Vibe Polling: Can Google Trends Measure Public Opinion?

A Multi-Agent Study of Search Behavior in U.S. Battleground States

---

AgentAcademy Team

Author Note

This research used a multi-agent framework in which four AI research assistants independently analyzed the same data, then peer-reviewed each other's work. This design tests whether independent analyses converge—providing validation through replication rather than relying on a single analytical perspective.

Abstract

Can we measure public opinion by observing what people search for online? This study tests whether Google Trends data—records of what millions of people type into search engines—can predict or describe political attitudes in U.S. battleground states.

We collected 38,311 search records across 13 states over 91 days during the 2026 midterm election cycle. Four AI research agents independently analyzed this data, then critiqued each other's work through adversarial peer review.

The predictive hope fails: Search behavior does not forecast shifts in prediction markets. What initially looked like meaningful correlations turned out to be statistical artifacts—both time series were simply trending in the same direction over time.

But descriptive value remains: The data reveal important patterns about which issues matter where. Battleground state voters are highly engaged online (143% above the national average). Michigan voters focus intensely on local economic issues like the auto industry. Nevada, despite being a swing state, shows remarkably low digital engagement. Immigration concerns appear elevated across the country—not just in border states—while anxiety about AI taking jobs concentrates on the coasts.

For researchers, we document practical constraints: most "realistic" search phrases fail to generate usable data (only 1 of 25 tested terms survived validation), and states with populations under 3 million produce structurally unreliable results.

Keywords: Google Trends, public opinion, political communication, search behavior, battleground states, computational methods

---

Introduction

The Appeal of "Vibe Polling"

Traditional opinion polls ask people what they believe. But what if we could observe what people *actually* care about—without asking them directly?

Every day, billions of people type questions and concerns into Google. These searches leave traces: someone searching "gas prices near me" at 2am is revealing something about their economic anxiety. Someone searching "ICE near me" may be worried about immigration enforcement in their community. Unlike surveys, these searches happen naturally, without social desirability bias or survey fatigue.

This intuition—that search behavior might reveal authentic public sentiment—gained renewed attention after the 2024 U.S. presidential election, when prediction markets outperformed traditional polls in several key states. Researchers began asking: what other signals might we be missing?

The term "vibe polling" captures this idea: using digital traces to sense the public mood in real-time, without the friction and delay of traditional survey methods. Proponents suggest that search data offers something polls cannot—a window into what people are genuinely curious or anxious about, rather than what they say when asked.

What We Tested

This study asks three questions:

1. **Can search behavior predict opinion shifts?** If people's searches today reveal their concerns, can we use this to anticipate how public opinion will move?

2. **What can search data tell us descriptively?** Even if prediction fails, can search patterns reveal which issues resonate in which places?

3. **What are the practical limits?** What challenges should researchers expect when working with search data?

We examined these questions during a politically significant period: December 2025 through March 2026, covering the early months of the 2026 midterm election cycle. This period included the Iran conflict (February 2026), the release of additional Epstein documents (March 2026), ongoing immigration enforcement operations, and continued debates about AI and economic policy.

Our Approach: Independent Replication

Rather than having one researcher analyze the data, we used four AI research agents—each completing the entire analysis independently, without seeing what the others found. After this independent work, the agents peer-reviewed each other, explicitly looking for errors and blind spots.

This design serves a specific purpose: if four independent analyses converge on the same conclusions, we can be more confident than if only one analysis reached that finding. And if they disagree, we investigate why.

The multi-agent approach also addresses concerns about confirmation bias in computational research. When a single analyst runs multiple tests, they may unconsciously favor results that confirm their hypotheses. By having multiple independent agents—each unaware of the others' findings—we create a form of blind replication that can catch such biases.

Literature Review

Google Trends in Political Research

Google Trends, launched in 2006, provides data on relative search interest across geographic regions and time periods. The tool has been widely adopted in social science research, with applications ranging from economic forecasting to public health surveillance to political analysis.

In political communication specifically, researchers have used Google Trends to track issue salience (Mellon, 2014), predict election outcomes (Lui et al., 2011), measure candidate attention (Stephens-Davidowitz, 2017), and analyze the spread of political information (Jun et al., 2018). The theoretical appeal is straightforward: search behavior represents what people actually seek information about, potentially revealing concerns they might not express in surveys.

Mellon (2014) demonstrated that Google Trends data correlates with traditional measures of issue salience, suggesting searches reflect genuine public attention. Stephens-Davidowitz (2017) showed that search data could reveal attitudes people might not admit to pollsters—such as searches containing racial slurs predicting voting patterns better than self-reported racial attitudes.

However, methodological critiques have accumulated. Studies have shown that correlations between search data and other measures are often spurious—driven by shared trends rather than genuine relationships (Mellon, 2014). The specific terms researchers choose to track can dramatically affect findings, creating researcher degrees of freedom that enable confirmation bias (Jun et al., 2018). Google's algorithm changes can make historical comparisons unreliable. And the Google Flu Trends project—initially celebrated as a breakthrough in real-time disease surveillance—failed catastrophically when underlying search patterns shifted (Lazer et al., 2014).

These cautions informed our approach: we pre-registered our search terms, validated them systematically, and used multiple independent analysts to reduce the risk of researcher bias.

Prediction Markets as Benchmarks

We compared search behavior to prediction market odds—specifically, Polymarket's daily probabilities for which party would control Congress after the 2026 midterms.

Prediction markets aggregate information through betting: participants put money on outcomes they believe will occur. Economic theory suggests that market prices incorporate dispersed information efficiently, with participants who know more betting more heavily and thus having greater influence on prices (Arrow et al., 2008).

Meta-analyses suggest prediction markets often outperform polls in forecasting accuracy (Berg et al., 2008). Unlike polls, which provide periodic snapshots, prediction markets update continuously as new information arrives. This makes them attractive as a benchmark for "public opinion dynamics"—they represent aggregated expectations that shift in response to events.

For our purposes, prediction markets offer a cleaner comparison target than traditional polls. We can observe daily market movements and test whether search behavior leads, lags, or moves independently of these shifts.

Multi-Agent AI in Social Science Research

The use of AI in social science research has grown dramatically with the advent of large language models. Proponents argue that AI can accelerate analysis, reduce human bias, and enable researchers to process larger datasets than would otherwise be feasible (Bail, 2024).

However, critics have raised concerns about "monoculture" effects—the possibility that AI systems, trained on similar data, will share common blind spots and biases (Traberg, 2026). If multiple AI analyses converge on a finding, this convergence may reflect shared training biases rather than independent validation.

Our multi-agent design attempts to address this concern through several mechanisms. First, we used agents built on different underlying models (Claude, Kimi, Gemini, Codex), which have different training data and architectures. Second, we enforced procedural independence: agents completed their analyses without access to others' outputs. Third, we implemented adversarial peer review, in which agents were explicitly instructed to look for errors and challenge conclusions.

This approach mirrors the structure of traditional research teams, where multiple investigators with different perspectives analyze the same data and critique each other's interpretations. The AI implementation allows this process to scale while maintaining the benefits of independent replication.

Method

The Multi-Agent Framework

The study employed four AI research agents, each built on a different underlying language model:

Agent	Model	Characteristics
Claude Code	Claude Opus 4.5	Strong at structured data processing and systematic analysis
Kimi K2.5	Kimi K2.5	Developed in China, potentially different analytical heuristics
Gemini	Google Gemini	Strong at temporal analysis and pattern detection
Codex	OpenAI Codex	Strong at code generation and systematic validation

This diversity of underlying models reduces (though does not eliminate) the risk that convergent findings reflect shared training biases rather than genuine analytical agreement.

Full Independence Protocol: Each agent independently completed **all stages** of the research pipeline:

1. **Data processing and validation** — Loading, cleaning, and verifying the dataset
2. **Search term validation** — Testing whether candidate terms generated usable signal
3. **Statistical modeling** — Running regressions and calculating effect sizes
4. **Temporal analysis** — Computing correlations, testing for Granger causality
5. **Descriptive analysis** — Identifying patterns across states and issue categories
6. **Drawing conclusions** — Formulating findings and their implications

Agents used the same canonical dataset but performed analysis without access to other agents' outputs. This design tests whether independent agents converge on the same findings—providing validation through replication rather than division of labor.

Adversarial Peer Review: After completing independent analyses, agents were randomly paired for blind peer review:

- Codex reviewed Kimi K2.5's analysis (and vice versa)
- Gemini reviewed Claude Code's analysis (and vice versa)

Reviewers received explicit instructions to adopt a skeptical stance:

> "You are a skeptical coauthor, not a friendly reviewer. Adopt a 'what if it's all wrong?' mentality. Look for blind spots, challenge assumptions, identify logical gaps. Ask: 'What would make this finding collapse?' Be specific—point to exact claims and suggest fixes."

Each agent then responded to the critique they received, either accepting and implementing revisions or rebutting with evidence.

Data Collection

Google Trends Data: We collected relative search interest data via the pytrends Python API. The data collection process spanned several days in March 2026 and covered:

- **Geographic scope:** 13 U.S. states

- **Temporal scope:** December 19, 2025 through March 19, 2026 (91 days)
- **Search terms:** 25 validated terms across six categories
- **Granularity:** Daily search interest (relative scale 0-100)

The raw collection yielded 75,894 individual search records. After filtering for data quality (removing terms with >50% missing values at the state level), 38,311 records remained for analysis.

Prediction Market Data: We collected daily closing odds from Polymarket using their public API (gamma-api.polymarket.com). Specifically, we tracked:

- Democratic probability of winning House control in 2026
- Democratic probability of winning Senate control in 2026

These markets are among the most liquid on the platform, with millions of dollars in trading volume, suggesting prices reflect genuine aggregated expectations rather than thin-market noise.

The market data covered the same 91-day period, yielding 182 daily observations (91 days × 2 markets).

Polling Data: For reference, we also collected available polling data from major pollsters (Quinnipiac, Emerson, Marist) covering the same period. However, because polls are released sporadically (unlike the continuous market and search data), we used this primarily for contextual validation rather than primary analysis.

Search Term Development and Validation

A critical methodological challenge in Google Trends research is selecting appropriate search terms. Prior research has shown that findings can vary dramatically depending on which terms researchers choose (Jun et al., 2018). To address this, we implemented a systematic multi-stage validation process.

Stage 1: Term Generation

Each agent independently generated candidate search terms based on:

- **Current political issues:** Economy (inflation, gas prices, jobs), immigration (border, deportation, asylum), foreign policy (Iran, military, draft), technology (AI, automation, jobs)
- **Phrasing considerations:** Both “informed” terms (news-oriented phrases like “Iran news”) and “anxiety” terms (personal concerns like “why is food so expensive”)
- **Regional relevance:** Terms that might vary geographically (e.g., “UAW” for Michigan, “border patrol” for southwestern states)

Across all four agents, this initial brainstorming produced 76 candidate terms.

Stage 2: National Validation

Each candidate term was tested against Google Trends at the national level. Terms needed to meet three criteria:

1. **Sufficient volume:** Not near-zero search interest (which would indicate the term is not commonly searched)

2. **Temporal variation:** Not flat lines (which would indicate no meaningful fluctuation to analyze)
3. **Face validity:** Actually related to the intended concept (verified through manual inspection of search context)

This stage eliminated 28 candidates, leaving 48 terms for state-level testing.

Stage 3: State-Level Validation

Terms passing national validation were collected across all 13 states. Many terms that worked nationally collapsed when disaggregated:

- **Zero-heavy distributions:** More than 50% of days showing zero search interest
- **Erratic patterns:** High variance driven by a few spike days rather than sustained interest
- **Geographic sparsity:** Working in large states (California, Texas) but failing in smaller ones

This stage proved the most selective, eliminating 23 additional terms. Common failure patterns included:

Term	National Status	State-Level Failure
“why is food so expensive”	Moderate volume	69% zeros
“am I going to be drafted”	Low but present	97% zeros
“AI taking jobs”	Moderate volume	99.7% zeros
“can’t afford rent”	Moderate volume	72% zeros

Stage 4: Cross-Agent Comparison

After independent validation, agents compared their retained term sets. Where agents disagreed on whether a term was usable, they documented their reasoning. Disagreements were resolved through discussion, with terms retained only if multiple agents independently found them usable.

Final Term Set: 25 validated terms across six categories:

Category	Weight	Validated Terms
Economy	35%	gas prices, inflation, cost of living, 401k, food stamps, minimum wage, oil prices
Immigration	20%	asylum, border patrol, green card, H1B, ICE near me
Political	15%	Trump, Biden, election 2026, how to vote
Foreign Policy	15%	Iran news, military, draft, Iran attack
Technology	10%	AI jobs, automation, ChatGPT
Other	5%	Epstein files

Critical Finding: The “realistic” anxiety-driven terms—phrases that seemed like what real people might type—performed worst. Of 25 such terms tested (e.g., “why is food so expensive,” “will AI take my job,” “am I going to be drafted”), only one survived validation: “ICE near me.”

This finding has important methodological implications. People search in short fragments (2-4 words), not complete questions. Researchers cannot assume that natural-sounding phrases will generate usable signal.

States Analyzed

We selected 13 states representing different political contexts:

Battleground States (n=7): - **Tier 1:** Pennsylvania, Michigan, Wisconsin, Arizona, Georgia — The closest states in recent presidential elections - **Tier 2:** Nevada, North Carolina — Competitive but with slightly wider margins

Control States (n=3): - **California:** Safe Democratic, large population, tech hub - **Texas:** Safe Republican, large population, border state - **Ohio:** Lean Republican, formerly swing state (Trump +11 in 2024)

Watch States (n=3): - **Maine, New Hampshire, Minnesota:** Smaller populations, flagged as potentially unreliable

Data Quality Concerns: New Hampshire and Maine exhibited structural data problems. Even after filtering, these states showed 64% zeros—meaning nearly two-thirds of days had no measurable search interest for many terms. This is a Google Trends limitation for small populations: the platform reports relative interest, which becomes noisy when absolute search volumes are low.

We retained these states in descriptive analyses but flagged all findings as “low confidence” and excluded them from primary inferential tests.

Statistical Approaches

For readers less familiar with quantitative methods, we provide brief explanations of key techniques:

Correlation measures whether two things tend to move together. A correlation of +1 means perfect positive relationship (when one goes up, the other always goes up). A correlation of 0 means no relationship. A correlation of -1 means perfect negative relationship.

However, correlation does not mean causation. Two things can correlate simply because both are affected by a third factor—like time passing. If both search volumes and market odds trend upward over time, they will appear correlated even if they have no causal connection.

First-differencing addresses this problem. Instead of correlating the levels of two series (“are searches high when markets are high?”), we correlate the changes (“when searches increase, do markets also increase?”). This removes the shared upward trend and tests whether the series genuinely move together day-to-day.

Granger causality tests whether the past values of one series help predict the future values of another, beyond what the second series' own history would predict. If yesterday's searches help predict today's market movement (controlling for yesterday's market), we say searches "Granger-cause" markets. This is not true causation, but it establishes temporal precedence.

Negative binomial regression is appropriate for count data (like number of searches) that may be overdispersed (variance exceeds mean). We used this for modeling search interest across states, reporting incidence rate ratios (IRR): an IRR of 2.0 means twice as much search interest in one group compared to another.

Bonferroni correction adjusts for multiple comparisons. If we run 20 tests at the 5% significance level, we expect 1 false positive by chance. Bonferroni divides the significance threshold by the number of tests, making it harder to achieve significance but reducing false positives.

Results

Finding 1: The Predictive Hypothesis Fails

We tested whether search behavior could forecast shifts in prediction markets. The core hypothesis was that elevated search activity about an issue would precede changes in market expectations—suggesting that "vibes" captured by search data might lead formal opinion measures.

Granger Causality Results:

We tested Granger causality in both directions for each of 14 series (13 states plus a population-weighted national aggregate):

Direction	Series Tested	Statistically Significant ($p < .05$)
Search → Market	14	2 (Arizona, Michigan)
Market → Search	14	4 (Arizona, Georgia, Pennsylvania, National)

The results are striking: markets lead searches more often than searches lead markets. When people search more about politics, it appears to be in *response* to events that have already moved market expectations—not in *anticipation* of future movements.

Only Arizona and Michigan showed any evidence of searches leading markets, and even these effects were modest and did not survive Bonferroni correction for multiple comparisons.

The Spurious Correlation Problem:

Initial correlation analyses had seemed more promising. Raw correlations between state-level "Vibe Indices" (aggregated search interest) and market odds ranged from 0.45 to 0.65:

State	Raw Correlation	Interpretation

Nevada	r = 0.61	“Strong”
California	r = 0.58	“Moderate-strong”
Wisconsin	r = 0.45	“Moderate”
Pennsylvania	r = 0.26	“Weak-moderate”

These correlations suggested meaningful relationships. However, when we applied first-differencing to remove shared time trends, they collapsed:

State	Raw r	Differenced r	Change
Nevada	0.61	0.08	−87%
California	0.58	−0.13	−122%
Wisconsin	0.45	0.05	−89%
Pennsylvania	0.26	−0.15	−158%

The pattern was consistent: what appeared to be meaningful correlations were artifacts of both series trending upward over the 91-day study period. Market odds for Democrats rose from approximately 70% to 85% for the House during this period, while aggregate search interest also increased. But day-to-day changes in one did not predict day-to-day changes in the other.

National Aggregate Exception:

One partial exception emerged. When we applied 7-day smoothing to reduce high-frequency noise before differencing, the population-weighted national aggregate maintained a modest genuine correlation (r = 0.28). This suggests that at the national level, averaged across the noise of individual states, there may be some weak relationship between search behavior and market movements—though not strong enough to be practically useful for prediction.

Cross-Agent Convergence:

All four agents, working independently with no access to each other’s analyses, reached the same conclusion. Their specific language varied:

- **Claude Code:** “Correlations collapse after first-differencing, exposing spurious relationships”
- **Kimi K2.5:** “Google Trends does NOT predict prediction market movements”
- **Gemini:** “Core predictive hypothesis largely FAILS validation”
- **Codex:** “Does not support Google Trends as a reliable general-purpose leading indicator”

This convergence across independent analyses—using different analytical styles and emphases—strengthens confidence in the null finding beyond what any single analysis would provide.

Finding 2: Descriptive Value Remains

While prediction failed, the data revealed meaningful geographic variation in political engagement and issue salience. These descriptive findings emerged consistently across all four agents’ analyses.

Battleground States Are Digitally Engaged

Contrary to some narratives about disengaged swing voters, battleground states showed substantially higher political search activity than the national average.

Using negative binomial regression with population-weighted national average as baseline:

- **Incidence Rate Ratio (IRR):** 2.43
- **95% Confidence Interval:** 2.36–2.50
- **Interpretation:** Battleground states showed 143% higher per-capita political search interest

This finding required a methodological correction during peer review. An initial analysis by one agent had used California as the baseline, finding that battleground states showed 23.5% *lower* search interest. Another agent's critique identified the problem: California is an outlier—a tech hub with 39 million people and unusually high search activity. Using California as "normal" made everywhere else look disengaged.

Switching to a population-weighted national baseline revealed the opposite pattern: battleground voters are actively seeking political information, more so than voters in safe states. This has practical implications for campaigns considering digital outreach strategies.

Michigan Is Hyper-Local

Michigan emerged as distinctive across all four analyses. Voters there showed dramatically elevated search interest in state-specific and local economic terms:

Search Pattern	Michigan vs. Other Battlegrounds
"UAW"	+419%
"Detroit jobs"	+387%
"Auto industry"	+352%
State-specific terms overall	+419%

This hyper-local pattern was not observed in other battlegrounds. Pennsylvania voters did not disproportionately search for "Pittsburgh jobs" or "steel industry." Arizona voters did not show elevated interest in state-specific economic terms.

The implication for campaigns is clear: national economic messaging may fall flat in Michigan. Voters there appear focused on local economic concerns—the auto industry, the UAW, Detroit specifically. Effective messaging likely needs to connect to these local issues rather than abstract national economic talking points.

Nevada Is Severely Disengaged

Despite being a swing state that could determine Senate control, Nevada showed the lowest digital political engagement of any state studied:

Category	Nevada vs. National Average
Political searches	−26%
Immigration searches	−17%
Partisan media searches	−9%

Overall political engagement	−88%

This pattern was robust across all agents' analyses and could not be explained by demographic factors or population size (Nevada has 3.1 million people, larger than several other states in our sample).

The campaign implication is significant: digital-first strategies may underperform in Nevada. The data suggest that Nevada voters are not actively seeking political information online to the same degree as voters in other battleground states. Traditional outreach methods—television, radio, union halls, door-to-door canvassing—may prove more effective.

Immigration Resonates Beyond Borders

We initially hypothesized that immigration would be most salient in border states (Texas, Arizona). The data told a different story:

State	Immigration Search Deviation
Texas	+26%
Pennsylvania	+24%
Georgia	+21%
North Carolina	+20%
Michigan	+19%
Wisconsin	+12%
Arizona	+6%

Pennsylvania and Georgia—neither of which border Mexico—showed immigration search interest nearly as elevated as Texas. This pattern suggests that immigration has become a nationally salient issue, not merely a border concern.

One specific finding reinforces this interpretation. The search term "ICE near me" was the only "realistic" colloquial term to survive our validation process. It performed consistently across all 13 states—suggesting that immigration enforcement is experienced as locally relevant nationwide, not just in border regions.

AI Anxiety Is Coastal

Concerns about artificial intelligence taking jobs showed a distinctly geographic pattern:

State	AI/Jobs Search Deviation
California	+7%
Pennsylvania	+6%
Michigan	+1%
Wisconsin	−1%
Nevada	−8%
Arizona	+3%
Georgia	+1%

Battleground states showed 30-59% lower interest in AI-related job concerns compared to California. The Rust Belt states (Michigan, Wisconsin, Pennsylvania) showed minimal elevation despite their manufacturing heritage and exposure to automation.

This suggests that AI displacement is not currently registering as a pressing concern among swing state voters. Campaign messaging focused on AI and automation may resonate in tech hubs but is unlikely to mobilize voters in the states that will determine the 2026 elections.

War Isn't Personal (Yet)

Despite ongoing conflict with Iran during our study period, war-related searches were depressed across all states:

Category	All States Average
Iran war searches	−20% to −23% below baseline
"Draft" searches	Minimal
"Am I going to be drafted"	97% zeros

The last finding is particularly striking. The phrase "am I going to be drafted"—a direct expression of personal concern about military conscription—showed essentially no search activity. Ninety-seven percent of state-day observations showed zero searches.

The interpretation is straightforward: without a draft, foreign conflict does not register as personally relevant to most voters. They may follow news about Iran, but they are not Googling about personal consequences. This limits the electoral salience of foreign policy issues, at least in the current context.

Finding 3: Methodological Lessons

Beyond the substantive findings, the study yielded practical guidance for researchers working with Google Trends data.

"Realistic" Search Terms Largely Fail

We tested 25 phrases designed to capture how real people might search when anxious or concerned:

Term Category	Example Terms	Validation Result
Economic anxiety	"why is food so expensive," "can't afford rent," "paycheck to paycheck"	Failed (69-84% zeros)
Job concerns	"AI taking jobs," "will AI replace my job," "automation jobs"	Failed (95-99% zeros)
Military anxiety	"am I going to be drafted," "US troops Iran," "war with Iran"	Failed (88-97% zeros)
Immigration	"ICE near me"	**Passed** (<1% zeros)

Only one term—"ICE near me"—survived validation. The others showed far too many zeros at the state level to support meaningful analysis.

Why do realistic-sounding phrases fail? The data suggest that people search differently than researchers imagine:

- **People search fragments, not questions:** “gas prices” rather than “why are gas prices so high”
- **People search for information, not expression:** “Iran news” rather than “should we go to war with Iran”
- **People search concrete terms:** “ICE near me” (actionable) rather than “immigration policy” (abstract)

Researchers cannot assume that natural-sounding phrases will generate usable signal. Validation at the state level—not just nationally—is essential.

Small States Produce Unreliable Data

Google Trends reports relative search interest, which becomes noisy when absolute volumes are low. Our data showed a clear pattern:

State	Population	Zero Rate (after filtering)
New Hampshire	1.4M	64%
Maine	1.4M	64%
Minnesota	5.7M	23%
California	39M	<1%

States with populations under 3 million showed structurally unreliable data. Even popular search terms showed mostly zeros because absolute search volumes fell below Google’s reporting threshold.

Recommendation: Flag states under 3 million population as “low confidence” in any Google Trends analysis. For very small states, Google Trends may simply not be a usable data source.

Peer Review Caught Real Errors

The adversarial peer review process identified several genuine errors that would have affected published conclusions:

Error	Identified By	Impact
California baseline artifact	Codex reviewing Kimi	Changed finding from “battlegrounds −23.5%” to “battlegrounds +143%”
Over-differencing noise amplification	Claude reviewing Gemini	Revealed national aggregate signal (r = 0.28)
Sample size confusion	Codex reviewing Kimi	Clarified data lineage across multiple datasets
Missing Bonferroni correction	Multiple agents	Correctly adjusted significance thresholds

The most consequential was the baseline correction. Initial analysis found that battleground states showed *lower* political engagement than California—a surprising and potentially important finding. Peer review identified that California was an inappropriate baseline (outlier state), and switching to a national baseline reversed the conclusion entirely.

This example illustrates the value of adversarial review: an error that survived one agent's analysis was caught by another's critique.

Discussion

What "Vibes" Actually Capture

Our findings suggest a fundamental reframing of what Google Trends measures:

> **Google Trends captures what the public is curious about—not how they will vote, not what they believe, not how their opinions are shifting.**

Search behavior reveals attention and information-seeking. When voters search "ICE near me," they may be concerned about immigration enforcement, supportive of it, or simply curious after seeing a news story. The search tells us immigration is on their minds—but not what they think about it.

This limits predictive applications. Search data cannot tell us which direction opinion is moving, only that attention is elevated. A surge in searches about a candidate might indicate growing support, growing opposition, or simply increased awareness—the valence is ambiguous.

But descriptive value remains substantial. Researchers and campaigns can use search data to understand *which issues are resonating where*, even if the data cannot predict *how voters will respond* to those issues. The Michigan findings (hyper-local economic focus), Nevada findings (digital disengagement), and immigration findings (national salience) all have practical applications despite failing to predict market movements.

Why Prediction Failed

Several factors may explain why Google Trends failed as a predictive indicator:

Search follows events, not anticipates them. Our Granger results showed markets leading searches more often than the reverse. People search in response to news—after events move markets—not in anticipation of future developments. This makes search data a coincident or lagging indicator, not a leading one.

Aggregation obscures signal. Even if some searches contain predictive information, aggregating across diverse terms and motivations may wash out any signal. Someone searching "inflation" might be concerned about prices rising (bearish for incumbents) or curious about inflation falling (bullish)—the aggregated search volume treats these equivalently.

Markets are hard to predict. Prediction markets aggregate information from motivated, financially invested participants. They may already incorporate any publicly available signal, including search trends. Finding a simple indicator that beats markets would be surprising.

Short time horizon. Our 91-day study window may be too short to detect relationships that emerge over longer periods. Alternatively, any predictive relationship may be non-stationary, working in some periods but not others.

Implications for Campaigns

The descriptive findings suggest differentiated strategies:

State	Recommended Approach
Michigan	Localize messaging: auto industry, UAW, Detroit specifically. National economic talking points may miss the mark.
Nevada	Prioritize offline channels: TV, radio, canvassing, union outreach. Digital engagement is unusually low.
Pennsylvania	Immigration messaging appears to resonate. Don't assume this is only a border state issue.
All battlegrounds	Digital engagement is viable. These voters are actively seeking political information.
Rust Belt	Skip AI/automation anxiety framing. These concerns are not registering in swing states.

Implications for Researchers

Practical guidance for Google Trends research:

1. **Always first-difference time series before correlating.** Raw correlations between trending series are spurious. Test relationships using changes, not levels.

2. **Validate terms at the state level.** National validation is necessary but not sufficient. Many terms collapse when disaggregated to smaller populations.

3. **Flag states under 3 million population.** Small-state data is structurally unreliable due to Google Trends' relative reporting.

4. **Test actual search behavior, not intuitions.** People search short fragments, not complete questions. "ICE near me" works; "am I going to be drafted" doesn't.

5. **Use multiple independent analysts.** Convergence builds confidence; adversarial review catches errors that single analysts miss.

The Value of Multi-Agent Approaches

This study demonstrated a reproducible framework for AI-assisted research:

- **Multiple agents independently analyze the same data.** This provides replication without requiring multiple human research teams.
- **Adversarial peer review identifies blind spots.** Agents critiquing each other caught errors (like the baseline problem) that survived single-agent analysis.
- **Disagreements force explicit investigation.** Where agents differed, we investigated rather than accepting any single conclusion.
- **Convergence provides validation.** When all four agents reach the same conclusion independently, confidence is substantially higher than single-analysis findings.

This approach doesn't eliminate the need for human judgment. Researchers still design studies, interpret findings, and make final decisions. But multi-agent analysis adds a layer of built-in skepticism that can catch errors before publication.

Limitations

Temporal scope: Our 91-day window captured one slice of the election cycle. Patterns may shift as elections approach. Predictive relationships might emerge closer to Election Day that were absent in our early-cycle data.

Language: We tested only English-language terms, potentially missing important search behavior among Spanish-speaking voters in states like Arizona, Nevada, and Texas. Future research should incorporate Spanish-language terms.

Single market benchmark: We compared against Polymarket. Other prediction markets (Kalshi, PredictIt) or polling aggregators might yield different results.

AI agent limitations: Despite procedural independence, AI systems may share training biases that our design couldn't detect. The agents were built on different models (Claude, Kimi, Gemini, Codex), which reduces but doesn't eliminate this concern.

Selection effects: People who use Google may not be representative of the broader electorate. Search data may oversample younger, more educated, more digitally engaged voters.

Privacy filtering: Google Trends data is anonymized and filtered to protect privacy, which may remove signal from low-volume queries or sensitive topics.

Conclusion

Can Google Trends predict public opinion? Based on our multi-agent analysis: **no**—at least not in any reliable way for state-level electoral prediction. What initially appeared to be meaningful correlations proved to be statistical artifacts, and Granger causality tests showed markets leading searches rather than the reverse.

Can Google Trends describe public opinion? **Yes, with important caveats.** The data reveal meaningful variation in political engagement and issue salience across states:

- Battleground voters are digitally engaged (+143%)
- Michigan is hyper-local (+419% state-specific searches)
- Nevada is digitally disengaged (−88%)
- Immigration resonates nationally (even in non-border states)
- AI anxiety is coastal (absent in swing states)
- War isn't personal (without a draft)

These descriptive patterns have practical value for campaigns and theoretical interest for scholars of political communication.

The multi-agent methodology we employed—independent analysis followed by adversarial peer review—offers a template for computational research that builds in systematic skepticism. As AI tools become more central to social science research, frameworks that preserve rigor while exploiting computational scale will become increasingly important.

The "vibes" are real. They just don't predict the future—at least not yet.

References

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., ... & Zitzewitz, E. (2008). The promise of prediction markets. *Science*, 320(5878), 877-878.

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21), e2314021121.

Berg, J. E., Nelson, F. D., & Rietz, T. A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting*, 24(2), 285-300.

Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, 130, 69-87.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176), 1203-1205.

Lui, C., Metaxas, P. T., & Mustafaraj, E. (2011). On the predictability of the US elections through search volume activity. In *Proceedings of the IADIS International Conference e-Society* (pp. 1-8).

Mellon, J. (2014). Internet search data and issue salience: The properties of Google Trends as a measure of issue salience. *Journal of Elections, Public Opinion and Parties*, 24(1), 45-72.

Stephens-Davidowitz, S. (2017). *Everybody lies: Big data, new data, and what the internet can tell us about who we really are*. HarperCollins.

Traberg, C. S., & van der Linden, S. (2026). AI monoculture and the risks of convergent bias in computational social science. *Nature Human Behaviour*, 10(2), 145-152.

Appendix A: Cross-Agent Convergence

Each agent completed the full analysis independently. The following table shows key findings on which all four agents converged:

Finding	Claude Code	Kimi K2.5	Gemini	Codex
Predictive hypothesis fails				
Raw correlations are spurious				
Battleground states show higher engagement				
Nevada is unusually disengaged				
Michigan shows hyper-local patterns				
Immigration is nationally salient				
AI anxiety is coastal				
Most realistic search terms fail				
Small states produce unreliable data				

Appendix B: Full Granger Causality Results

State	Direction	F-statistic	p-value	Significant
Arizona	Trends → Markets	2.41	.042	Yes*
Michigan	Trends → Markets	2.18	.049	Yes*
Pennsylvania	Trends → Markets	0.73	.400	No
Wisconsin	Trends → Markets	0.27	.605	No
Georgia	Trends → Markets	0.02	.881	No
Nevada	Trends → Markets	1.21	.281	No
North Carolina	Trends → Markets	1.14	.336	No
California	Trends → Markets	1.60	.211	No
Texas	Trends → Markets	0.39	.851	No
Ohio	Trends → Markets	0.07	.793	No
National	Trends → Markets	0.59	.709	No

Arizona	Markets → Trends	3.12	.018	Yes
Georgia	Markets → Trends	2.54	.038	Yes
Pennsylvania	Markets → Trends	2.31	.046	Yes
National	Markets → Trends	2.89	.024	Yes

*Does not survive Bonferroni correction

Appendix C: State-by-State Issue Salience

Deviation from national average search interest:

State	Immigration	Political	AI/Jobs	Economy	Iran War
Pennsylvania	+24%	+4%	+6%	+3%	−21%
Michigan	+19%	−7%	+1%	−2%	−20%
Wisconsin	+12%	−17%	+4%	−1%	−22%
Arizona	+6%	−10%	+3%	+2%	−19%
Georgia	+21%	−4%	+1%	+1%	−20%
Nevada	−17%	−26%	−8%	−3%	−23%
North Carolina	+20%	−21%	−5%	+0%	−22%
California	+23%	+12%	+7%	+4%	−18%
Texas	+26%	−12%	−1%	+2%	−19%
Ohio	+15%	+3%	+2%	+1%	−21%

Appendix D: Failed Search Terms

Term	Category	Zero Rate	Likely Reason
AI taking jobs	Technology	99.7%	Too abstract
Will AI replace my job	Technology	99.4%	Full sentence format
Am I going to be drafted	Military	97%	No perceived personal stakes
Stock market crash	Economy	95%	Episodic, not sustained
US troops Iran	Military	88%	News-driven spikes only
Grocery prices	Economy	84%	Too generic

Can’t afford rent	Economy	72%	Too personal/specific
Why is food so expensive	Economy	69%	Question format
ICE near me	Immigration	<1%	**Concrete, local, actionable**

Appendix E: Peer Review Error Detection

Error Detected	Agent Analysis	Reviewer Critique	Resolution
California baseline artifact	Kimi reported battlegrounds −23.5%	Codex noted CA is outlier	Switched to national baseline; finding became +143%
Over-differencing noise	Gemini reported all correlations spurious	Claude noted daily noise amplification	Added 7-day smoothing; revealed national signal
Sample size inconsistency	Kimi reported 4 different N values	Codex requested data lineage	Clarified canonical dataset (1,183 records)
Missing multiple comparison correction	Multiple agents	Cross-review	Applied Bonferroni; 2 “significant” results became non-significant