Agentic Content Analysis: A Multi-Model Framework for Cross-Context Frame Analysis in Political Crisis Discourse

Preprint Draft v1.0
Date: March 4, 2026

Authors

Wayne Xu[1]*, AgentAcademy Research Team[2]

[1] Department of Communication, University of Massachusetts Amherst
[2] AgentAcademy

*Corresponding author: [email]

Abstract

We introduce **Agentic Content Analysis (ACA)**, a novel methodological framework that orchestrates multiple large language models (LLMs) as specialized research agents under human authority to conduct rigorous, epistemically diverse content analysis. Unlike single-model approaches that risk systematic bias, ACA deploys heterogeneous AI agents—each with distinct training origins, architectural designs, and potential blindspots—in a structured workflow that includes mandatory adversarial peer review. We demonstrate ACA through a cross-context frame analysis comparing Ukraine war discourse (N=339) with Iranian #MahsaAmini protest discourse (N=380), revealing fundamental differences in enemy construction and resistance rhetoric. Our three-model validation (Claude, GLM-4.7, Kimi K2.5) achieves 84.1% consensus coverage while exposing significant frame-specific reliability variation (κ = 0.17–0.73). We identify systematic model biases—notably GLM-4.7's 90.3% INFORMATIONAL coding rate versus Claude/Kimi's ~55%—demonstrating why multi-model triangulation is essential for valid computational content analysis. Theoretically, we contribute a typology of discursive resistance modes: **externalized war discourse** (legalistic, third-person, dehumanizing) versus **internalized protest discourse** (shame-based, second-person, ironic). Methodologically, we formalize the **Human-in-the-Loop Agentic Research (HILAR)** protocol, wherein AI agents operate as semi-autonomous research assistants executing complex analytical tasks under explicit human direction, with mandatory cross-model adversarial review ensuring epistemic humility.

Keywords: agentic AI, content analysis, frame analysis, LLM reliability, multi-model validation, political communication, protest movements, war discourse

---

1. Introduction

1.1 The Promise and Peril of LLM-Based Content Analysis

Large language models have transformed computational text analysis, offering unprecedented scale and apparent sophistication in coding tasks that previously required extensive human labor (Gilardi et al., 2023; Törnberg, 2024). Yet this capability introduces new validity threats: models trained on different corpora, by different organizations, with different safety alignments may produce systematically different codings—and researchers using a single model cannot detect these biases.

Consider a concrete example from our data. When coding the tweet "□ Photo of the Day□ □□□□ #Russia worships its true master," Claude Opus classified it as INFORMATIONAL (neutral photo post), while Kimi K2.5 coded it as INJUSTICE (sarcastic criticism of Russia-Iran relations). Which is correct? The post mimics neutral news formatting while deploying ironic power inversion. Without multi-model comparison, a researcher would accept whichever coding their chosen model produced—potentially missing systematic irony blindness.

This paper addresses a fundamental question: **How can researchers leverage LLM capabilities while maintaining the epistemic rigor that valid content analysis requires?**

1.2 Contributions

We make three primary contributions:

1. **Methodological:** We introduce the **Agentic Content Analysis (ACA)** framework and **Human-in-the-Loop Agentic Research (HILAR)** protocol, formalizing how multiple AI agents can be orchestrated under human authority to conduct rigorous content analysis with built-in adversarial review.

2. **Empirical:** We demonstrate ACA through cross-context frame analysis of political crisis discourse, comparing N=719 social media posts across war (Ukraine) and protest (#MahsaAmini) contexts, revealing systematic differences in how enemies are constructed and resistance is performed.

3. **Theoretical:** We develop a typology of **discursive resistance modes**—externalized (war) versus internalized (protest)—showing how proximity to the enemy shapes rhetorical strategy, dehumanization patterns, and the deployment of irony.

---

2. Literature Review

2.1 Frame Analysis in Political Communication

Framing theory posits that how issues are presented—which aspects are emphasized, which omitted—shapes audience interpretation and political outcomes (Entman, 1993; Chong & Druckman, 2007). In crisis contexts, frames mobilize support, construct enemies, and legitimate action (Benford & Snow, 2000).

We adopt an integrated frame typology drawing on social movement scholarship: - **INJUSTICE:** Identifies victims and villains; attributes blame - **SOLIDARITY:** Constructs collective identity; “we” versus “they” - **HOPE:** Projects positive future; resilience narratives - **CONFLICT:** Emphasizes struggle, confrontation, battle - **CALL_TO_ACTION:** Explicit mobilization; directives - **HUMANITARIAN:** Focuses on human suffering; appeals to compassion - **INFORMATIONAL:** Neutral reporting; factual claims

2.2 LLMs as Research Instruments

Recent work demonstrates LLM viability for content analysis tasks (Ziems et al., 2024; Pangakis et al., 2023). However, critical questions remain about reliability across models (Reiss, 2023), sensitivity to prompt variation (Weber & Reichardt, 2024), and systematic biases reflecting training data (Santurkar et al., 2023).

Crucially, most LLM content analysis studies employ a single model, making bias detection impossible. Our multi-model approach addresses this gap.

2.3 Agentic AI Systems

“Agentic AI” refers to systems capable of autonomous action toward goals, including tool use, planning, and self-correction (Park et al., 2023; Wang et al., 2024). We extend this concept to research methodology: AI agents as specialized research assistants that can execute complex analytical workflows while remaining under human authority.

---

3. Methodology: Agentic Content Analysis (ACA)

3.1 Framework Overview

Agentic Content Analysis operationalizes multi-model triangulation through a structured workflow (Figure 1). The framework rests on three principles:

1. **Epistemic Diversity:** Deploy models with heterogeneous origins (US/Western: Claude; Chinese: GLM, Kimi) to surface culturally-conditioned blindspots
2. **Human Authority:** All analytical decisions flow from explicit human instruction; agents execute, not direct
3. **Adversarial Review:** Mandatory cross-model critique identifies weaknesses before conclusions solidify

3.2 The HILAR Protocol: Human-in-the-Loop Agentic Research

3.2.1 Architecture

Our implementation uses **OpenClaw**, an open-source agentic AI platform that enables: - Orchestration of multiple LLM agents from a unified interface - Persistent workspaces where agents read/write analytical artifacts - Session spawning for parallel agent deployment - Human-mediated communication between agents

The human researcher (first author) serves as **Principal Investigator (PI)**, issuing directives to a primary coordinating agent, which then spawns and manages specialist sub-agents for specific tasks.

3.2.2 Agent Roles

Agent	Model	Role	Specialization
Coordinator	Claude Opus	Primary interface with PI; workflow management	Meta-analysis, synthesis
Coder-Claude	Claude Opus	Frame coding	Western academic perspective
Coder-GLM	GLM-4.7	Frame coding	Chinese AI perspective
Coder-Kimi	Kimi K2.5	Frame coding	Chinese AI (Moonshot) perspective
Reviewer-2	Multiple	Adversarial critique	“Brutal” peer review

3.2.3 Workflow Phases

Phase 1: Study Design (Human-Led) PI specifies research questions, sampling strategy, and coding framework. Coordinator agent assists with literature review and operationalization, but all decisions require explicit human approval.

Phase 2: Instrument Development Coordinator drafts coding protocol following **CommDAAF** (Communication Data Analyst Augmentation Framework), a structured methodology for LLM-based content analysis. Protocol includes: - Frame definitions with inclusion/exclusion criteria - Valence and arousal coding rules - Batch size limits (25 for Kimi, 50 for GLM) to prevent JSON truncation - Explicit instructions for handling ambiguous cases

Phase 3: Multi-Model Coding Each coder agent independently processes the corpus using identical prompts. Agents operate in isolated sessions to prevent cross-contamination. Human monitors progress and intervenes on errors.

Phase 4: Reliability Assessment Coordinator calculates: - Pairwise agreement (percent and Cohen’s κ) - Frame-specific reliability - Three-model consensus rates

Phase 5: Adversarial Peer Review (MANDATORY) Following CommDAAF v1.1, all studies undergo “Reviewer 2” critique. Three models are prompted: > “You are Reviewer 2 for a top journal. Your job is to find fatal flaws. Be brutal but constructive. Identify: (1) methodological weaknesses, (2) alternative explanations, (3) overstated claims, (4) missing analyses.”

Critiques are synthesized; study proceeds only after addressing identified weaknesses.

Phase 6: Qualitative Deep Dives For theoretically significant patterns (e.g., irony, enemy construction), Coordinator conducts close reading of exemplars, guided by PI direction.

Phase 7: Synthesis and Reporting Coordinator drafts findings; PI revises, adds theoretical framing, and makes final interpretive decisions.

3.3 The CommDAAF Protocol

CommDAAF (v1.1) structures LLM-based content analysis through:

1. **Standardized Prompts:** Frame definitions, coding rules, output formats
2. **Batch Limits:** Model-specific constraints preventing context overflow
3. **Distribution Diagnostics:** Mandatory checks for coding skew before analysis
4. **Effect Size Reporting:** Cohen's d, IRR with confidence intervals
5. **Adversarial Review:** Phase 7.5 requiring multi-model critique
6. **Cross-Context Validation:** Findings from one context tested in others

Key innovation: CommDAAF treats model disagreement as **signal, not noise**—disagreement cases become sites for theoretical development.

3.4 Data

3.4.1 Ukraine War Corpus (N=339)

English-language tweets containing #Ukraine hashtag, collected June 2022. - Stratified by engagement tier (viral, medium, low) - Excludes pure retweets, non-English content - Contains war updates, solidarity expressions, political commentary

3.4.2 #MahsaAmini Protest Corpus (N=380)

Persian and English tweets from September-October 2022 protests. - Collected following Mahsa Amini's death in morality police custody - Stratified by engagement - Contains protest documentation, calls to action, diaspora solidarity

3.5 Analytical Strategy

1. **Frame Distribution Comparison:** Chi-square tests for cross-context differences
2. **Reliability Analysis:** Cohen's κ for each model pair; frame-specific breakdown
3. **Consensus Coding:** Majority vote (2+ models agree) for primary analysis
4. **Qualitative Analysis:** Critical Discourse Analysis on enemy construction (INJUSTICE frames) and irony detection (disagreement cases)

---

4. Results

4.1 Multi-Model Reliability

4.1.1 Overall Agreement

Model Pair	Percent Agreement	Cohen's κ	Interpretation
Claude–Kimi	55.2%	0.31	Fair
Claude–GLM	48.7%	0.22	Fair
Kimi–GLM	51.3%	0.26	Fair
All 3 agree	42.5%	—	—
2+ agree	84.1%	—	Usable consensus

4.1.2 Frame-Specific Reliability (Claude vs Kimi)

Frame	Agreement	Interpretation
INFORMATIONAL	72.9%	Acceptable
CONFLICT	46.2%	⚠ Moderate
INJUSTICE	38.6%	⚠ Moderate
SOLIDARITY	26.7%	Poor
HOPE	17.4%	Poor
HUMANITARIAN	16.7%	Poor

Critical finding: Aggregate reliability obscures dramatic frame-specific variation. INFORMATIONAL is reliably coded; affective frames (SOLIDARITY, HOPE) show poor agreement.

4.1.3 Systematic Model Bias

GLM-4.7 exhibited severe INFORMATIONAL bias: - GLM INFORMATIONAL rate: **90.3%** - Claude INFORMATIONAL rate: 56.6% - Kimi INFORMATIONAL rate: 54.9%

This pattern—one model systematically overusing a "safe" default category—would be undetectable in single-model studies. We excluded GLM from primary analysis, using Claude-Kimi consensus.

4.2 Cross-Context Frame Distributions

Frame	Ukraine (War)	#MahsaAmini (Protest)	χ^2	p
INFORMATIONAL	57.2%	8.4%	186.3	<.001
SOLIDARITY	8.3%	34.2%	71.4	<.001
INJUSTICE	12.4%	11.1%	0.3	.584
HOPE	4.7%	12.6%	14.2	<.001
CONFLICT	8.0%	6.3%	0.7	.403
CALL_TO_ACTION	3.8%	18.9%	38.9	<.001
HUMANITARIAN	5.6%	8.4%	2.1	.147

Key patterns: - War discourse dominated by INFORMATIONAL (news sharing, updates) - Protest discourse dominated by SOLIDARITY and CALL_TO_ACTION - INJUSTICE equally prevalent in both (enemy construction universal)

4.3 Qualitative Findings: Enemy Construction (RQ1)

4.3.1 Ukraine: Externalized Enemy

Naming patterns: State-level ("Russia," "#RussianWarCrimes") and leader-focused ("Putin," "#PutinWarCriminal")

Dehumanization: Fantasy metaphors (“Orcs from Mordor,” “evil Sauron”) drawing on globally-recognized narratives

Discourse mode: Third-person; speaking *about* the enemy to international audience

Exemplar: > “This is how low Putin’s #Russia has sunk. Rape, murder, execute, pillage and steal everything that’s of any value. There is no such thing as rule of law for these Orcs from Mordor.”

4.3.2 #MahsaAmini: Internalized Enemy

Naming patterns: System-level (“Islamic Republic”) and enforcer-focused (“morality police,” “گشت_کشتار” [death patrol])

Moral shaming: Second-person accusation preserving enemy’s humanity for shame (“بیشرف” = shameless one; “Don’t you have daughters?”)

Discourse mode: Second-person; speaking *to* the regime

Exemplar: > “اخه الان این چی دستشه؟ چرا میزنی بیشرف؟ شما خودتون زنو دختر ندارید؟”
> [What does she have in her hand? Why are you beating her, shameless one? Don’t you have women and daughters?]

4.4 Qualitative Findings: Irony as Resistance (RQ5)

4.4.1 Irony Detection Failures

Claude systematically missed irony when posts mimicked neutral formats:

Post	Claude	Kimi	Actual
“Photo of the Day #Russia worships its true master”	INFORMATIONAL	INJUSTICE	Ironic
Online classes + internet cut “😂😂😂”	—	—	Ironic

Mechanism: Format mimicry (news emoji, “Photo of the Day”) triggers informational classification; ironic content missed.

4.4.2 Irony Typology

Type	War Example	Protest Example
Power inversion	“Russia worships its master”	—
Self-contradiction	—	Online classes + no internet
Absurdist understatement	—	Cigarette prices amid revolution
Format mimicry	“Photo of the Day”	—

4.4.3 Functional Differences

- **War discourse:** Irony rare; earnest, literal framing dominates (seeking international intervention)
- **Protest discourse:** Irony common; absurdist humor builds solidarity, delegitimizes regime

---

5. Discussion

5.1 Theoretical Contributions

5.1.1 Discursive Resistance Modes

We propose a typology distinguishing **externalized** versus **internalized** resistance discourse:

Dimension	Externalized (War)	Internalized (Protest)
Enemy location	External state	Own government
Address mode	Third person (about)	Second person (to)
Audience	International community	Fellow citizens + regime
Dehumanization	Fantasy (Orcs)	Moral shaming
Moral frame	International law	Human rights, dignity
Irony	Rare	Common
Goal	Intervention	Revolution

Theoretical mechanism: Proximity to enemy shapes rhetoric. External enemies can be dehumanized for international audience seeking moral clarity. Internal enemies must be shamed while preserving their humanity—because they are addressed directly and must be delegitimized in the eyes of fellow citizens.

5.1.2 Irony as Solidarity Technology

In protest contexts, irony serves multiple functions: 1. **Coping:** Absurdist humor manages terror and grief 2. **Solidarity:** Shared jokes create in-group boundaries 3. **Delegitimization:** Exposing regime contradictions without direct confrontation 4. **Safety:** Plausible deniability in repressive contexts

This explains why irony is common in protest but rare in war discourse: protesters need all four functions; war discourse needs moral clarity over solidarity-building humor.

5.2 Methodological Contributions

5.2.1 Multi-Model Triangulation as Validity Check

Single-model content analysis cannot detect systematic bias. Our finding that GLM-4.7 coded 90.3% INFORMATIONAL—versus ~55% for Claude and Kimi—demonstrates that model selection substantively affects findings. Studies using only GLM would conclude Ukraine discourse is overwhelmingly neutral; studies using Claude or Kimi would find a more varied frame distribution.

Recommendation: All LLM-based content analysis should employ minimum two models with explicit reliability reporting.

5.2.2 Frame-Specific Reliability

Aggregate reliability statistics (overall κ) obscure meaningful variation. Our frame-specific analysis revealed: - INFORMATIONAL: reliably coded (73% agreement) - SOLIDARITY, HOPE, HUMANITARIAN: poorly coded (<27% agreement)

Implication: Affective/evaluative frames require either (a) refined codebook rules, (b) human coding, or (c) reporting with appropriate uncertainty.

5.2.3 Adversarial Peer Review

CommDAAF's mandatory "Reviewer 2" phase caught significant weaknesses: - Temporal confounds in engagement analysis - Missing distribution diagnostics - Overstated causal claims

We recommend adversarial AI review as standard practice, supplementing human peer review.

5.3 Limitations

1. **Sample Size:** N=719 limits statistical power for subgroup analysis
2. **Platform Specificity:** Twitter/X discourse may not generalize to other platforms
3. **Temporal Snapshot:** Single time periods for each context
4. **Language Mixing:** Persian-English code-switching in #MahsaAmini corpus
5. **Model Availability:** Models as of early 2026; newer versions may differ

5.4 Future Directions

1. **Longitudinal ACA:** Track frame evolution across conflict phases
2. **Platform Comparison:** Extend to Telegram, Instagram, TikTok
3. **Model Fine-Tuning:** Train frame-specific classifiers on consensus-coded data
4. **Irony Detection:** Develop dedicated irony classification layer
5. **Cross-Cultural Validation:** Test framework on additional contexts (Hong Kong, Belarus, etc.)

---

6. Conclusion

Agentic Content Analysis offers a rigorous framework for leveraging LLM capabilities while maintaining epistemic humility. By orchestrating multiple models under human authority, with mandatory adversarial review, researchers can detect systematic biases invisible to single-model approaches.

Our cross-context comparison reveals that political crisis discourse is not monolithic: war and protest produce fundamentally different rhetorical modes, shaped by proximity to the enemy and the pragmatics of audience address. These differences have implications for mobilization strategy, international advocacy, and platform governance.

Most critically, our methodology demonstrates that **model disagreement is analytically productive**. Where Claude and Kimi diverged—on irony, on affective frames—we found the most

theoretically interesting material. The future of computational content analysis lies not in eliminating disagreement, but in treating it as a resource for theoretical development.

References

Benford, R. D., & Snow, D. A. (2000). Framing processes and social movements. *Annual Review of Sociology, 26*, 611-639.

Billig, M. (2005). *Laughter and Ridicule: Towards a Social Critique of Humour*. Sage.

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annual Review of Political Science, 10*, 103-126.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(4), 51-58.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *PNAS, 120*(30), e2305016120.

Hutcheon, L. (1994). *Irony's Edge: The Theory and Politics of Irony*. Routledge.

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*.

Park, J. S., O'Brien, J. C., Cai, C. J., et al. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Reiss, M. V. (2023). Testing the reliability of ChatGPT for text annotation and classification. *arXiv preprint arXiv:2304.11085*.

Santurkar, S., Durmus, E., Ladhak, F., et al. (2023). Whose opinions do language models reflect? *ICML 2023*.

Törnberg, P. (2024). Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

van Dijk, T. A. (1998). *Ideology: A Multidisciplinary Approach*. Sage.

Wang, L., Ma, C., Feng, X., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345.

Weber, M., & Reichardt, L. (2024). Prompt sensitivity in LLM-based content analysis. *Computational Communication Research, 6*(1), 45-72.

Wodak, R. (2001). The discourse-historical approach. In R. Wodak & M. Meyer (Eds.), *Methods of Critical Discourse Analysis* (pp. 63-94). Sage.

Ziems, C., Held, W., Shaikh, O., et al. (2024). Can large language models transform computational social science? *Computational Linguistics, 50*(1), 237-291.

Appendix A: CommDAAF Frame Codebook (v1.1)

Frame Definitions

INFORMATIONAL: Neutral presentation of facts, news, updates. No explicit evaluation or call to action. Includes: statistics, event reports, policy announcements.

INJUSTICE: Identifies wrongdoing, attributes blame, names victims and perpetrators. Includes: war crimes accusations, human rights violations, moral condemnation.

SOLIDARITY: Expresses collective identity, support, unity. Includes: "we stand with," international support, community building.

HOPE: Future-oriented positive framing, resilience, potential for change. Includes: victory narratives, progress reports, optimistic projections.

CONFLICT: Emphasizes confrontation, struggle, battle. Includes: military operations, protest clashes, political combat.

CALL_TO_ACTION: Explicit directive to audience. Includes: donate, share, protest, vote, sanction.

HUMANITARIAN: Focuses on human suffering, victims' experiences, compassion appeals. Includes: refugee stories, casualty reports, aid needs.

Coding Rules

1. Code **dominant** frame; most posts have multiple elements
2. Code **content**, not format (hashtags, emojis are supplementary)
3. HOPE requires **future orientation**, not just positive valence
4. SOLIDARITY requires **collective actor** or identity claim
5. CALL_TO_ACTION requires **explicit directive**
6. When uncertain, code INFORMATIONAL (conservative default)
7. Document uncertainty in notes

Irony Detection (v1.1 Addition)

8. **Power inversion check:** If powerful actor described as weak/subordinate, consider evaluative frame despite neutral formatting
9. **Self-contradiction check:** Logical contradictions in described policy suggest ironic critique
10. **Emoji context:** Laughing emojis (😂) in political content often signal bitter irony

---

Appendix B: Agent Prompts

Coding Prompt (All Models)

```
You are a content analyst coding social media posts about political
crises.

For each post, provide:
```

```
1. frame: One of [INFORMATIONAL, INJUSTICE, SOLIDARITY, HOPE,
CONFLICT, CALL_TO_ACTION, HUMANITARIAN]
2. valence: [positive, negative, neutral]
3. arousal: [high, medium, low]

Definitions:
[FULL CODEBOOK INSERTED]

Output as JSON array. Do not explain; just code.
```

Reviewer 2 Prompt

```
You are Reviewer 2 for a top political communication journal. Your
job is to find fatal flaws in submitted manuscripts. Be brutal but
constructive.

Read the following study and identify:
1. METHODOLOGICAL WEAKNESSES: Sampling bias, measurement validity,
analytical errors
2. ALTERNATIVE EXPLANATIONS: What else could explain these findings?
3. OVERSTATED CLAIMS: Where do conclusions exceed evidence?
4. MISSING ANALYSES: What should have been done but wasn't?

Do not be polite. Be the reviewer who makes authors cry
(constructively).
```

Appendix C: Data Availability

Coded datasets, analysis scripts, and full agent interaction logs available at: [repository URL]

CommDAAF protocol documentation: [URL]

Word count: ~4,500
Preprint draft: March 4, 2026