiveness scores to a random sample of petitions. We recruited 14 Cornell Tech graduate students to participate in the study as labelers. We randomly sampled 432 English language petitions from our dataset, such that 216 were from the pre-AI period and 216 were from the post-AI period. Post-hoc we applied our AI detection method to predict if our sampled petitions written with AI or not. We show in SI Appendix Figure 3 that the ensemble detection model labeled 9.4% of pre-AI petitions as AI-written and 53.3% of post-AI petitions as AI written. We selected a sample size of 432 to achieve power=0.8, $\alpha = 0.05$ to detect an estimated moderate effect size of 0.3. We adjusted our sample size calculation to account for the three ratings per petition under a conservative value of $ICC = 0.6$, and $ICC_{k=3} = 0.81$. We sampled from petitions with fewer than 5,000 characters, which accounted for more than 98% of petitions, to account for Qualtrics word limits. To maintain consistency with petitions included in our main analysis, we also excluded petitions with fewer than five words or two sentences. We then removed 15 petitions from the sample of 432 that used offensive language or included personal identifying information of petition authors (e.g., email address, online handles). This left us with a sample of 417 petitions (203 pre-AI, 214 post-AI). We assigned each rater a random set of 80-90 petitions written both pre- and post-AI. The Cornell Institutional Review Board (Cornell University’s FWA is FWA00004513) deemed the labeling task exempt from formal ethics review. Human labelers provided informed consent for the task, and were compensated \$20.00 per hour (\$60.00 for a three-hour task).

We asked participants to rate petitions for writing quality and persuasiveness. We adapted the measures for writing quality from the measures of email writing quality used by Noy and Zhang [9]. We share these measures in Table 4. We drew from prior work on message credibility to inform our measures for persuasiveness [10]. We adapt them for the purposes of the annotation task, in which our labelers were unfamiliar with the context or communities of sampled petitions and thus could not assess petition authenticity. We ask about two specific forms of influence that a petition can exert: influencing readers and decision-makers. We then take the average of the writing quality measures as a writing quality score and the average of the persuasiveness measures as the persuasiveness score.

We find that on average, petition writing quality and persuasiveness is significantly higher in petitions written post-AI, i.e. those who had in-platform AI access, compared to petitions written before AI access. SI Appendix Figure 4 shows the distribution of writing quality and persuasiveness scores among petitions written pre-AI and post-AI. The average writing quality in the pre-AI petitions was 2.73, while the average writing quality in the post-AI petitions was 3.43. A Welch two-sample t-test rejects the null hypothesis of no difference in writing quality between the periods, with a large effect size ($\mu_{post-pre} = 0.70$, $t(412.46) = 9.76$, $p < 0.001$, Cohen’s $d = 0.96$, 95% CI=[0.75, 1.16]). The average persuasiveness score for pre-AI petitions was 2.46, while the average persuasiveness score post-AI was 3.12. A Welch two-sample t-test rejects the null hypothesis of no difference in persuasiveness scores between periods $\mu_{post-pre} = 0.66$, $t(414.56) = 8.75$, $p < 0.001$, Cohen’s $d = 0.86$, 95% CI=[0.65, 1.06]). In sum, we find that petitions written with access to the AI feature have, on average, higher writing quality and persuasiveness scores.

Our analysis also shows that prior to in-platform AI, higher petition writing quality and persuasiveness scores were positively associated with petition outcomes. Post-AI, when predicting likelihood of receiving 1+

comment in 30 days, the predictive strength of text persuasiveness weakened, and the predictive relationship between writing quality and petition outcomes reversed. SI Appendix Figure 5a demonstrates that the predictive relationship between writing quality and outcome (1+ comment in 30 days) observed pre-AI is weaker when petitions are written post-AI. We fit a logistic regression model to predict the likelihood of a petition receiving at least 1 comment in 30 days, regressed on the full interaction of average petition writing quality and treatment period. We show model coefficients in Table 6. We observe a similar significant interaction effect between treatment period and persuasiveness score, as shown in Table 5. SI Appendix Figure 5b shows that while pre-AI, higher persuasiveness scores were predictive of higher probability of a positive outcome (1+ comment in 30 days), post-AI, persuasiveness was a weaker predictor of outcome. We observe consistent results for our other main outcome metric (10+ signatures), as shown in SI Appendix Figure 6, with model coefficients in Tables 7 and 8. With in-platform AI, the predictive relationship between text quality (i.e., writing quality, persuasiveness) and outcomes weakened or reversed: the interaction terms indicate that higher writing quality and persuasiveness no longer increase, and in the case of writing quality, may decrease the probability of receiving a comment. Writing quality and persuasiveness cease to be strong signals of petition outcomes; instead they become weakly, or negatively, associated with outcomes post-AI.

To understand the relationship between lexical features and our measures of petition quality, we consider correlations between variables. SI Appendix Figure 7 shows the linear relationship between lexical features (MATTR, log-transformed word count, Flesch-Kincaid grade level) and our measures of text quality (writing quality score, persuasiveness score). Pearson correlation coefficients with writing quality range from modest (MATTR) to moderate (log-transformed word count, Flesch-Kincaid grade) positive relationships. The same pattern holds for correlations with persuasiveness scores.

Sentiment Analysis

Our analysis shows that petition sentiment is more positive after the introduction of the in-platform AI tool, as measured as the average ratio of positive/negative/neutral sentiment per sentence. SI Appendix Figure 8 shows the distribution of petition sentiment, as calculated with VADER (Valence Aware Dictionary and sEntiment Reasoner) tool. VADER is a tool for sentiment analysis that uses lexical and rule-based features to assign sentiment scores to short text [11]. We use the VADER python library to assign sentiment scores to each petition. To do this, we first decompose each petition into sentences, as is the recommended approach for longer text. VADER provides the ratio of each sentence that is positive, neutral, and negative. To retrieve average per-sentence sentiment for each petition, we then take the average of each positive, neutral, and negative score across the sentences within a petition. SI Appendix Figure 8 shows the distribution of average sentence sentiment for each English language petition written pre- and post- AI in the United States, Canada, Great Britain, and Australia. When petitions are written with access to in-platform AI, the average share of neutral sentiment in petition sentences decreases, as shown by the distribution shifting to the left in SI Appendix Figure 8c. Conversely, the average share of positive sentiment in petition sentences increases, as the distribution shifts to the right in SI Appendix Figure 8a. The distribution of negative sentiment does not shift, but rather narrows about the peak of 0.1, as shown in SI Appendix Figure 8b, thus demonstrating a concentration of average sentence negativity ratio at 0.1. Previous literature has documented a positivity bias in text produced by generative AI tools, which supports our finding of a modest increase in positive sentiment [12, 13].

Title Concreteness

Our data analysis shows that the language used in petition titles is less concrete when petitions are written with access to in-platform AI. We study the specificity of words used in petition titles through the headline concreteness measure developed by Aubin le Quéré and Matias [14]. The authors define concreteness as defined the “degree to which the concept denoted by a word refers to a perceptible entity.” The measure combines Named Entity Recognition tagging with a psychologically validated scale of English word concreteness. We apply the measure to English Change.org petition titles, and display the distribution of title concreteness in SI Appendix Figure 9. We apply the measure to Change.org petitions posted from the US, GB, CA, and AU during the pre-AI period (January 2022-March 2023) and post-AI period (January 2024-December 2024). The measure crafted by Aubin le Quéré and Matias excludes text that does not have

at least one word tagged with their concreteness score. We observe a small decrease in median petition title concreteness with in-platform AI access, as the median title concreteness decreased from 3.63 to 3.50 (Kruskal-Wallis test $H(1)=5341.433$, $p < 0.001$, $\eta^2 = 0.011$). We hypothesize that the in-platform AI tool may contribute to a decline in petition title specificity if users fail to prompt the in-platform AI tool with specific enough information.

Difference-in-Differences

In the main text, we report on static and dynamic difference-in-difference estimates. Here, we provide a series of further analyses that evidence the robustness of our findings. We use the software provided by Callaway and Sant’Anna to create our static and dynamic difference-in-difference estimations [15]. We use the R `did` package (version 2.1.2) to estimate the dynamic and static difference-in-difference analyses. For the static difference-in-differences analysis we clustered standard errors at the month level. For the dynamic difference-in-differences, we calculated standard errors with a multiplier bootstrap (1,000 iterations) to create a uniform confidence band. In the main text, we found that lexical features changed significantly, homogeneity increased, and petition outcomes were either unchanged or worsened by access to the AI tool. While we include figures in the main text to visualize these results, we also provide tables for the static difference-in-differences outcomes here. We do not include tables for the dynamic difference-in-differences, as each table has as many terms as there are time periods (76). However, code for producing these tables is publicly available on OSF.¹ Tables 9-10 include the main text outcome variables (percent of petitions with 1+ comment (30d) and percent of petitions with 10+ signatures). Tables 11-13 include the static difference-in-differences estimates for lexical features in the main text (MATTR, Flesch-Kincaid Grade Level, and word count). Table 14 includes the static difference-in-differences estimation of petition similarity.

In order to satisfy the assumption that trends in treated and control countries were similar prior to the in-platform AI intervention, we used ordinary least squares regression to fit trend lines to the pre-AI period in each group of countries. SI Appendix Figure 10 shows that the interaction term was insignificant in all cases, with no significant difference in slope for pre-AI trends in treated and control countries.

Difference-in-Differences Robustness Checks

We demonstrate the robustness of our results through several additional checks. First, in SI Appendix Figure 11 we demonstrate that dynamic and static difference-in-differences analysis using additional outcome thresholds reveal similar results and the raw count of comments and signatures. SI Appendix Figure 11a and SI Appendix Figure 11b show that for a range of comment and signature thresholds, the share of petitions that reach these thresholds is either significantly lower or unchanged for treated countries with AI access. In SI Appendix Figure 11c and SI Appendix Figure 11d, we log-transformed raw counts of comments and signatures per week, and then normalized these values by the number of petitions produced in each week. We find that this metric, which leverages raw signatures and comments rather than thresholds, reveals the same patterns of unimproved outcomes in the post-AI period.

In addition to estimating the impact of AI on additional outcome measures, we also estimate the impact of AI on additional lexical features. In SI Appendix Figure 12 we also conduct dynamic and static difference-in-differences using time series of additional median weekly lexical features. In particular, we consider words per sentence, characters per word, readability as measured by the Gunning Fog index, and the count of sentences [16]. The Gunning Fog index is a measure of readability that is calculated as $0.4 \times (\frac{\text{words}}{\text{sentences}}) + 100 \times (\frac{\text{complex words}}{\text{words}})$, with complex words defined as select multi-syllabic words. We find that across these four additional measures, after AI access the median values are significantly increased, as is consistent with the shifts observed in the main text.

As our pre-AI period is substantially longer than the post-AI period for which we have differential treatment, we repeat the main text difference-in-differences specification using a subset of the pre-AI data, to demonstrate that our estimates are not indicative of any one pre-AI treatment period. In SI Appendix Figure 13, we show that dynamic and static estimates for lexical and outcome features that use a shorter time frame for the pre-AI period (25 weeks pre-AI, 11 weeks post-AI) recover the same results as in our main text, where we use our full data. Again, lexical metrics are significantly increased, and in this case,

¹https://osf.io/q3hj5/?view_only=3fc5b89d05eb4741957af431b107cf2a

outcome metrics are either unchanged or significantly decreased. This shows that our results are robust to pre-AI time period length.

We end our main analysis difference-in-differences with December 15th, in order to prevent including weeks in which Australia gained access to the AI feature. To demonstrate that results would hold even for the full month of December included in the period of differential access, we repeat our main analysis with the end date of December 31, 2023. SI Appendix Figure 14 we use the full month of December, rather than curtailing our study with the last full week prior to December 15th. In this study, we show that we recover similar results when using the full month of December for the static difference-in-differences results. However, during the final two weeks of December we observe heterogeneous effects for lexical features, with insignificant weekly estimates of change to lexical features. This is potentially indicative of the time in which Australian users received AI access. Outcomes retain the same patterns as in the main text, with significantly reduced share of petitions reaching the signature threshold and unchanged share of petitions reaching the comment threshold.

In SI Appendix Figure 15, we show static and dynamic difference-in-difference estimates had we included short text. In our main report, we drop petitions that have fewer than 2 sentences or 5 words. This is for two reasons: first, such short petitions are unlikely to be legitimate petitions seeking to incite change; second, our lexical metrics require at least two sentences to be coherent. For instance, our readability metric accounts for average number of words per sentence, which would not be meaningful on a one-sentence petition. Likewise, extremely short text is likely to have high lexical diversity by virtue of using few words overall. We show through SI Appendix Figure 15 that results are consistent with our main text, even when including short petitions. Namely, lexical features increase significantly, and outcomes are unchanged.

The difference-in-differences model in the main text does not include covariates. We use a doubly robust [15] difference-in-difference estimator with both inverse probability weighting and ordinary least-squares regression to control for covariates. We include the following covariates: the share of petitions per week and country that were (a) posted on the weekend (b) had decision-makers specified, and the share of users per week and country that (c) were new, and (d) had commented on a petition before. We use these covariates as they were available to us from our scrape of the Change.org platform. However, it is possible that these covariates are post-treatment (impacted by whether or not a petition was written with AI access). For example, perhaps the share of petition-writers that had themselves commented on another user’s petition is *lower* in the post-AI period, if users are less likely to comment on petitions when they know other authors may use AI. For this reason, we interpret these results with caution. From this analysis we again recover estimates for change to style and outcomes as we did in the main analysis. SI Appendix Figure 16 shows that lexical features increase significantly. However, the analysis using this model is slightly different with regard to the outcomes than the main analysis: whereas the main text shows that outcomes become worse with access to AI, this doubly robust estimate shows no directional difference in outcomes. Overall, this analysis supports the claim that access to in-platform AI changes lexical features but does not improve platform-level petition outcomes.

In SI Appendix Figure 17 and Tables 15-19, we present a static difference-in-differences analysis with standard errors clustered by country. In our main analysis, we cluster standard errors by month to account for temporal shocks that would affect all countries [17]. Another approach to clustering standard errors would be to cluster at the unit of treatment assignment, which would be country in this case [18]. However, as we have only four countries used in our main analysis, this would result in too few units for reliable clustering and increase the risk of severe over-rejection [19]. As expected, the 95% confidence intervals calculated with country-clustered standard errors remain significant for lexical features, and are insignificant for outcome features. These results are consistent with our main finding: access to the in-platform AI tool significantly changed platform-level trends in petition lexical style, without improving outcomes.

In SI Appendix Figure 19 we repeat our main text metrics on a subset of petitions that we labeled as being about government and politics topics (we describe methods in the Petition Topic section). While many petitions are posted on Change.org, some pertain to trivial topics or appear to be made without any serious intent to enact change. We therefore sought to understand changes within a subset of petitions that pertain to topics relevant to civic engagement and democratic participation. We found that outcomes and most lexical features were consistent with main text results: lexical features increased significantly while outcomes did not change significantly. The one notable deviation from main text results is that word count is no longer significantly increased when petitions are written with access to AI. Instead, these petitions

are unchanged in length. This is likely because government petitions tended to be longer than the average petition in the pre-AI period. As a result, the in-platform AI feature did not make a difference, even if it was highly utilized, for word count.

In SI Appendix Figure 20 we conduct a placebo static difference-in-differences. It is unnecessary to conduct a placebo dynamic difference-in-differences, as we can understand week-to-week changes in average treatment effects by time period in the pre-AI period by simply observing the pre-AI part of the dynamic figures in the main text. Our placebo test uses the pre-AI data, with the intervention date of October 1, 2022. This is to ensure that any seasonal effects in the main report would be replicated here. To maintain consistency, we also include a placebo A/B test period (which we skip), from April 1 - October 1, 2022. We therefore specify the placebo pre-intervention period as January 1, 2022 - March 31, 2022 and the post-intervention period as October 1, 2022 - March 31, 2023. We stop the placebo post-period before the actual A/B test intervention. The placebo test, as expected, does not recover significant estimates in lexical features or outcomes for the placebo static difference-in-differences.

Verb Frequency

We demonstrate the changes to the linguistic features of petitions written with access to the AI tool in SI Appendix Figure 21. In particular, we show time series of the frequency of select verbs in petition titles. We focus on verbs as these are less specific to the subject or location of the petitions. In SI Appendix Figure 21a, we show the 10 words with the highest frequency in the pre-AI period. We observe that the word “stop” spikes in frequency in March, 2023: this is because of the viral #StoptheWillowProject movement, in which many individuals took to Change.org to create petitions about stopping an oil drilling project that would threaten the environment. Otherwise, we observe consistent usage of these top 10 words across the time period. In SI Appendix Figure 21b, we show the 10 words that have the highest post-AI to pre-AI frequency ratio. One notable case is the word “implement”, which has the greatest jump in pre to post AI frequency, with utilization hovering around 4 titles per 100 petitions after full AI release. In the post-AI period we can also observe the influence of changes to the underlying model that powers the AI tool: in mid 2024, the Change.org AI tool was briefly updated to a different large language model, which produced different text. This period falls outside of our main difference-in-differences analysis.

The influence of the brief model update in 2024 is also apparent in SI Appendix Figure 23, in which the word count trends dip substantially. The time series figures demonstrate that trends prior to AI release were parallel in treated and control countries. Furthermore, after all countries have AI access we observe parallel trends. The only period for which lexical features vary in treated and control countries is during the period of differential AI access.

In SI Appendix Figure 22 we demonstrate that this increase in occurrence of the word “implement” in petition titles was consistent across countries, with <1% of petitions per month using the word pre-AI, and 4-5% of petitions per month using the tool post-AI access. The United States, Great Britain, and Canada (blue lines) had an earlier spike in occurrence of “implement” in October 2023, which was the full roll-out of the AI tool. Australia then had a spike in word usage around December 2023, which was when Australian users gained access to the tool. In 2024 when all four countries had access to the AI tool we observe similar share of petitions in each country used the term “implement” in the titles.

Petition Language Variant

The current edition of the Change.org AI tool does not appear personalized by country (as of time of writing: January 2026). While the AI tool produces petitions in the language of the user’s input text, our exploratory interactions with the tool suggested that if the user inputs text in English varieties or dialects other than American English, the AI tool may yet produce a petition in American English. SI Appendix Figure 24b supports this claim, showing that the percent of petitions classified as American English, using a simple spelling and vocabulary based classifier, increases in Canada, Great Britain, and Australia when petitions are written with access to the AI tool. We use the `abclf` Python package to classify petition English variety. The package uses spelling and vocabulary to classify English variety (American vs. British). The `abclf` package classifies text as either British English, American English, unknown, or as a mix of British and American English. The classifier checks if words are present in the VarCon dictionary, and if not, assigns

unknown. The classifier assigns a text to a variant if it has more than twice as many identified words as the other variant. Otherwise it classifies the text as a mix [20]. As shown in SI Appendix Figure 24, access to the AI tool increases the share of petitions that are assigned either the American English or British English label. This is likely because many petitions do not use any spelling or vocabulary that is specific to a single English variety, as identified by `abclf`. With access to the AI tool, however, petitions had longer and more complicated vocabularies, which were more likely to include words that trigger a vocabulary-based classifier (e.g. “organize” or “organise”). SI Appendix Figure 24e shows that conditional on receiving a label at all (British or American English), the percent of petitions labeled British English decreased while in SI Appendix Figure 24d, the percent of petitions that use American English increased. As SI Appendix Figure 24b shows, the *relative* increase of American English petitions seems to be driven by an increase in petitions labeled American English rather than a decrease in petitions labeled British English. SI Appendix Figure 24c shows that the overall share of petitions labeled British English remains consistent.

The increase in percent of petitions labeled as American English could amount to a quality of service harm [21], as users may find that the petition produced with the tool does not reflect their intended dialect, language, and grammar. However, for the purposes of our study, we do not expect that the tool’s predilection for American English responses to factor into the results we observe. While users in our main control country, Australia, may find the tool to be misaligned with their Australian English, it is also the case that two of our treatment countries (Canada and Great Britain) would experience analogous misalignment. As shown in SI Appendix Figure 24b, the percent of petitions labeled as American English has a parallel increase in all countries. While this classification is likely an underestimate considering the low detection rate of American English petitions among petition produced in the US, the trends in English varieties suggest that English variety does not explain the patterns that we observe with AI access. In SI Appendix Figure 25 we conducted a difference-in-differences analysis where we exclude the United States as a treated country, to compare changes in lexical features and outcomes between countries where the local variety of English is non-American. We find results are consistent with those that include the US as a treated country (as in the main text).

SI Appendix Figure 26 further demonstrates that it is unlikely that quality of service harms by English language variety drive our results. Here, we show the trends reported in the main text are consistent with the single-country comparisons between Australia to Great Britain; Australia to Canada; Australia to the United States. Since Australian and British English similarly diverge from American English, if it were the case that country-level varieties of English affected the outcomes reported in our main analysis, we would expect to see differences in countries’ pairwise comparisons. While it may be the case that the AI tool fails to generate petitions in dialects and varieties of English other than Standard American English (SAE), we do not anticipate that our causal estimates were driven by this aspect of the AI tool.

We also compare petition outcomes by English variety, before and after the introduction of AI to the platform. SI Appendix Figure 27 shows that there is no single shift in all countries’ outcome metrics for petitions written in American English. In the United States, Canada, and Australia, the share of American English petitions that reach 1+ comment in 30 days is lower when petitions are written with access to AI. On the other hand, we see a much more minor difference in the share of British English petitions that reach 1+ comment before and after AI access. If there were a quality-of-service harm introduced by the AI tool in countries for which the English variety is non-American, we would expect to see a consistent decrease in outcomes for American English petitions written post-AI in Canada, Great Britain, and Australia. Though we do see decreases in Canada and Australia, there is no notable difference in Great Britain. Furthermore, there is a similar sharp difference in pre- and post- AI outcomes in the United States for American English petitions, which we would not attribute to a quality-of-service harm from national language variety.

Petition Topic

Change.org petitions cover a wide array of topics, from education to corporate policies, to pop culture and more. The petition topic distribution did not meaningfully shift between the periods we study in the main analysis difference-in-differences. Further, the difference-in-differences results replicate even when focusing the analysis on specific topics that are related to civic action and politics.

To analyze topics, we followed the topic extraction pipeline used by Costello et al., who took open-ended participant text, summarized the text, created embeddings for summarizations, and then clustered

the embeddings to define topics [22]. We created one-sentence petition summaries with the OpenAI Chat Completions API with the model gpt-4o-mini. Our prompt read: “You are summarizing Change.org petitions. Summarize the following petition, which describes the author’s claim or request, in a single sentence. Do not say that it is a petition, and do not add any normative judgment. Merely accurately describe the content in a way the author who wrote the petition would concur with. Frame it as an assertion. If the petition content is already short, no need to change it very much. If it is quite long and detailed, be sure to capture the core, high-level points, and key words relevant to the topic. Do not focus on the evidence provided for the petition, but merely focus on the basic assertion. Return ONLY strict JSON with keys: petition_id, title, claim.” To control input size and cost, we truncated petition text to a maximum of 500 words prior to summarization, while always including the petition title to preserve context. Only 4.37% of petitions were long enough to be truncated.

We then generated embeddings for each summarization with the OpenAI model text-embedding-3-small, which produces 1536-dimensional embeddings. During verification, we identified that approximately 0.01% of English language petitions (115 of 964,751 English language petitions) had missing/empty embeddings. Manual inspection revealed that these petitions had empty summaries, typically originating from petitions with placeholder titles or non-substantive text (e.g., “.”, “ignore”, “null”, “ ”). These likely stem from the nature of the Change.org platform, which does not allow users to delete their petitions. Instead, users might simply edit the text to contain non-substantive placeholder text, like the examples above.

We clustered embeddings with MiniBatchKMeans with K=20, which approximates standard K-means by repeatedly updating cluster centroids based on small random batches of observations. We used both the inertia (within-cluster sum of squares) and the silhouette score to determine the value of K. We used this method due to the size of our dataset (964,636 English language petitions) and the high dimensionality of embeddings. Following clustering, clusters were interpreted by manually inspecting random samples of petition claims within each cluster. We assigned each cluster a descriptive topic label based on its dominant theme, with labels determined independently by two annotators, then combined. In a final consolidation step, we grouped clusters reflecting closely related subject matter under broader thematic categories. For example, clusters related to schooling were consolidated under “School and Education,” entertainment-related clusters under “Pop Culture and Media,” environmental and animal-related clusters under “Conservation and Environment.” This manual review and consolidation process mirrors the qualitative validation emphasized by Costello et al., ensuring that unsupervised clusters correspond to substantially meaningful topics [22].

SI Appendix Figure 28 shows the distribution of topics for 361,635 petitions in the main analysis during the pre-AI, A/B test, and post AI periods. The distribution of topics across the three periods remains consistent, with the most frequent topics being Pop Culture and Media, School and Education, Criminal Justice, and Community Planning. A 13x3 chi-squared test comparing the distribution of topic by period (pre-AI, A/B test, post-AI) had a negligible effect size, as shown by Cramer’s $V < 0.1$ ($\chi^2(24) = 2950.5$, $p < 0.0001$, Cramer’s $V=0.064$), suggesting little to no relationship between topic and time period. Relatedly, SI Appendix Figure 29 shows a consistent share (about one-third) of repeat-petition writers return to the platform to write about the same topic. This trend is consistent across cohorts (Pre/Pre, Pre/Post, Post/Post).

We then focused our analysis on a set of topics related to civic engagement (Conservation and Environment, Community Planning, International Politics, Social and Economic Welfare, Civil Rights and Diversity, Criminal Justice, Public Safety). To verify the topic selection, the first author (IC) sampled 100 petitions randomly (50 civic engagement topics and 50 other topics) and manually classified them as related to civic engagement, or not. The manual ratings had 82% agreement with the label based on the topic assignment via clustering (i.e., civic engagement topic or other). We used this civic engagement label to repeat the static difference-in-differences analysis. In SI Appendix Figure 30, we show that the trends reported in the main analysis, by which lexical features change significantly and outcomes do not show evidence of improvement, are consistent when limited to the set of civic engagement petitions. We also repeat the static difference-in-differences analysis with other petition topics (Pop Culture and Media, Consumer Products and Brands, Leadership Accountability, School and Education, Personal Requests and Inside Jokes, Ban Offensive Content), and observe consistent trends in SI Appendix Figure 31.

In addition to labeling petitions through unsupervised clustering, we sought to understand trends around the specific topic of political petitions, as classified through a large language model. Previous literature in petitioning has often focused on petitions related to politics and government, as these have the potential

for the greatest impact. In order to understand if the share of petitions produced on Change.org that related to government topics changed after the introduction of AI, we labeled petitions as pertaining to a government or non-government topic. We used zero-shot classification with the OpenAI gpt-4o-mini model. Each petition’s text (truncated to 8,000 tokens) is sent to the model with a system prompt plus a category-specific user prompt, and the model responds with just “A” (in cluster) or “B” (not in cluster). The decoding is deterministic (temperature=0, no fine-tuning), and we run the requests in parallel with up to 50 threads. We then conducted a dynamic difference-in-differences analysis to understand how the share of petitions pertaining to government topics changed after the release of the in-platform AI feature. As shown in SI Appendix Figure 18, there was no change in trend around the share of petitions pertaining to government topics. Instead, the share of petitions that concerned government topics in the post-AI period (blue, right of week 26) was consistently insignificantly different from the baseline week (Week -1) prior to A/B testing of the AI feature.

The system prompt is as follows: “You are a highly efficient information detection and extraction engine, specialized in analyzing natural language data. You value accuracy: when the user asks you to extract certain information from given text data, you will try your best to adhere to what is directly mentioned in the text and the extraction criteria. You value efficiency: your response will be concise and follow formatting conventions specified in the extraction prompt.”

The user prompt is as follows: “The following text is a petition. Is the petition topic Government and Public Safety? This includes petitions that are about matters related to local government, government and politics, or public safety. Output a letter only. A) Yes, government and public safety topic B) No, other topic Text: text Answer:”

We then validated responses through a manually labeled set of 153 petitions (2 raters, cohen’s kappa = 0.583). The LLM-annotated results used 104 of the labeled petitions to achieve 86.5% accuracy, 82.3% precision, 89.3% recall on this set.

Predictive Strength of Lexical Features

As we show AI, significantly changes lexical features, and it is expected that the distribution of each feature would be different in the pre-AI and post-AI periods. We show this through the histograms captured in each panel of SI Appendix Figure 32. For instance, the distribution of petition word count shifts from the 5-45 word decile to the 200-228 word decile. The post-AI word count mode corresponds to the word count that the AI feature tends to produce, which is the 100-250 word range. SI Appendix Figure 32 also shows the marginal means of word count deciles in the pre-AI and post-AI periods. The marginal mean of each decile corresponds to the predicted probability of a positive outcome, as measured by the presence of at least one comment within 30 days of petition post. We calculated CIs with HC3 heteroskedasticity-consistent standard errors [23]. Before access to AI, higher word count deciles corresponded to higher average predicted probability of a positive outcome. Thus, with the shift of petition word count to the 100-300 word range, we would expect to see petitions that have average mean predicted probability of positive outcome in the 0.3-0.4 range. However, with access to the feature, the predictive strength of word count plateaus once it enters the AI-range: lengthier petitions are no longer more likely to receive a positive outcome than their shorter counterparts. In fact, petitions in the 97-130 word range have higher, though insignificantly so, average marginal means than petitions in the 167-200 word range. We observe similar patterns for the Flesch-Kincaid grade level and the MATTR.

Repeat Petition Writers

We studied a group of 4,611 repeat petition writers that wrote multiple petitions between January 2022 and November 2024. We exclude the final month of 2024 to ensure all petitions had at least one month before time of data collection, since primary outcome metric is indication of receiving a comment within 30 days. We drop observations that have missing log-transformed Flesch-Kincaid grade levels, which are those petitions for which the Flesch-Kincaid grade level is negative. Furthermore, we only consider repeat petition-writers whose second petitions were written at least 6 months after their first petitions, and exclude the A/B testing period.

We conduct an exploratory descriptive analysis of repeat petition writers. The main text includes the distribution of word count and the share of petitions that reach 1 comment in 30 days across the three cohorts of petition writers. In Figure 33 we show the distribution of MATTR, Flesch-Kincaid grade level, and the share of petitions that reach 10 signatures. We find consistent results to the main text: in the case of lexical measures, the average MATTR and Flesch-Kincaid grade level is higher in users' second petitions when users write their first petition without AI access and their second petition with AI access (Pre/Post cohort). In the case of the 10+ signature outcome, users' second petitions have reduced outcomes when their first petition was written without AI access and their second petition was written with AI access (Pre/Post cohort).

In our main analysis, we design a mixed effects model for inference into the predictive strength of user cohort among repeat petition writers. We show tables with the model summary for outcomes in Table 20, and for lexical features in Table 21. In both cases we use the `lme4` R package to fit our models [24]: we use the BOBYQA optimizer with 200,000 maximum iterations to fit the model for outcomes. In Table 20, the first column shows the coefficients and standard deviation in parentheses for each term of the model trained to predict if a petition received at least 1 comment in 30 days. Likewise, the second column shows the coefficients and standard error for each term of the model trained to predict if the petition received at least 10 signatures. Both models were generalized logistic regression models with random effects for user ID. In this model, the intercept reflects the first petitions written in the Pre/Pre cohort. In Table 21, we show the coefficients and standard error for linear regression model trained to predict petition MATTR, Flesch-Kincaid grade level, and word count. We log transform Flesch-Kincaid and word count, but not MATTR, as it is normally distributed without transformation. Similarly, the intercept reflects the first petitions written in the Pre/Pre cohort. We then use pairwise contrasts with a Bonferroni adjustment to compare users' first and second petitions, across the three cohorts.

Synthetic Control

Synthetic control is an approach that allows us to compare trends in treated countries to trends in control countries without a strict requirement of parallel trends. To confirm the findings we report in the main text, we conduct a synthetic control analysis in which we use the `synth` package version 1.1.8 in R to compare a single treated unit against a synthetic control [25]. With the synthetic control, we can leverage data from countries that did not qualify for the difference-in-differences analysis in the main text. In particular, we use trends from English petitions written in Australia, India, and New Zealand to create our synthetic control.

The `synth` package identifies country-level weights that minimize the pre-treatment gap between treated and control countries with the Broyden-Fletcher-Goldfarb-Shanno algorithm. This synthetic control is created by using a weighted combination of control units to approximate the characteristics of the treated units. We then use the actual time series for each control country, weighted by the optimized weights, to produce a synthetic counterfactual. In order to create the weighted synthetic controls we provide the same set of covariates as are included in the full-covariate difference-in-differences robustness check. These are as follows: the share of petitions per week and country that were (a) posted on the weekend (b) had decision makers specified, and the share of users per week and country that (c) were new, and (d) had commented on a petition before. We also include the lagged outcome (lagged at 4, 8, and 12 weeks) for matching across groups [26]. We can then compare the time series in the treated countries to the time series in the synthetic counterfactual, which provides us with an estimate of the causal effect of the intervention (in-platform AI access). SI Appendix Figures 34, 35, 36 show the time series of the actual treated countries against the synthetic control estimates, which are consistent with our difference-in-differences analysis. In particular, the word count, lexical diversity, and readability of petitions written in treated countries are higher than in the synthetic control counterfactual trend during the period of differential access. Outcome metrics (i.e., the share of petitions that reach 1+ comment in 30 days or 10+ signatures overall) in treated countries are lower than the synthetic control counterfactual during the period of differential access.

Through the synthetic control analysis, we can see that after full AI access, the actual treated petition lexical features (MATTR, Grade Level, Word Count) are greater than the synthetic control estimates. Likewise, the actual treated outcome metrics are lower than the synthetic control outcome metrics. These trends are consistent with the patterns we observed in our main differences-in-differences analysis.

Bayesian Structural Time Series

Another form of synthetic control is the Bayesian Structural Time Series (BSTS) model, which is leveraged through the `CausalImpact` R software [27] to create a synthetic control that accounts for both trends in control countries as well as patterns of seasonality, and general trends. We choose to use the BSTS approach to predict participation, because unlike outcomes and lexical features, trends in participation appear to have a strong seasonal component. We use participation in control countries (same as earlier synthetic control: Australia, India, New Zealand) to predict synthetic control outcomes. We fit a BSTS model with two components: a semi-local linear trend, and a seasonal component. We fit the BSTS model with the `bsts` R package using package-default priors throughout (including a spike-and-slab prior on the control-series regression coefficient) and 1,000 MCMC iterations. In SI Appendix Figure 37 we show three sub-figures that describe platform participation on Change.org after AI access. In SI Appendix Figure 37a, we show that weekly count of users that produce petitions (normalized by pre-AI average and standard deviation) does not appear to shift from the 2022 mean. In Figure 37b, we show raw user participation counts in a solid line for the period before AI access, during the A/B period, and after the full launch. Figure 37c further elucidates this result, through the actual vs. predicted time series (top), point-wise estimate, for which the effect is estimated for every week, (middle), and cumulative estimate, for which the overall impact over the post-AI time period is shown (bottom). We find that user participation fell well within a 95% credible interval of predictions. The estimated effect of the intervention, according to the BSTS model, is insignificant ($\Delta_{treated-control} = -354$ users (s.d. = 901), $p = 0.343$, 95% credible interval [-2167, 1263]).

We repeat the BSTS analysis process on petition production, rather than user participation, to see if results are consistent. Indeed, Figure 38 demonstrates that the actual trends in petition production do not diverge from predicted trends estimated using pre-AI seasonal trends and production from control countries. This is not surprising as only 6.38% of users wrote multiple petitions, and as such we would expect to see similar trends in user participation and petition production over time.

Interrupted Time Series

In our preregistration we planned to use an interrupted time series analysis through a regression discontinuity in time method to support our analysis. Figure 39 shows the output of this analysis, in which we plot changes to the slope of fit lines in lexical features and outcomes, before and after the start of A/B testing. In this case, the intervention is the start of A/B testing rather than the full launch. We use the start of A/B testing as the intervention date for two reasons. First, this date enables us to have a clean threshold for the discontinuity, rather than a 6-month gap when comparing pre-AI to post-AI. Second, because we do not know how the A/B testing period differed from the full-launch, we would not be able to interpret changes between the A/B testing and the post-AI period: shifts to level or slope of the trend line could be a result of expanding access or shifting features during experimentation. However, comparing pre-AI period to A/B testing period we can hypothesize that we would expect to see a change in slope if A/B testing occurred incrementally and changed lexical features and outcomes. In Figure 39a, we show that the slope of the line fit to predict lexical features in treated countries (top to bottom: MATTR, Flesch-Kincaid grade level, word count) increased after the introduction of A/B testing for the in-platform AI feature. Conversely, Figure 39b shows that the slope of the line fit to predict lexical features in control countries, meaning those that were not included in the A/B test, did not have the same increase. Figure 39c and Figure 39d show that the slope of the line fit to predict outcomes seems to increase post-A/B testing, for both treated and control countries. These results are consistent with those found with our difference-in-differences analysis and synthetic control methods studying the full-launch.

References

- [1] Change.org. Change.org Home. *Change.org*. <https://www.change.org>, 2023. Accessed June 5, 2026. Archived at *Internet Archive*, <https://web.archive.org/web/20231101012225/https://www.change.org/>.
- [2] Change.org. Change.org Home. *Change.org*. <https://www.change.org>, 2025. Accessed February 14, 2025.

- [3] Change.org. Change.org Home. *Change.org*. <https://www.change.org>, 2026. Accessed January 26, 2026.
- [4] David M. J. Lazer, Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062, 2020. doi: 10.1126/science.aaz8170. URL <https://www.science.org/doi/abs/10.1126/science.aaz8170>.
- [5] Joseph B. Bak-Coleman, Stephan Lewandowsky, Philipp Lorenz-Spreen, Arvind Narayanan, Amy Orben, and Lisa Oswald. Moving towards informative and actionable social media research, 2025. URL <https://arxiv.org/abs/2505.09254>.
- [6] Joseph Bak-Coleman, Jevin West, Cailin O’Connor, and Carl T. Bergstrom. Industry influence in high-profile social media research, 2026. URL <https://arxiv.org/abs/2601.11507>.
- [7] Steven Bird and Edward Loper. Natural language toolkit (nltk), 2004. URL <https://www.nltk.org/>.
- [8] Desklib. Desklib AI Text Detector, 2025. URL <https://huggingface.co/desklib/ai-text-detector-v1.01>.
- [9] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586. URL <https://www.science.org/doi/abs/10.1126/science.adh2586>.
- [10] Alyssa Appelman and S. Shyam Sundar. Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, 93(1):59–79, 2016. doi: 10.1177/1077699015606057. URL <https://doi.org/10.1177/1077699015606057>.
- [11] C. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014. doi: 10.1609/icwsm.v8i1.14550.
- [12] Esra Dönmez, Maximilian Maurer, Gabriella Lapesa, and Agnieszka Falenska. AI argues differently: Distinct argumentative and linguistic patterns of LLMs in persuasive contexts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34583–34614, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1755. URL <https://aclanthology.org/2025.emnlp-main.1755/>.
- [13] David M. Markowitz, Jeffrey T. Hancock, and Jeremy N. Bailenson. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1):63–82, 2024. doi: 10.1177/0261927X231200201. URL <https://doi.org/10.1177/0261927X231200201>.
- [14] Marianne Aubin Le Quéré and J. Nathan Matias. When curiosity gaps backfire: effects of headline concreteness on information selection decisions. *Scientific Reports*, 15(1):994, January 2025. ISSN 2045-2322. doi: 10.1038/s41598-024-81575-9. URL <https://www.nature.com/articles/s41598-024-81575-9>.
- [15] Brantly Callaway and Pedro H. C. Sant’Anna. did: Difference in differences. R package, 2021. URL <https://bcallaway11.github.io/did/>. R package version 2.1.2.
- [16] Robert Gunning. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13, 1969. doi: 10.1177/002194366900600202. URL <https://doi.org/10.1177/002194366900600202>.
- [17] Maria del Rio-Chanona, Nadzeya Laurentsyeve, and Johannes Wachs. Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow. *PNAS Nexus*, 3(9):pgae400, September 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae400.

- [18] Jonathan Roth, Pedro H.C. Sant’Anna, Alyssa Bilinski, and John Poe. What’s trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2): 2218–2244, 2023. ISSN 0304-4076. doi: 10.1016/j.jeconom.2023.03.008.
- [19] James G. MacKinnon and Matthew D. Webb. Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics*, 218(2):435–450, October 2020. ISSN 0304-4076. doi: 10.1016/j.jeconom.2020.04.024.
- [20] Sreekanth Malyala. abclf, version 1.0.1. <https://pypi.org/project/abclf/1.0.1/>, 2025. Python package.
- [21] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N’Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’23, page 723–741, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604673. URL <https://doi.org/10.1145/3600211.3604673>.
- [22] Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024. doi: 10.1126/science.adq1814. URL <https://www.science.org/doi/abs/10.1126/science.adq1814>.
- [23] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000. doi: 10.1080/00031305.2000.10474549. URL <https://www.tandfonline.com/doi/abs/10.1080/00031305.2000.10474549>.
- [24] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4, 2015.
- [25] Jens Hainmueller aut cre, Alexis Diamond aut, and Alberto Abadie aut. Synth: An r package for synthetic control methods in comparative case studies, 2011. URL <https://www.jstatsoft.org/v42/i13/>.
- [26] Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, 2021. ISBN 9780300251685. URL <http://www.jstor.org/stable/j.ctv1c29t27>.
- [27] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247 – 274, 2015. doi: 10.1214/14-AOAS788. URL <https://doi.org/10.1214/14-AOAS788>.