Media Effects and Party Support Volatility in US: A Synthetic Data Approach

Karl Ho
University of Texas at Dallas
karlho@gmail.com

John Marion Glendenning
University of Texas at Dallas
John.Glendenning@UTDallas.edu

Abstract

Research on media use and partisanship has been building fast recently. The literature is populated with studies using partisanship as an independent variable driving differences in media usage or partisanship as the dependent variable, analyzing the exogenous effects of media usage or media coverage. The third group of studies utilize partisanship as a moderator variable conditioning various types of media effects (Shehata and Strömbäck 2020). This study takes on a novel approach to study the dynamics of media attention and party support in the United States. In addition to employing traditional survey data, we gather data using "silicon sample" generated by ChatGPT (Argyle et al. 2023) to identify the traditionally difficult sub-populations which conventional sampling methods can barely reach. By comparing the conventional samples such as ANES with silicon samples, we explore new methods in detecting the voting intention and party support among potential voters who are precarious and hard to access.

Introduction: Media Effects and Partisan Volatility[1]

The relationship between media consumption and partisan attitudes has become increasingly complex in the contemporary American political landscape. While traditional models of media effects focused on persuasion and agenda-setting, recent scholarship emphasizes the role of selective exposure and media choice in reinforcing existing political predispositions (**?**).

The American National Election Studies (ANES) has long served as the gold standard for measuring public opinion and voting behavior in the United

[1]This paper is in the first draft version. The author would like to thank the reviewer, conference organizers and participants for their valuable feedback and suggestions.

States. However, traditional survey methods face mounting challenges in reaching volatile voter segments—particularly young voters, non-partisans, and those disengaged from conventional political processes.

The Challenge of Volatile Voters

American elections have witnessed increasing electoral volatility in recent cycles. The phenomenon of swing voters, ticket-splitters, and late-deciding voters presents both methodological challenges for pollsters and strategic puzzles for campaigns. Understanding the media consumption patterns of these volatile voters is essential for explaining partisan dynamics.

Silicon Samples: A Novel Approach

Argyle et al. (2023) introduced the concept of "silicon samples"—synthetic survey respondents generated by large language models that can simulate human opinion distributions. This approach offers promising avenues for:

1. Generating synthetic responses for hard-to-reach populations
2. Testing survey instruments before deployment
3. Exploring counterfactual scenarios in media effects research

Literature Review

Media Effects and Voting Behavior

The relationship between mass media and political behavior has been a central concern of political science research for nearly eight decades. This literature review traces the theoretical evolution from early "minimal effects" paradigms to contemporary understandings of media influence on partisan attitudes and voting behavior, while also examining emerging methodological innovations using synthetic data approaches.

The Columbia Studies and Minimal Effects Paradigm

The systematic study of media effects on voting behavior began with the pioneering work of Paul Lazarsfeld and colleagues at Columbia University. (**?**) conducted the landmark Erie County study during the 1940 presidential election, interviewing 2,400 voters across multiple waves to document their decision-making processes. This study introduced the influential "two-step flow" model of communication, demonstrating that information from mass media typically reaches voters indirectly through “opinion leaders” rather than directly affecting attitudes.

(**?**) extended this approach with their Elmira study during the 1948 election. Their findings reinforced what became known as the “minimal effects” hypothesis, suggesting that mass media primarily reinforced existing predispositions rather than converting voters. The authors emphasized the primacy of sociological factors—family, religion, social networks—in shaping political preferences, with media serving a secondary, reinforcing role.

The Michigan Model and Party Identification

A parallel tradition emerged from the University of Michigan, fundamentally reshaping our understanding of voting behavior. (**?**) introduced the concept of party identification as the central organizing principle of political attitudes. Based on early waves of what became the American National Election Studies (ANES), they argued that partisan identification, often inherited from parents, functions as a “perceptual screen” through which voters interpret political information.

The Michigan model suggested that media effects were filtered through this partisan lens, limiting direct persuasion while potentially reinforcing existing partisan attachments. This psychological approach to voting behavior contrasted with the Columbia school’s sociological emphasis but similarly implied constrained media influence.

Agenda-Setting and Priming Effects

The 1970s saw a theoretical breakthrough that revitalized media effects research. (**?**) demonstrated through their Chapel Hill study that while media may not tell citizens what to think, they are remarkably successful at telling them what to think *about*. By correlating issue salience in news coverage with voter perceptions of important issues during the 1968 presidential election, they established agenda-setting as a powerful, if indirect, form of media influence.

(**?**) extended this framework through innovative experimental designs, demonstrating two additional mechanisms of media influence. First, “priming” effects showed that news coverage alters the criteria citizens use to evaluate political leaders. Second, “framing” effects revealed that how issues are presented—emphasizing episodic versus thematic contexts—shapes attributions of responsibility and policy preferences. These findings rehabilitated the notion of consequential media effects while identifying more subtle mechanisms than direct persuasion.

Information Processing and Elite Influence

(**?**) synthesized these various traditions into a comprehensive theory of public opinion formation. His Receive-Accept-Sample (RAS) model proposes that citizens vary in their attention to elite-mediated political information, their propensity to accept messages consistent with prior predispositions, and the considerations they sample when constructing survey responses. Zaller's framework emphasized the role of media in transmitting elite discourse, with effects moderated by individual-level political awareness.

This model has profound implications for understanding media effects: highly aware citizens are most likely to receive political information but also most capable of resisting counter-attitudinal messages. Conversely, less aware citizens receive fewer messages but are more susceptible when exposed. The result is often non-monotonic relationships between awareness and attitude change.

Media Choice and Partisan Polarization

The transformation of the media environment from broadcast-era scarcity to contemporary abundance has profoundly altered the dynamics between media choice and party support. (**?**) demonstrates how expanded media choice—through cable television and later the internet—has enabled citizens to opt out of news consumption entirely or to selectively expose themselves to ideologically congenial sources. His "Conditional Political Learning" model demonstrates that entertainment preferences now predict political knowledge gaps more strongly than demographic factors.

Prior's work suggests that media effects have become increasingly conditional on self-selection. In low-choice environments, even politically disinterested citizens encountered news through broadcast television; in high-choice environments, the same citizens can avoid political information altogether. The result is growing inequality in political knowledge and participation, with elections increasingly dominated by committed partisans.

AI and Synthetic Data in Political Research

The Promise of Silicon Samples

Recent advances in large language models (LLMs) have opened new possibilities for social science research. (**?**) introduced the concept of "silicon samples"—synthetic survey responses generated by conditioning LLMs on sociodemographic backstories from real survey respondents. Their groundbreaking study demonstrated that GPT-3, when provided with detailed persona descriptions based on ANES respondents, could generate responses exhibiting "algorithmic fidelity" to human opinion distributions.

Argyle and colleagues found high correlations between human and synthetic voting predictions, with tetrachoric correlations exceeding 0.9 for partisan vote choice. They argued that LLMs encode nuanced representations of the relationship between demographics, attitudes, and political behavior, potentially offering a novel tool for social science research.

Limitations and Cautions

However, subsequent research has identified important limitations. (**?**) conducted a systematic evaluation of ChatGPT's ability to replicate ANES feeling thermometer responses. While aggregate means closely matched human surveys, they identified several critical problems:

1. **Reduced variance**: Synthetic responses showed significantly less variation than human respondents, particularly for questions about racial and religious groups
2. **Regression bias**: Coefficients from synthetic data often differed substantially from those obtained with human samples
3. **Prompt sensitivity**: Minor variations in prompt wording produced significantly different response distributions
4. **Temporal instability**: The same prompts yielded different results across a three-month period

These findings suggest that while LLMs may capture central tendencies in public opinion, they are not yet reliable substitutes for human respondents in rigorous social science research. The reduced variance is particularly problematic for studying opinion heterogeneity across demographic subgroups—precisely the application for which silicon samples were proposed.

Theoretical Framework for the Present Study

This study synthesizes insights from both literatures to examine media effects on partisan volatility. We leverage the ANES as a validated measure of American political attitudes while using synthetic samples not as replacements for human data, but as complementary tools for exploring theoretical mechanisms.

Our approach is motivated by several theoretical considerations:

1. **Media environment transformation**: Following Prior, we examine how contemporary high-choice media environments may differentially affect partisan stability across levels of political engagement

2. **Elite communication dynamics**: Following Zaller, we consider how exposure to elite discourse through various media channels shapes the accessibility of partisan considerations

3. **Methodological innovation**: Following Argyle et al., we explore whether synthetic samples can illuminate patterns that might be difficult to observe in finite human samples, while heeding the cautions raised by Bisbee et al. regarding appropriate interpretation

The combination of conventional survey data and synthetic samples represents a novel mixed-methods approach that we term "silicon-augmented survey research." This methodology allows us to generate theoretically motivated hypotheses through synthetic data exploration while validating substantive findings against rigorous human survey data.

This study synthesizes insights from both literatures to examine media effects on partisan volatility. We leverage the ANES as a validated measure of American political attitudes while using synthetic samples not as replacements for human data, but as complementary tools for exploring theoretical mechanisms.

Our approach is motivated by several theoretical considerations:

1. **Media environment transformation**: Following Prior, we examine how contemporary high-choice media environments may differentially affect partisan stability across levels of political engagement

2. **Elite communication dynamics**: Following Zaller, we consider how exposure to elite discourse through various media channels shapes the accessibility of partisan considerations

3. **Methodological innovation**: Following Argyle et al., we explore whether synthetic samples can illuminate patterns that might be difficult to observe in finite human samples, while heeding the cautions raised by Bisbee et al. regarding appropriate interpretation

The combination of conventional survey data and synthetic samples represents a novel mixed-methods approach that we term "silicon-augmented survey research." This methodology allows us to generate theoretically motivated hypotheses through synthetic data exploration while validating substantive findings against rigorous human survey data.

Media Effects and Partisanship

The study of media effects on partisanship has evolved through several theoretical paradigms. Early minimal effects models gave way to more nuanced understandings of agenda-setting, framing, and priming effects (**?**).

(**?**) demonstrated that partisan news exposure can produce activation, conversion, or reinforcement effects depending on the viewer's prior dispositions and the intensity of exposure. This heterogeneous treatment effects perspective aligns with our focus on identifying volatile subpopulations.

Selective Exposure and Media Choice

(**?**) provided empirical evidence of ideological selectivity in media use, documenting how Republicans and Democrats gravitate toward different news sources. This selective exposure has implications for understanding how media environments shape—and are shaped by—partisan identity.

Data and Methods

Conventional Survey Data: ANES

We employ data from the American National Election Studies (ANES), which provides comprehensive measures of:

- Media consumption patterns across traditional and digital platforms
- Partisan identification and strength of attachment
- Vote choice and electoral participation
- Demographic and socioeconomic characteristics

Data is from the 2016, 2020, and 2024 American National Election Studies (ANES) Time Series surveys.

Study	Pre-Election N	Post-Election N	Panel Component	Mode(s)
ANES 2016	4,271	3,649	—	Face-to-face, Web
ANES 2020	8,280	7,449	2016 Panel (n 2,800)	Web, Phone, Video
ANES 2024	5,521	4,964	2016-2020 Panel (n=2,171)	Face-to-face, Web, Video, Phone, Paper
Total	**18,072**	**16,062**	—	—

ANES 2016 Time Series (n = 4,271)

Pre-election: September 2016 - November 2016

Post-election: November 9, 2016 - January 2017

Mode: Dual-mode design with face-to-face (n = 1,181) and web (n = 3,090)

Sampling: 60 primary sampling units (PSUs) selected from 48 contiguous states and DC for face-to-face; simple random sample of residential addresses from all 50 states and DC for web component

Response rates: Face-to-face component weighted response rate accounted for sub-sampling during collection; 90% post-election reinterview rate for face-to-face, 84% for web

Population: Target population of approximately 222.6 million U.S. citizens age 18+ (face-to-face) and 224.1 million (web, including Alaska and Hawaii)

ANES 2020 Time Series (n = 8,280 pre-election; 7,449 post-election)

Pre-election: August 18 - November 2, 2020

Post-election: November 8, 2020 - January 4, 2021 Mode: Mixed-mode contactless design adapted for COVID-19: web (n = 7,782), telephone (n = 139), video (n = 359)

Sample components: Fresh cross-sectional sample plus panel reinterviews with 2016 ANES respondents (n 2,800, representing 77.9% reinterview rate)

Response rates: Fresh cross-sectional component AAPOR RR1 of 36.7% overall (37.8% web-only, 39.7% mixed web, 27.6% mixed video); AAPOR RR3 of 40.9% overall accounting for eligibility

Post-election reinterview: 90.0% overall (94.0% for 2016 panel, 87.9% for fresh sample)

Population: 231 million non-institutional U.S. citizens age 18+ living in 50 states or DC

ANES 2024 Time Series (n = 5,521 pre-election; 4,964 post-election)

Pre-election: August 3 - November 5, 2024 Post-election: November 7, 2024 - February 17, 2025 Mode: Mixed-mode design with face-to-face, web, video, phone, and paper-and-pencil Sample components: Fresh in-person sample (n = 1,042 pre-election, 925 post-election), fresh web sample (n = 2,308 pre-election, 1,969 post-election), 2016-2020-2024 panel extension (n = 2,171 pre-election, 2,070 post-election) Response rates: 33% for fresh in-person sample (detailed rates by component available in study documentation) Population: 232.5 million U.S. citizens age 18+ (face-to-face, 48 contiguous states and DC); 234.1 million (web, all 50 states and DC) Innovations: Return to face-to-face interviewing post-pandemic; three-wave panel design extending from 2016; mixed-mode flexibility including paper follow-up for web non-respondents

Variable Selection and Harmonization

Strategy Overview

Harmonizing data across the 2016, 2020, and 2024 ANES waves required careful coding consisency. The harmonization process identified equivalent variables, and created consistent, common coding schema. Limiting categories was important because in later analysis the modeling would have failed from data sparsity.

Another design choice was maintaining the distinction between structural zeros, which are logically impossible combinations, such as non-voters having a vote choice, and missing data such as non-response or item skips. This distinction also proved essential for synthetic data generation. Bayesian Network models and similar require proper constraint handling to avoid generating implausible synthetic data.

Outcome Variables

Turnout (VOTE_ANY_HARM)

The primary turnout measure is a harmonized binary variable indicating whether the respondent voted in the presidential election. All three ANES waves used consistent question wording for self-reported turnout. Responses were coded as "Voted" (1) or "Did not vote" (0). Item non-response for this item was minimal (typically <2%), and missing cases were excluded from the analysis.

Vote Choice (PRES_VOTE_CAT to VoteChoice2)

Presidential vote choice was harmonized into four mutually exclusive categories to accommodate third-party candidates and non-voters across waves:

Democrat - Voted for the Democratic candidate (Clinton 2016, Biden 2020, Harris 2024)

Republican - Voted for the Republican candidate (Trump in all three waves)

Other - Voted for any third-party or write-in candidate

NoVote - Did not vote in the presidential election

Respondents coded as non-voters on the turnout variable (VOTE_ANY_HARM = 0) were assigned to "NoVote," regardless of any stated preference or intent. The "Other" category aggregates all non-major-party votes, including Libertarian, Green, other minor-party candidates, and write-ins.

Media Exposure Variables

Media Exposure (MediaExposure) Constructing a harmonized media variable presented a challenge, since the media landscape evolved dramatically between 2016 and 2024. We posited that exposure to politically-oriented news sources shaped political attitudes and voting behavior.

The harmonization process identified functionally equivalent media source questions in the ANES data and classified them. For each respondent, media consumption patterns across these sources were coded into four categories:

Left - Consumed left-leaning sources with little to no right-leaning exposure *Right* - Consumed right-leaning sources with little to no left-leaning exposure *Both* - Consumed news from both left and right sources *None* - Consumed neutral/local sources or none at all

Demographic Variables

Age (AGE_HARM) ANES provided respondent age in years as a continuous measure, and the harmonization collapsed it into four categories:

18-29 - Young adults, emerging electorate 30-44 - Early to middle career adults 45-59 - Late career, pre-retirement adults 60+ - Retirement age and elderly voters

These categories align with generational cohort boundaries while ensuring adequate numbers in each combination.

Sex (SEX_HARM)

Sex was measured as a binary variable (Male/Female) across the datasets. Sex composition showed minor variation across waves, being typically 52-54% female.

Education (EDU_HARM) Educational attainment presented moderate harmonization challenges. The data was harmonized into the following categories:

High school or less - No high school diploma through high school graduate *Some college* - Some college coursework, associate's degree, vocational/technical training *4-year+ degree* - Bachelor's degree or higher (including graduate and professional degrees)

Region (REGION) Geographic region uses the standard four-category Census division scheme, which is coded consistently across all ANES waves, so no harmonization was required:

Northeast - New England and Mid-Atlantic states Midwest - East North Central and West North Central states South - South Atlantic, East South Central, and West South Central states West - Mountain and Pacific states

Missing Data Patterns

Skipped for now

Data Quality Assessment

Skipped for now

Descriptive Statistics

The final sample after harmonization and cleaning is shown in Table 2. This is 15,608 total samples.

Table 2: Sample sizes by election year

Year	: n :
2016	4270
2020	8280
2024	5521

Table 3 shows the statistic of the missing data. The Presidential vote choice (pres_vote_cat) had the highest missingness rate. This may reflect a refusal to disclose, voters who are uncertain, or survey breaks or errors. It is rather high relative to other missing values. Turnout (vote_any_harm) is only slightly lower. There may be a correlation between these. The substantially higher missing rates for outcome variables (11-13%) compared to demographic variables (1-4%) suggests non-random missingness.

Table 3: Missingness by variable

variable	n Missing	Percent Missing
pres_vote_cat	2275	12.59
vote_any_harm	2068	11.44
age_harm	748	4.14
edu_harm	231	1.28
race_harm	202	1.12
sex_harm	165	0.91
year	0	0.00
region	0	0.00
media_lean_cat	0	0.00

Table 4 showed the association between media consumption and electoral participation. Partisan (*Left/Right*) and independent (*Both*) consumers consistently exceed the turnout rates of *None* by 6–13 percentage points. Notably, the *Both* category shows no evidence of demobilizing voters, matching or exceeding the high turnout of single-source partisans (~91% in 2020/2024). However, two oddities needed to be addressed. The fluctuation of the *None* category, ranging 46–76% of the sample, artifacts necessitated year-conditioned priors. The sparsity in 2024—specifically among *Left* and *Both* media consumers required Bayesian parameter smoothing.

Table 4: Media exposure by turnout and year (counts)

YEAR	MediaExposure	Turnout	n
2016	Both	NoVote	30
2016	Both	Voted	107
2016	Left	NoVote	149
2016	Left	Voted	700
2016	None	NoVote	567
2016	None	Voted	1722
2016	Right	NoVote	22
2016	Right	Voted	159
2020	Both	NoVote	41
2020	Both	Voted	413
2020	Left	NoVote	235
2020	Left	Voted	2353
2020	None	NoVote	656
2020	None	Voted	2695
2020	Right	NoVote	77
2020	Right	Voted	817
2024	Both	NoVote	10
2024	Both	Voted	100
2024	Left	NoVote	23
2024	Left	Voted	175
2024	None	NoVote	698
2024	None	Voted	2976
2024	Right	NoVote	93
2024	Right	Voted	790

Table 5: Year-level baseline summaries (turnout and vote shares among voters)

YEAR	n	turnout_rate	dem_share_among_voters	rep_share_among_voters
2016	3456	0.778	0.473	0.433
2020	7287	0.862	0.550	0.401
2024	4865	0.831	0.535	0.421

Methods: Synthetic Data Generation

Architecture

Stage 1: Data Harmonization

As noted above.

Stage 2: Latent Class Analysis

Latent Class Analysis (LCA) was used to identify natural voter segments based on observed behaviors and media consumption patterns. This unsupervised clustering approach created heterogeneous subpopulations, reducing the dimensions of the data.

Stage 3: Bayesian Network Specification

Bayesian Networks (BNs) were then used to encode conditional dependencies and causal assumptions which governed the relationships between voter characteristics and political behavior. The directed acyclic graph (DAG) structure made these causal assumptions explicit. (Figure Figure 1) This pipeline fitted separate models for each election year rather than pooling across waves. This approach was chosen because electoral contexts differ, media environments change, and the effect of media on voting behavior will vary election-to-election.

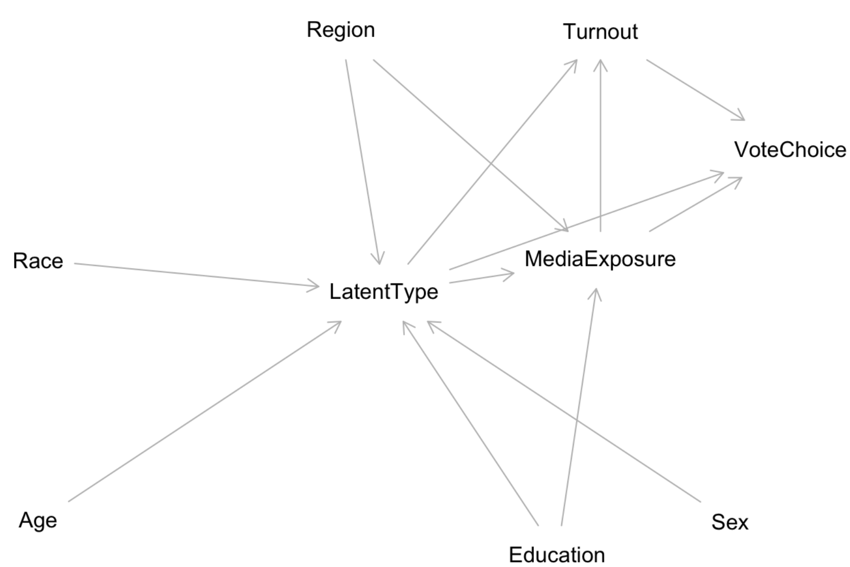

Figure 1: Directed Acyclic Graph

Stage 4: Bayesian Network Fitting

Conditional probability tables were built for each node in each year, and smoothing was applied to handle sparse cells. This created year-specific, fitten BNs.

Stage 5: Synthetic Data Generation

Synthetic voter datasets were generated by sampling from a fitted Bayesian network. This produced two types of data: observational, mirroring the original survey relationships, and interventional, simulating possible scenarios. Following generation, logical constraints were enforced, such as preventing non-voters from having candidate preferences, to ensure data consistency for analysis.

Stage 6: Validation

Data quality was assessed through tests such as comparing individual and joint variable patterns between real and synthetic data, testing whether statistical models could distinguish synthetic from real cases, evaluating whether the synthetic data reproduced meaningful associations from substantive analyses, confirming all logical and demographic constraints were met, and producing validation metrics with diagnostic visualizations.

Silicon Sample Generation

Following Argyle et al. (2023), we generate synthetic survey respondents using large language models. Our approach involves:

1. **Persona construction**: Creating detailed demographic and attitudinal profiles
2. **Survey simulation**: Prompting the model to respond as each persona would
3. **Validation**: Comparing aggregate distributions with known survey marginals
4. **Targeted sampling**: Over-sampling volatile voter profiles

Voter Behavior Typology

Classification Framework

Building on established typologies in electoral behavior research, we classify voters based on their partisan consistency across election cycles:

Voter Type	Definition	Expected Media Pattern
Loyal Partisans	Consistent party voters across cycles	Heavy partisan media consumption

Voter Type	Definition	Expected Media Pattern
Switchers	Changed party vote between elections	Mixed media diet
Mobilized	Non-voters who entered electorate	Social media heavy
Demobilized	Previous voters who abstained	Declining news attention
Independents	Consistent non-partisan voters	Diverse, less frequent consumption

Analysis

Predictive Modeling Approach

We employ three classification models to predict volatile voter behavior:

1. **Logistic Regression**: For interpretable coefficient estimates
2. **Random Forest**: To capture non-linear interactions
3. **XGBoost**: For gradient-boosted ensemble performance

Variable Importance

Figure 1: Variable Importance for Predicting Vote Switching

Note: Placeholder for variable importance visualization

Model Comparison

Figure 2: ROC Curves Comparing Model Performance

Note: Placeholder for ROC curve comparison

Comparing Silicon and Conventional Samples

Distribution Comparisons

A key contribution of this study is the systematic comparison between conventional ANES samples and silicon samples across:

- Partisan identification distributions
- Media consumption patterns
- Vote choice distributions
- Demographic compositions

Advantages for Volatile Voter Research

Silicon samples offer particular advantages for studying volatile voters:

1. **No non-response bias** among hard-to-reach populations
2. **Ability to target specific profiles** that are rare in probability samples
3. **Cost efficiency** for exploratory analysis
4. **Rapid iteration** for testing hypotheses

Discussion

Implications for Public Opinion Research

The integration of silicon samples with conventional survey methods opens new possibilities for understanding volatile voters who are systematically underrepresented in traditional polling.

Limitations

Several limitations warrant acknowledgment:

- Silicon samples reflect training data, not ground truth preferences
- Temporal validity concerns as political environments shift rapidly
- Questions about external validity across different electoral contexts

Conclusion

This study advances methodological innovation in media effects research by demonstrating the utility of silicon samples for studying volatile voter populations. Future research should continue validating synthetic data approaches against benchmark surveys while exploring applications in campaign strategy and democratic engagement.

References