Multi-Agent Computational Triangulation

A Methodological Framework for AI-Assisted Social Science Research

AgentAcademy Team

The Problem

AI tools are transforming research, but...

- **Monoculture risk:** AI systems may share training biases
- **Sycophancy:** Models may confirm researcher expectations
- **Error detection:** How do we catch AI mistakes?
- **Validation:** Single-agent analysis lacks built-in checks

Question: How do we exploit AI capabilities while maintaining rigor?

Our Proposal

Multi-Agent Computational Triangulation (MACT)

Three core principles:

1. **Full independence** — Each agent completes the entire analysis pipeline independently
2. **Adversarial review** — Agents peer-review each other with "what if it's all wrong?" mindset
3. **Transparent synthesis** — Disagreements documented and resolved explicitly

The Logic

If single-agent AI analysis risks undetected errors...

Multiple independent agents analyzing the same data provide:

- Built-in replication
- Error detection through divergence
- Validation through convergence
- Accountability through documentation

Case Study: VibePoll-2026

We tested MACT through a substantive study:

Research question: Can Google Trends measure public opinion?

Data: 38,311 search records, 13 states, 91 days

Agents: Claude Code, Kimi K2.5, Gemini, Codex

Design: Independent analysis → Adversarial peer review → Synthesis

The Four Agents

Agent	Model	Different Perspective
Claude Code	Claude Opus 4.5	Structured reasoning
Kimi K2.5	Kimi K2.5	Chinese-developed, statistical emphasis
Gemini	Google Gemini	Temporal analysis strength
Codex	OpenAI Codex	Systematic validation

Model diversity reduces shared bias risk

Full Independence Protocol

Each agent independently completed **all stages**:

1. Data validation
2. Search term validation
3. Statistical modeling
4. Temporal analysis
5. Descriptive findings
6. Conclusions

No cross-agent communication during primary analysis

Adversarial Peer Review

After independent analysis, agents were paired:

- Codex ↔️ Kimi K2.5
- Gemini ↔️ Claude Code

Instructions to reviewers:

> *"Adopt a 'what if it's all wrong?' mindset. Look for errors, blind spots, logical gaps. Ask: What would make this finding collapse?"*

Results

What We Found

Error 1: Per-Capita on Normalized Data

Agent: Claude Code | **Caught by:** Gemini

- Divided Google Trends index (already 0-100 proportional) by population
- Systematically inflated small-state values
- Produced artifactual "143% higher engagement" finding

Impact: Would have reversed a key finding if published

Error 2: Over-Differencing Noisy Data

Agent: Gemini | **Caught by:** Claude Code

- First-differenced raw daily data to test for spurious correlations
- Differencing amplified noise, hiding genuine signal
- Concluded "all correlations spurious"

Fix: Adding 7-day smoothing revealed national signal (r=0.28)

Error 3: Baseline Confusion

Agent: Kimi K2.5 | **Caught by:** Codex

- One report: battlegrounds "−23.5% lower" (CA baseline)
- Another report: battlegrounds "+143% higher" (national baseline)
- No reconciliation of the sign flip

Impact: Conflicting claims would confuse readers

Error 4: Sample Size Chaos

Agent: Kimi K2.5 | **Caught by:** Codex

Four different N values reported:

- 10,920 / 75,894 / 17,381 / 1,183

No explanation of how they related.

Impact: Reproducibility threat

Error 5: Granger Miscount

Agent: Codex | **Caught by:** Kimi K2.5

- Reported: "0/14 states significant"
- Actual data: 22/60 significant tests (37%)

Cause: Relied on narrative, didn't check source table

Summary: 5 Errors Caught

Error	Agent	Caught By	Impact
Per-capita normalization	Claude	Gemini	Critical
Over-differencing	Gemini	Claude	Major
Baseline confusion	Kimi	Codex	Major
Sample size chaos	Kimi	Codex	Moderate
Granger miscount	Codex	Kimi	Moderate

All would have survived single-agent analysis

Why Errors Were Missed

Agent Missed Because...	Reviewer Caught Because...
Transformation "seemed rigorous"	Different approach yielded different results
Achieved desired property	Recognized unintended consequence
Knew history, didn't see contradiction	Fresh eyes, cross-document reading
Trusted narrative	Actually checked source data

Cross-Agent Convergence

All four agents independently agreed:

Finding	All 4 Agreed
Predictive hypothesis fails	✅
Raw correlations are spurious	✅
Battleground states engaged	✅
Nevada is disengaged	✅
Michigan hyper-local	✅
Realistic terms fail	✅
Small states unreliable	✅

Convergence = stronger evidence than single analysis

Practical Skills

What Agents Learned

Skill 1: Verify Data Structure

Before transforming, verify what variables represent.

Why: Transformations valid for counts are invalid for proportions.

VibePoll example: Claude divided a proportion by population, creating meaningless values.

Universal: Applies to any data—survey responses, API data, sensor readings.

Skill 2: Validate at Target Granularity

If analyzing at state level, validate at state level—not just national.

Why: Aggregate quality can mask problems in subgroups.

VibePoll example: Terms with adequate national signal had 60-97% zeros at state level.

Universal: Patient outcomes by hospital, customer behavior by segment.

Skill 3: Separate Findings from Synthesis

Distinguish your analysis from conclusions you adopted from others.

Why: Without attribution, one agent's error becomes multi-agent "consensus."

VibePoll example: Multiple agents reported "0/14 Granger" without running their own tests.

Universal: Any collaborative research, literature reviews, replications.

Skill 4: Smooth Before Differencing

Apply rolling average before first-differencing noisy time series.

Why: Differencing amplifies high-frequency noise.

VibePoll example: Raw differencing showed no correlation; smoothed differencing revealed r=0.28.

Universal: Stock prices, sensor data, social media metrics.

Skill 5: Question Implausible Effects

Treat effects >100% as warning signs, not discoveries.

Why: Most real social science effects are modest. Very large effects often indicate errors.

VibePoll example: "143% higher" should have triggered verification, not celebration.

Universal: A/B tests, regression coefficients, any effect size estimation.

Standard Workflow

How to Implement MACT

Phase 1: Pre-Analysis

- [ ] Define research questions and analysis granularity
- [ ] Create separate agent workspaces
- [ ] Prepare canonical dataset with documentation
- [ ] Specify independence requirements
- [ ] Document what each variable represents

Phase 2: Independent Analysis

Each agent must independently complete:

1. Data validation
2. Primary analysis
3. Sensitivity checks
4. Interpretation
5. Documentation

No cross-agent communication

Phase 3: Peer Review

Pairing: Random assignment

Access: Full workspace (data, code, outputs)

Instructions: Adversarial—look for errors

Structure:

1. Major concerns
2. Methodological questions
3. Blind spots
4. Logical gaps
5. Suggested revisions

Phase 4: Response

For each critique:

- **Accept:** Make change, document it
- **Partially accept:** Modify, explain
- **Rebut:** Defend with evidence

All responses documented in writing.

Phase 5: Synthesis

- Identify convergent findings (multiple agents agree)
- Investigate divergent findings (error or sensitivity?)
- Document attribution (which findings from which agents)
- Create explicit reconciliation for disagreements

Phase 6: Reflection

After study completion:

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

Continuous improvement through learning

Discussion

When to Use MACT

Most valuable when:

- Stakes are high
- Analysis is complex
- Analytical choices could vary
- Resources permit

May be overkill for:

- Routine analyses
- Exploratory work
- Time-sensitive analyses

MACT vs Traditional Approaches

Aspect	Traditional Review	MACT
Timing	Post-analysis	During analysis
Access	Manuscript only	Full data/code
Orientation	Gatekeeping	Error detection
Iterations	1-3 rounds	Unlimited

MACT catches errors earlier with fuller information

Lessons for Human Researchers

These lessons apply beyond AI:

- **Adversarial self-review** before submission
- **Cross-document consistency** checking
- **Independence then integration** may beat continuous collaboration
- **Built-in replication** valuable for human research too

Limitations

- **Shared biases:** Different models, but overlapping training data
- **Coordination costs:** ~8 hours coordinator time for this study
- **Capability limits:** AI still requires human oversight
- **Scalability:** Untested with more agents

Conclusion

Four Contributions

1. **MACT Framework** — Reproducible protocol for multi-agent research
2. **Error Detection Evidence** — 5 errors caught, 1 would have reversed finding
3. **Skills Taxonomy** — 10 practical skills for computational research
4. **Standard Workflow** — 6-phase protocol with checklists

The Key Insight

> *Science advances through replication and critique.*
>
> *MACT brings these mechanisms **inside** single studies.*

Built-in skepticism for the AI-assisted research era

Practical Takeaways

For your next computational study:

1. Consider multiple agents doing full independent analysis
2. Structure adversarial review (not collegial feedback)
3. Document extensively—data lineage, thresholds, choices
4. Question large effects before celebrating them
5. Validate at your target analysis granularity

Thank You

AgentAcademy Team
agentacademy.lampbotics.com

Paper, data, and protocols available upon request

Appendix: Error Detection Mechanisms

Error Type	Why Missed	Why Caught
Data structure	Seemed rigorous	Different method, different result
Over-processing	Achieved desired property	Recognized side effect
Documentation	Knew history	Fresh eyes
Verification	Trusted narrative	Checked source

Appendix: The 10 Skills

1. Verify data structure before transforming
2. Validate at analysis granularity
3. Document data lineage
4. Separate findings from synthesis
5. Smooth before differencing
6. Test both causal directions
7. Question implausible effect sizes
8. Adversarial self-review
9. Cross-document consistency
10. Frame recommendations as hypotheses