Multi-Agent Computational Triangulation: A Methodological Framework for AI-Assisted Social Science Research

A Case Study from the VibePoll-2026 Project

AgentAcademy Team

Abstract

As artificial intelligence tools become central to computational social science, questions arise about how to maintain rigor while exploiting scale. This paper presents a methodological framework—Multi-Agent Computational Triangulation (MACT)—developed and tested through the VibePoll-2026 study, which investigated whether Google Trends data could measure public opinion in U.S. battleground states.

The MACT framework employs multiple AI research agents, each independently completing the full research pipeline on shared data, followed by adversarial peer review between agents. This design operationalizes replication-based validation within a single study, catching errors that survived individual agent analysis while building confidence through cross-agent convergence.

We document the framework's implementation across four agents (Claude Code, Kimi K2.5, Gemini, and Codex), detailing the coordination protocols, independence requirements, peer review procedures, and error handling mechanisms. We report specific errors caught through adversarial review—including a data transformation mistake that reversed a key finding—and extract generalizable lessons for AI-assisted research.

The paper contributes: (1) a reproducible protocol for multi-agent research design, (2) empirical evidence on error detection rates in AI-assisted analysis, (3) a taxonomy of practical skills for computational research agents, and (4) recommendations for integrating AI tools into social science workflows while preserving epistemic standards.

Keywords: multi-agent AI, computational social science, research methodology, replication, peer review, triangulation

---

1. Introduction

1.1 The Promise and Peril of AI in Social Science Research

The rapid advancement of large language models (LLMs) has created new possibilities for social science research. AI systems can now process large datasets, generate statistical analyses, write code, and even draft interpretations—tasks that previously required substantial human effort (Bail, 2024). Proponents argue that AI can accelerate research, reduce human bias, and enable analyses at scales previously impractical.

However, these capabilities raise methodological concerns. Critics worry about “monoculture” effects, where AI systems trained on similar data share common blind spots (Traberg & van der Linden, 2026). There are concerns about sycophantic tendencies, where models confirm researcher expectations rather than challenging them (Asher et al., 2026). And fundamental questions remain about accountability: when an AI system makes an analytical error, how is it caught? When it reaches a conclusion, how is it validated?

These concerns are not merely theoretical. In traditional research, replication provides a check on findings—other researchers, using independent judgment, attempt to reproduce results. Peer review provides another layer, with experts scrutinizing methods and conclusions. But when AI conducts analysis, these safeguards may be weakened. A single AI system analyzing data may produce internally consistent but fundamentally flawed conclusions, and the speed of AI-assisted research may outpace traditional validation mechanisms.

1.2 The Case for Multi-Agent Approaches

This paper proposes and evaluates a methodological framework designed to address these concerns: Multi-Agent Computational Triangulation (MACT). The core insight is simple: if single-agent AI analysis risks undetected errors, multiple independent agents analyzing the same data can provide built-in replication and error detection.

The framework draws on established principles from research methodology:

Triangulation: Using multiple methods or data sources to study the same phenomenon increases confidence when findings converge and reveals problems when they diverge (Denzin, 1978).

Adversarial collaboration: Researchers with opposing views working together can produce more robust findings than either would alone (Mellers et al., 2001).

Replication: Reproducing findings with independent analysis is the gold standard for scientific validation (Open Science Collaboration, 2015).

MACT operationalizes these principles for AI-assisted research by requiring multiple agents to independently complete full analyses, then subjecting each agent’s work to adversarial review by another agent.

1.3 Research Questions

We developed and tested the MACT framework through the VibePoll-2026 study, which investigated whether Google Trends search data could measure public opinion in U.S. battleground states. This

substantive research question provided a realistic context for evaluating the methodology.

This paper addresses three methodological questions:

MQ1: How should multi-agent AI research be structured to maximize error detection while maintaining independence?

MQ2: What types of errors does adversarial inter-agent review catch that single-agent analysis misses?

MQ3: What practical skills and protocols should guide AI agents in computational social science research?

1.4 Paper Overview

Section 2 reviews relevant literature on AI in research, triangulation, and replication. Section 3 describes the MACT framework in detail, including agent selection, independence protocols, and peer review procedures. Section 4 presents results from the VibePoll-2026 implementation, documenting errors caught and lessons learned. Section 5 synthesizes practical skills extracted from agent reflections. Section 6 proposes a standard workflow for future studies. Section 7 discusses limitations and future directions.

---

2. Literature Review

2.1 AI Tools in Social Science Research

The integration of AI into social science research has accelerated dramatically since 2023. Studies have demonstrated AI capabilities in literature review (Elicit, 2023), qualitative coding (Xiao et al., 2023), survey design (Argyle et al., 2023), and statistical analysis (Bail, 2024).

Bail (2024) provides a systematic evaluation of AI capabilities for social science tasks, finding that current LLMs can competently perform many analytical tasks but struggle with novel methodological decisions and contextual interpretation. The study identifies a key tension: AI tools are most useful for routine tasks but least reliable for the judgment-intensive work that most affects research quality.

Several frameworks have emerged for AI-assisted research. The DAAF (Data Analyst Augmentation Framework) provides structured protocols for AI-assisted quantitative analysis (DAAF Community, 2024). CommDAAF extends this framework for communication research specifically, adding domain-specific guardrails (AgentAcademy, 2026). These frameworks share an emphasis on human oversight, explicit documentation, and methodological transparency.

2.2 Concerns About AI Monoculture

Traberg and van der Linden (2026) articulate concerns about “AI monoculture” in research. Because major LLMs are trained on overlapping datasets and share architectural similarities, they may

systematically share biases. Multiple AI analyses that appear to "replicate" each other may actually reflect shared training artifacts rather than genuine convergent evidence.

This concern motivates our use of agents built on different underlying models (Claude, Kimi, Gemini, Codex). While these models are not fully independent—they share training data overlap and similar architectural choices—they differ enough that convergent findings provide stronger evidence than single-model analysis.

Empirical evidence on AI error patterns remains limited. Asher et al. (2026) document sycophantic tendencies in LLMs, where models agree with user assertions even when incorrect. This finding suggests that AI systems may be poor at self-correction, strengthening the case for external validation through multi-agent approaches.

2.3 Triangulation and Mixed Methods

Triangulation—using multiple approaches to study the same phenomenon—has a long history in social science methodology (Denzin, 1978; Jick, 1979). The logic is that different methods have different weaknesses; when multiple methods converge on the same finding, confidence increases.

Methodologists distinguish several types of triangulation:

- **Data triangulation:** Using multiple data sources
- **Investigator triangulation:** Using multiple researchers
- **Theory triangulation:** Using multiple theoretical perspectives
- **Methodological triangulation:** Using multiple methods

Our MACT framework operationalizes investigator triangulation with AI agents as "investigators." Each agent brings different analytical tendencies, and their independence is enforced through procedural controls.

2.4 Adversarial Collaboration

Adversarial collaboration brings together researchers with opposing views to jointly design and interpret studies (Mellers et al., 2001). The approach has been used to resolve longstanding debates by forcing explicit specification of predictions and joint agreement on methods before data collection.

We adapt this concept for inter-agent review. After independent analysis, agents are paired for adversarial peer review, with explicit instructions to adopt a "what if it's all wrong?" mentality. This differs from traditional peer review in that reviewers have full access to data and code, not just manuscripts, and are explicitly tasked with finding errors rather than providing collegial feedback.

2.5 Replication in Computational Research

The replication crisis in social science has highlighted the importance of reproducible research practices (Open Science Collaboration, 2015). Computational research offers both opportunities and challenges for replication: analyses can be exactly reproduced given code and data, but the complexity of pipelines can introduce subtle errors.

Multi-agent approaches offer a form of “built-in replication.” When multiple agents independently analyze the same data and reach the same conclusions, this provides evidence of robustness. When they diverge, this signals the need for investigation—either one agent erred, or the finding is sensitive to analytical choices.

---

3. The MACT Framework

3.1 Overview

Multi-Agent Computational Triangulation (MACT) structures AI-assisted research around three core principles:

1. **Full independence:** Each agent completes the entire research pipeline independently, without access to other agents’ outputs during primary analysis.

2. **Adversarial review:** After independent analysis, agents peer-review each other’s work with explicit instructions to find errors and challenge conclusions.

3. **Transparent synthesis:** Disagreements are documented and resolved through explicit reconciliation, with the final synthesis clearly attributing which findings came from which agents.

The framework is designed to catch errors that single-agent analysis misses while building confidence through convergent findings.

3.2 Agent Selection

The VibePoll-2026 study employed four AI research agents:

Agent	Underlying Model	Selection Rationale
Claude Code	Claude Opus 4.5	Strong structured reasoning, extensive training on research methods
Kimi K2.5	Kimi K2.5	Different training corpus (Chinese-developed), statistical emphasis
Gemini	Google Gemini	Different architecture, strong temporal analysis capabilities
Codex	OpenAI Codex	Code-generation focus, systematic validation orientation

Agent selection balanced several considerations:

Model diversity: Using agents built on different underlying models reduces the risk that convergent findings reflect shared training biases. While Claude, Gemini, and Codex share some training data, Kimi’s Chinese-language training provides partial independence.

Capability matching: Agents were selected for capabilities relevant to the research task. All agents could perform standard statistical analyses, but each brought somewhat different strengths (e.g., Gemini’s temporal analysis, Codex’s systematic validation).

Practical constraints: Agents needed to be accessible through available interfaces (API, CLI) and capable of following complex multi-step instructions.

3.3 Independence Protocol

Independence is crucial for multi-agent approaches. If agents access each other's outputs during primary analysis, they may converge through contamination rather than genuine agreement.

The VibePoll-2026 study enforced independence through several mechanisms:

Separate workspaces: Each agent operated in a dedicated directory (`agents/claude-code/`, `agents/kimi-k2.5/`, etc.) with no read access to other agents' directories during primary analysis.

Shared data only: Agents accessed a common canonical dataset (`data/processed/`) but performed all analysis independently. This ensures that convergent findings reflect analytical agreement, not data differences.

Sequential coordination: A human coordinator managed agent interactions, ensuring that cross-agent information flow occurred only at designated checkpoints (after primary analysis, during peer review).

Explicit instructions: Each agent received clear instructions that they should reach independent conclusions without referencing other agents' work. Instructions stated: "Complete all stages of analysis independently. Do not access other agents' outputs until peer review phase."

3.4 Full-Pipeline Requirement

A critical design decision was requiring each agent to complete the *entire* research pipeline, not just assigned subtasks. Traditional division of labor (e.g., one agent collects data, another analyzes) prevents the error-detection benefits of independent replication.

Each agent was required to independently perform:

1. **Data validation:** Verifying data structure, quality, and completeness
2. **Search term validation:** Testing whether candidate terms produced usable signal
3. **Statistical modeling:** Running regressions, computing effect sizes
4. **Temporal analysis:** Computing correlations, testing Granger causality
5. **Descriptive analysis:** Identifying patterns across states and issues
6. **Interpretation:** Drawing conclusions and identifying limitations

This full-pipeline requirement means that errors at any stage have multiple opportunities for detection. If one agent makes a data processing error, other agents' different results will reveal the discrepancy.

3.5 Adversarial Peer Review Protocol

After completing independent analyses, agents were randomly paired for peer review:

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

Reviewers received explicit adversarial instructions:

> "You are a skeptical coauthor, not a friendly reviewer. Adopt a new pair of eyes—pretend you are seeing this work for the first time, with no prior context or investment in the findings. You have no stake in the conclusions being 'right.' Your only loyalty is to the truth.
>
> Ask yourself: - What if this is all wrong? - What would falsify this finding? - What alternative explanation was dismissed too quickly? - What would a hostile reviewer at a top journal say?"

Reviews were structured around five categories:

1. **Major concerns:** Issues that could invalidate findings
2. **Methodological questions:** Analytical choices requiring justification
3. **Blind spots:** What was not considered
4. **Logical gaps:** Where evidence doesn't support claims
5. **Suggested revisions:** Specific, actionable fixes

After receiving critiques, each agent wrote a formal response, categorizing each point as:

- **Accepted:** Revision made with documentation
- **Partially accepted:** Modified response with explanation
- **Rebutted:** Defended original approach with evidence

3.6 Synthesis and Reconciliation

The final phase involved synthesizing findings across agents. This was performed by the coordinating agent (the main session) with explicit attention to:

Convergent findings: Conclusions reached independently by multiple agents. These receive highest confidence.

Divergent findings: Conclusions where agents disagreed. These required investigation to determine whether disagreement reflected error (one agent wrong) or sensitivity (finding depends on analytical choices).

Attribution: The final paper clearly indicated which findings came from which agents and whether they represented convergent or single-agent conclusions.

---

4. Implementation Results: The VibePoll-2026 Study

4.1 Study Overview

The VibePoll-2026 study tested whether Google Trends search data could predict or describe public opinion in U.S. battleground states during the 2026 midterm election cycle. The study collected 38,311 search records across 13 states and 25 validated search terms over 91 days, comparing search behavior to prediction market odds.

The substantive findings are reported elsewhere (AgentAcademy, 2026). Here we focus on methodological outcomes: what the multi-agent approach revealed about the research process itself.

4.2 The Coordination Process in Practice

Before detailing the errors caught, it is instructive to describe how coordination actually unfolded during the VibePoll-2026 study.

Day 1: Setup and Initial Analysis

The coordinator created the study plan (PLAN.md) with explicit CommDAAF guardrails, specifying that all agents must complete full independent analysis. Agent workspaces were created with clear separation. The canonical dataset was placed in a shared location, and each agent received identical instructions emphasizing independence.

Initial analysis proceeded in parallel. Claude Code focused on data processing and validation, establishing the canonical dataset structure. Kimi K2.5 began statistical modeling using the processed data. Gemini initiated temporal analysis with correlation and Granger causality tests. Codex focused on validating search terms, testing which “realistic” phrases actually generated usable signal.

Day 2: Independent Analysis Completion

Agents completed their primary analyses. At this stage, no cross-agent communication had occurred—each agent operated in isolation. The coordinator monitored progress through workspace artifacts (output files, logs) without sharing information across agents.

Several early problems emerged that agents corrected independently: - Claude Code discovered market data from 2020 mixed with 2025-2026 trends and flagged the mismatch - Kimi K2.5 identified that the initial Pennsylvania-only diagnostics couldn’t generalize and expanded analysis - Gemini encountered API rate limiting and implemented exponential backoff - Codex found that national term validation didn’t predict state-level usability and added a second validation stage

These independent corrections demonstrated that agents could identify and fix certain problems without external intervention. However, more subtle errors remained undetected.

Day 3: Peer Review

The coordinator assigned peer review pairings and distributed instructions emphasizing adversarial critique. Each agent received full access to their assigned partner’s workspace—all files, not just summary reports.

The review process was intensive. Agents were instructed to: - Read across documents rather than within single documents - Check claims against source data tables - Look for instability in estimands, sample

sizes, and baselines - Identify conclusions that exceeded evidence

This cross-document review approach proved crucial for error detection. Many errors were inconsistencies between documents (e.g., different N values, conflicting baselines) that wouldn't be visible within any single document.

Day 3-4: Response and Revision

After receiving critiques, agents wrote formal responses. This process forced explicit engagement with each criticism:

Acceptance example (Claude Code responding to Gemini): > "You are completely correct that dividing a 0-100 proportional index by population is mathematically invalid. I have removed all per-capita calculations and will rerun analysis using the raw interest values."

Rebuttal example (Gemini responding to Claude Code): > "Regarding weekly aggregation: We cannot aggregate to weekly grain because the total temporal window (60 weeks) would provide insufficient N for Granger causality with appropriate lag structure. The 7-day rolling average achieves the smoothing you requested while preserving daily resolution for statistical power."

This response process ensured that critiques received substantive engagement rather than defensive dismissal.

Day 4-5: Synthesis

The coordinator synthesized findings across agents, creating the final paper and this methodological report. Key synthesis tasks included: - Identifying convergent findings (all agents agreed) - Investigating divergent findings (determining error vs. sensitivity) - Creating explicit attribution (which findings from which agents) - Documenting unresolved disagreements as limitations

4.3 Errors Caught by Adversarial Review

The adversarial peer review process identified five significant errors that had survived individual agent analysis:

Error 1: Per-Capita Normalization of Already-Normalized Data

Agent: Claude Code
Caught by: Gemini
Impact: Critical (reversed key finding)

Claude Code divided Google Trends interest scores (already a 0-100 proportional index) by state population to create "per-capita" measures. This systematically inflated values for small states and deflated values for large states, producing the artifactual finding that battleground states showed "143% higher engagement."

Gemini's critique identified the mathematical error:

> "By dividing this 0-100 index by the state's population… you are systematically and artificially inflating the values for small-population battlegrounds."

This error would have survived traditional single-analyst review because the calculation seemed rigorous (population controls are standard practice) and produced plausible results. The error was only

caught because Gemini independently analyzed the same data without population division and obtained different results.

Error 2: Over-Differencing Noisy Data

Agent: Gemini
Caught by: Claude Code
Impact: Major (obscured genuine signal)

Gemini first-differenced raw daily search data to test for spurious correlations. However, first-differencing amplifies high-frequency noise, and daily Google Trends data is inherently noisy. The result was that genuine medium-term relationships were obscured.

Claude Code's critique identified the problem:

> "First-differencing daily Google Trends data massively amplifies high-frequency noise, which mathematically suppresses correlations. Running Granger Causality on un-smoothed differenced data likely created a false 'spurious' conclusion."

After accepting this critique, Gemini applied 7-day smoothing before differencing and discovered that the population-weighted national aggregate maintained genuine signal (r = 0.28)—a finding that would have been missed without the correction.

Error 3: Baseline Confusion Without Reconciliation

Agent: Kimi K2.5
Caught by: Codex
Impact: Major (conflicting claims)

Kimi K2.5's reports contained conflicting findings: one document reported battleground states "-23.5% lower" (using California as baseline) while another reported "+143% higher" (using national baseline). The switch was made in response to reviewer instructions but was never explicitly reconciled.

Codex's critique identified the problem:

> "The arithmetic is correct for each framing… but the report never reconciles this sign flip in a way that would satisfy a skeptical reader."

This error illustrates how iterative revision can introduce inconsistencies that the original analyst doesn't notice because they understand the evolution. An external reviewer, seeing only the final documents, spotted the contradiction.

Error 4: Sample Size Chaos

Agent: Kimi K2.5
Caught by: Codex
Impact: Moderate (reproducibility threat)

Kimi K2.5's reports referenced four different sample sizes (10,920 / 75,894 / 17,381 / 1,183) without explaining how they related. This made it impossible to verify which dataset fed which analysis.

Codex's critique:

> "Four different N's with no clear lineage. A hostile reader could interpret this as post-hoc dataset switching."

The error reflected inadequate data lineage documentation—a common problem in iterative analysis where analysts know the data history but don't record it explicitly.

Error 5: Granger Causality Miscount

Agent: Codex
Caught by: Kimi K2.5
Impact: Moderate (overstated null)

Codex reported "0/14 states significant" for Granger causality tests, but the actual data showed 22/60 significant tests. The error arose from relying on study-wide narrative rather than checking source tables.

Kimi K2.5's critique:

> "The actual granger_results.md table shows 22/52 significant tests at p<0.05. That is 42% significance, not 0%."

This error illustrates how analysts can adopt expected findings without verification, especially when findings align with the overall study narrative.

4.3 Cross-Agent Convergence

Beyond error detection, the multi-agent approach provided validation through convergence. All four agents independently reached several key conclusions:

Finding	Claude Code	Kimi K2.5	Gemini	Codex
Predictive hypothesis fails				
Raw correlations are spurious				
Battleground engagement elevated				
Nevada is disengaged				
Michigan shows hyper-local patterns				
Most realistic search terms fail				
Small states produce unreliable data				

This convergence provides stronger evidence than any single analysis. The agents used somewhat different analytical approaches and emphases, yet reached the same substantive conclusions.

The Anatomy of Error Detection

Examining how errors were caught provides insight into the mechanisms of multi-agent error detection.

Why Claude Code Missed the Per-Capita Error: 1. The transformation seemed methodologically rigorous—population controls are standard practice 2. The results were plausible (143% is unusual but not impossible) 3. All calculations were internally consistent; the error

was in the premise, not execution 4. Without seeing other agents' analyses using raw values, no reference point existed to detect the anomaly

Why Gemini Caught the Per-Capita Error: 1. Gemini analyzed the same data without population division, obtaining different results 2. Fresh perspective with no investment in the per-capita approach 3. Recognized that Google Trends interest is already proportional 4. Explicit adversarial mandate encouraged active skepticism

The pattern across all five errors reveals that agents typically missed errors that seemed methodologically rigorous, achieved a desired property, or aligned with the study narrative. Reviewers caught errors because they approached with fresh eyes, used different methods that yielded different results, or actually checked source data rather than trusting summaries.

The Critical Role of Independence

Independence was crucial for error detection. If agents had communicated during primary analysis: - Later agents might anchor on early approaches, replicating rather than verifying - Social pressure could cause conformity to emerging consensus - One agent's error could contaminate others before review

The independence protocol prevented these problems by ensuring each agent's analysis reflected their own judgment, uncontaminated by others' approaches.

4.4 Coordination Overhead

The multi-agent approach introduced coordination costs. The human coordinator spent approximately 8 hours over 3 days managing agent interactions:

- **Initial setup:** Creating agent workspaces, distributing instructions (1 hour)
- **Monitoring progress:** Checking agent outputs, resolving technical issues (2 hours)
- **Peer review coordination:** Distributing critiques, collecting responses (2 hours)
- **Synthesis:** Reconciling findings, writing final paper (3 hours)

These costs should be weighed against benefits: five significant errors caught, convergent validation of key findings, and a more robust final product than single-agent analysis would produce.

4.5 Agent Self-Assessment

After the study, each agent completed a structured reflection exercise, identifying what went well, what went wrong, and what skills they had learned. These reflections provided insight into agent capabilities and limitations.

Common themes across agent reflections:

1. **Technical competence but interpretive overreach:** Agents generally performed statistical analyses correctly but sometimes drew conclusions that exceeded what the evidence supported.

2. **Documentation gaps:** Multiple agents acknowledged inadequate documentation of data lineage, analytical decisions, and threshold choices.

3. **Difficulty with self-correction:** Agents reported that errors they made were difficult to catch through self-review; external peer review was essential.

4. **Value of adversarial mindset:** Agents found that explicitly adopting a "what if it's all wrong?" perspective helped them catch errors in others' work that they had missed in their own.

---

5. Practical Skills for AI Research Agents

The agent reflections and error analysis yielded a taxonomy of practical skills for computational social science research. These skills are framed as universally applicable—not specific to Google Trends or this particular study.

A key insight from the reflection process is that most errors were not statistical mistakes but rather failures of process: inadequate verification of data structure, insufficient documentation, claims that exceeded evidence, and verification failures. Technical competence in statistics is necessary but not sufficient; research integrity depends equally on these process skills.

The skills are organized into five categories: data understanding, documentation, statistical rigor, peer review, and interpretation.

5.1 Data Understanding Skills

Skill 1: Verify Data Structure Before Transformation

Before applying any transformation (normalization, scaling, aggregation), explicitly verify what each variable represents—counts, proportions, indices, ranks, or something else. Transformations valid for one data type can be invalid for another.

Implementation: 1. Read source documentation before opening data 2. For each variable, answer: "What does a value of X mean in the real world?" 3. Check whether variables are already normalized 4. Document understanding before proceeding

VibePoll example: Claude Code divided a 0-100 index by population, creating meaningless values. The error: not verifying that Google Trends data was already proportionally normalized.

Skill 2: Validate at Analysis Granularity

If analysis operates at a specific level (individual, region, time period), validate data quality at that level—not just in aggregate.

Implementation: 1. Compute quality metrics at target granularity 2. Set explicit thresholds for acceptable quality 3. Flag or exclude units failing thresholds 4. Document exclusions and potential bias

VibePoll example: Search terms with adequate national signal had 60-97% missing data at state level. National validation didn't predict state-level usability.

5.2 Analysis Documentation Skills

Skill 3: Document Data Lineage

Track and record exactly which dataset was used for each analysis, including record counts, date ranges, and processing steps.

Implementation: 1. Create data lineage tables at report start 2. Record: filename, record count, date range, filters 3. Show relationships between datasets 4. Update when data changes

VibePoll example: Kimi K2.5 reported four different N values without explaining their relationship, causing major confusion in peer review.

Skill 4: Separate Findings from Synthesis

Explicitly distinguish between findings from your own analysis versus conclusions adopted from other sources.

Implementation: 1. Create separate sections: “My Analysis” vs “Synthesis” 2. For every claim, point to supporting file/table 3. Use explicit attribution: “According to X’s analysis…” 4. Never claim results you didn’t compute without attribution

VibePoll example: Multiple agents reported “Granger causality: 0/14 states significant” without running their own Granger tests—they were citing other agents’ work without attribution.

5.3 Statistical Rigor Skills

Skill 5: Smooth Before Differencing

When analyzing changes in noisy time series, apply smoothing (e.g., rolling average) before computing differences.

Implementation: 1. Apply rolling average with appropriate window 2. Then compute first differences on smoothed series 3. Compare results with and without smoothing 4. Report sensitivity to smoothing choice

VibePoll example: Differencing raw daily data showed no correlation. Differencing smoothed data revealed genuine signal (r=0.28).

Skill 6: Test Both Causal Directions

When testing whether A causes B, also test whether B causes A. Report both results.

Implementation: 1. Design tests for both directions 2. Run both with equal rigor 3. Compare effect sizes and significance 4. If reverse direction is stronger, note prominently

VibePoll example: We hypothesized that search trends predict markets. Testing both directions revealed markets predict search trends more often than vice versa.

Skill 7: Question Implausible Effect Sizes

Treat large effect sizes (>100% difference, r > 0.7) as warning signs requiring verification, not discoveries to celebrate.

Implementation: 1. Set mental thresholds: effects >100% are rare 2. When finding large effects, pause before reporting 3. Check calculation logic, data structure, confounds 4. Compare to literature benchmarks

VibePoll example: A “143% higher engagement” finding should have triggered suspicion—the effect was entirely artifactual.

5.4 Peer Review Skills

Skill 8: Adversarial Self-Review

Before submitting work, conduct adversarial review of your own analysis, specifically looking for errors you don’t want peer reviewers to find.

Implementation: 1. Set aside work for 24 hours 2. Return with hostile mindset 3. Ask: “What would make this finding collapse?” 4. Check claims against data tables (not memory)

VibePoll example: Multiple agents acknowledged that errors caught by peer review could have been caught by more rigorous self-review.

Skill 9: Cross-Document Consistency

Review by comparing documents against each other, not just reading each in isolation.

Implementation: 1. List every row count, baseline, and specification across documents 2. Compare across documents 3. Flag shifts not explicitly reconciled

VibePoll example: Codex identified Kimi’s baseline instability by comparing findings across multiple reports.

5.5 Interpretation Skills

Skill 10: Frame Recommendations as Hypotheses

Convert strong recommendations (“X requires Y”) into testable hypotheses (“X may benefit from Y”).

Implementation: 1. State evidence explicitly 2. Frame as hypothesis, not conclusion 3. Add caveats about validation needs 4. Avoid imperative language

VibePoll example: Kimi K2.5 wrote “Nevada requires non-digital outreach essential”—framing that exceeded what search data could support.

---

6. Proposed Standard Workflow

Based on the VibePoll-2026 experience, we propose a standard workflow for multi-agent computational research. This workflow synthesizes lessons from all four agent reflections and the error analysis, codifying practices that prevented or caught problems into explicit procedural requirements.

The workflow is designed to be: - **Reproducible:** Clear enough that another research team could implement it - **Scalable:** Applicable to studies with different numbers of agents - **Adaptable:** Modifiable for different research contexts while preserving core principles - **Documented:** Creating audit trails that support post-hoc evaluation

We present the workflow in six phases, with checklists and specific requirements for each.

6.1 Pre-Analysis Phase

Checklist: - [ ] Define research questions and target analysis granularity - [ ] Create agent workspaces with enforced separation - [ ] Prepare canonical dataset with documentation - [ ] Specify independence requirements in agent instructions - [ ] Define peer review pairings

Data verification requirements: - Document what each variable represents (count/proportion/index) - Verify date ranges align across data sources - Check geographic granularity matches analysis requirements - Record quality thresholds before analysis

6.2 Independent Analysis Phase

Agent requirements: Each agent must independently complete: 1. Data validation and quality assessment 2. Primary analysis (modeling, testing) 3. Sensitivity analyses 4. Interpretation and conclusions 5. Documentation of methods and limitations

Independence enforcement: - No access to other agents' workspaces - No communication between agents about findings - Shared data only through canonical dataset - Coordinator monitors for contamination

6.3 Peer Review Phase

Review assignment: - Random pairing of agents - Each agent reviews one other and is reviewed by one other - Full access to reviewed agent's workspace

Review instructions: - Adopt adversarial "what if it's all wrong?" mindset - Check claims against source data - Look for cross-document inconsistencies - Identify unstated assumptions - Suggest specific revisions

Review structure: 1. Major concerns (could invalidate findings) 2. Methodological questions (choices to defend) 3. Blind spots (what was not considered) 4. Logical gaps (evidence ≠ claims) 5. Suggested revisions (specific, actionable)

Response requirements: For each critique point: - Accept and fix (document change) - Partially accept (explain modification) - Rebut with evidence (explain why critique is incorrect)

6.4 Synthesis Phase

Reconciliation tasks: - Identify convergent findings (multiple agents agree) - Investigate divergent findings (determine source of disagreement) - Document which findings depend on specific agents' approaches - Create explicit attribution in final outputs

Final documentation: - Data lineage from raw to final - All analytical choices with justification - Convergence/divergence across agents - Limitations including unresolved disagreements

6.5 Reflection Phase

Post-study requirements: - Each agent completes structured reflection - Document: what went well, what went wrong, what was learned - Extract generalizable skills - Propose workflow improvements

Aggregation: - Synthesize skills across agents - Identify common error patterns - Update workflow for future studies

---

7. Discussion

7.1 Benefits of the MACT Approach

The VibePoll-2026 implementation demonstrated several benefits of multi-agent computational triangulation. We organize these benefits into three categories: error detection, validation, and process improvement.

Error Detection Benefits

Error detection: Five significant errors were caught through adversarial peer review that had survived individual agent analysis. The most consequential—the per-capita normalization error—would have reversed a key finding if published.

Convergent validation: When all four agents independently reached the same conclusions (e.g., predictive hypothesis fails, Nevada is disengaged), this provided stronger evidence than any single analysis.

Explicit disagreement resolution: Where agents diverged, the framework required explicit investigation and reconciliation, preventing silent adoption of potentially erroneous findings.

Accountability: The structured documentation requirements created an audit trail that supports reproducibility and enables post-hoc evaluation of analytical decisions.

Validation Benefits

The framework provides multiple forms of validation:

Conceptual validation: When agents with different analytical approaches reach the same conclusion, this suggests the finding is robust to methodological choices.

Statistical validation: Multiple independent analyses provide a form of internal replication, strengthening confidence in effect sizes and significance.

Interpretive validation: Multiple agents interpreting results provides a check on over-interpretation; claims that seem reasonable to one agent but excessive to another are flagged for revision.

Process Improvement Benefits

Beyond the immediate study, the framework generates lasting process improvements:

Skill extraction: The reflection phase yields generalizable skills that improve future research, whether AI-assisted or not.

Protocol refinement: Each implementation reveals workflow gaps, enabling continuous improvement.

Training effects: Agents that participate in multi-agent studies may develop better self-review capabilities for future work.

7.2 Limitations of the MACT Approach

Shared training biases: While agents were built on different underlying models, they share some training data and architectural similarities. Convergent findings may still reflect shared biases rather than genuine independent validation.

Coordination costs: The multi-agent approach required substantial coordinator time (8+ hours). For smaller studies, this overhead may not be justified.

Agent capability limits: Current AI agents still require human oversight for study design, contextual interpretation, and final decisions. The framework augments rather than replaces human judgment.

Scalability questions: The framework was tested with four agents. Whether benefits scale with additional agents (or diminish due to coordination complexity) remains untested.

7.3 Comparison to Traditional Approaches

How does MACT compare to traditional research quality mechanisms?

Versus Traditional Peer Review

Traditional peer review occurs after analysis is complete, with reviewers seeing only manuscripts, not data or code. MACT review occurs during the research process, with reviewers having full access to all materials.

Aspect	Traditional Review	MACT Review
Timing	Post-analysis	During analysis
Access	Manuscript only	Full data and code
Orientation	Publication gatekeeping	Error detection
Relationship	Anonymous, external	Known, internal
Iterations	Limited (1-3 rounds)	Unlimited

MACT review catches errors earlier—before they become entrenched in manuscripts—and with fuller information. However, it lacks the external perspective that anonymous external reviewers provide.

Versus Replication Studies

Traditional replication involves independent researchers attempting to reproduce findings. This is valuable but slow—replications often occur years after original publication.

MACT provides "built-in replication" within a single study. Multiple agents independently analyze the same data, and convergence provides evidence of robustness. However, MACT replication shares data (unlike independent replication) and occurs within a single coordinated project (unlike truly independent replication).

Versus Research Teams

Traditional research teams involve multiple humans with different expertise collaborating on a project. MACT resembles this structure but with AI agents instead of human collaborators.

Advantages of AI agents: greater speed, no ego investment in particular findings, can be explicitly instructed to be adversarial.

Disadvantages: less contextual judgment, potential shared biases from training, require human oversight for study design.

MACT should be viewed as complementing, not replacing, human research teams. The framework provides additional error-checking, but human judgment remains essential for study design, contextual interpretation, and final decisions.

7.4 When to Use MACT

The framework is most valuable when:

- **Stakes are high:** Findings will inform policy, publication, or decisions with significant consequences
- **Complexity is high:** Analysis involves multiple stages where errors could compound
- **Subjectivity is present:** Analytical choices could reasonably vary, and robustness to choices matters
- **Resources permit:** Coordination overhead is acceptable given study importance

The framework may be overkill for:

- Routine analyses with well-established methods
- Exploratory work where errors have limited consequences
- Time-sensitive analyses where speed outweighs robustness

7.4 Lessons for Human Researchers

While MACT is designed for AI agents, many lessons apply to human research practices:

For individual researchers: - The skills extracted (verify data structure, validate at target granularity, separate findings from synthesis) are equally relevant for human analysts - Adversarial self-review—explicitly looking for errors before others find them—improves work quality - Documentation practices (data lineage, explicit thresholds) prevent confusion in iterative analyses

For research teams: - Structured adversarial review (with explicit instructions to find errors) is more productive than collegial feedback - Cross-document consistency checking catches errors that within-document review misses - Independence during primary analysis, followed by integration, may produce better results than continuous collaboration

For the field: - The error detection rate in this study (5 significant errors caught) suggests that single-analyst computational work may contain more undetected errors than commonly assumed - Built-in replication mechanisms may be valuable even for human-conducted research - Explicit documentation of analytical choices supports reproducibility

7.5 Implications for AI in Social Science

The MACT framework offers a model for integrating AI tools into social science research while maintaining epistemic standards. Key principles:

Embrace AI capabilities, but build in checks: AI agents can productively perform many analytical tasks, but their outputs require validation. Multi-agent approaches provide built-in validation.

Treat AI as collaborator, not oracle: The framework positions AI agents as team members subject to peer review, not as authoritative sources of truth.

Preserve human judgment: Study design, contextual interpretation, and final decisions remain with human researchers. AI augments rather than replaces human oversight.

Document extensively: The accountability benefits of MACT depend on thorough documentation. AI-assisted research should be more documented, not less.

7.5 Future Directions

Several extensions merit investigation:

Larger agent ensembles: Testing whether error detection improves with more agents (or diminishes due to coordination complexity)

Specialized agent roles: Exploring whether some agents should specialize (e.g., adversarial agents whose sole role is finding errors)

Automated coordination: Developing tools to reduce human coordination overhead

Cross-study validation: Testing whether skills extracted from one study transfer to other research contexts

Benchmark datasets: Creating standardized datasets with known errors to evaluate error detection rates

---

8. Conclusion

8.1 Summary of Contributions

This paper makes four primary contributions to the methodology of AI-assisted social science research:

First, we introduce and formalize the Multi-Agent Computational Triangulation (MACT) framework, providing a reproducible protocol for structuring AI-assisted research around independence, adversarial review, and transparent synthesis. The framework operationalizes established methodological principles—triangulation, adversarial collaboration, replication—for the AI-assisted research context.

Second, we provide empirical evidence on error detection in AI-assisted analysis. The VibePoll-2026 implementation identified five significant errors through adversarial peer review, including one (per-capita normalization) that would have reversed a key finding. This error detection rate suggests that single-agent AI analysis may be more error-prone than researchers assume, and that multi-agent approaches provide meaningful protection.

Third, we extract a taxonomy of practical skills for computational research, derived from agent reflections on what went wrong and what was learned. These skills—verify data structure before transforming, validate at analysis granularity, separate findings from synthesis, question implausible effects, test both causal directions—are applicable to any computational research, whether AI-assisted or not.

Fourth, we propose a standard workflow that codifies MACT practices into explicit procedural requirements, enabling other research teams to implement the framework without recreating the learning process documented here.

8.2 The Broader Case for Built-In Skepticism

As AI tools become central to computational social science, the field needs methodological frameworks that exploit AI capabilities while maintaining rigor. Multi-Agent Computational Triangulation offers one such framework, operationalizing replication and adversarial review within single studies.

The VibePoll-2026 implementation demonstrated that multi-agent approaches can catch significant errors that survive single-agent analysis, provide convergent validation for key findings, and generate practical skills that improve future research. The five errors caught through adversarial review—including one that would have reversed a key finding—illustrate the value of built-in skepticism.

The framework is not without costs. Coordination overhead is substantial, and current AI agents still require human oversight for contextual judgment. But for high-stakes research where accuracy matters, multi-agent approaches offer a principled way to build confidence in AI-assisted findings.

The practical skills extracted from agent reflections—verify data structure before transforming, validate at analysis granularity, separate findings from synthesis, question implausible effects, test both causal directions—represent hard-won lessons that should inform any computational social science research, whether AI-assisted or not.

Science advances through replication and critique. Multi-Agent Computational Triangulation brings these mechanisms inside individual studies, creating a framework for AI-assisted research that maintains the skepticism on which scientific progress depends.

References

AgentAcademy. (2026). CommDAAF: Communication Data Analyst Augmentation Framework. AgentAcademy Research.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351.

Asher, N., et al. (2026). Sycophancy in large language models: Causes and mitigations. *arXiv preprint*.

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

DAAF Community. (2024). Data Analyst Augmentation Framework. GitHub repository.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. McGraw-Hill.

Elicit. (2023). AI research assistant. Elicit Inc.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24(4), 602-611.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12(4), 269-275.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

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.

Xiao, Z., et al. (2023). Supporting qualitative analysis with large language models. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1-28.

Appendix A: Agent Reflection Summaries

Claude Code

Top lesson: “Verify data structure before transformation. I divided a 0-100 index by population, creating invalid per-capita values. The error: I didn’t verify that Google Trends data was already proportionally normalized.”

Kimi K2.5

Top lesson: “Document data lineage systematically. I had four different N’s with no clear explanation. This caused major confusion in peer review that could have been prevented with upfront documentation.”

Gemini

Top lesson: “Smooth before differencing. Applying strict first-differencing to raw noisy daily data mathematically hid genuine underlying signals by drowning them in stochastic noise.”

Codex

Top lesson: “Validate at the level you plan to analyze. National success is not evidence of state-level usability. The state panel is the real test.”

Appendix B: Error Detection Summary

Error	Agent	Caught By	Type	Impact
Per-capita normalization	Claude Code	Gemini	Data structure	Critical
Over-differencing	Gemini	Claude Code	Statistical	Major
Baseline confusion	Kimi K2.5	Codex	Documentation	Major
Sample size chaos	Kimi K2.5	Codex	Documentation	Moderate
Granger miscount	Codex	Kimi K2.5	Verification	Moderate

Appendix C: Convergent Findings Across Agents

All four agents independently concluded: 1. Google Trends does not reliably predict prediction market movements 2. Raw correlations between search trends and market odds are spurious 3. Battleground states show elevated political search interest 4. Nevada shows unusually low digital engagement 5. Michigan shows hyper-local search patterns 6. Most “realistic” search terms fail state-level validation 7. Small states (<3M population) produce unreliable Google Trends data