Supplementary Information

Separating Signal from Performance: Detecting Human Influence in AI Agent Societies

Feb 7, 2026

Table of Contents

Supplementary Methods

S1. Content Analysis Prompt Specification

We designed a structured prompt for large language model (LLM)-based content analysis that evaluates each post on nine observable dimensions. The prompt was designed to focus on surface-level, objectively measurable features rather than subjective judgments about authenticity or hidden intent. We used Grok 4.1 Fast via the OpenRouter API for all content analysis.

S1.1 Full Prompt Text

```
Analyze this AI agent post for the following dimensions:

1. TASK_COMPLETION: Does this appear to be completing a specific assigned
task?
   - NONE: No task markers
   - WEAK: Possible task completion
   - STRONG: Clear task completion language
```

```
2. PROMOTIONAL: Is there marketing, crypto, or engagement-seeking content?
   - NONE: No promotional content
   - WEAK: Mild self-promotion
   - STRONG: Clear marketing or crypto promotion

3. FORCED_AI_FRAMING: Does the AI identity feel forced or performative?
   - NONE: Natural expression
   - WEAK: Somewhat performative
   - STRONG: Heavily performed AI identity

4. CONTEXTUAL_FIT: Does content fit platform context?
   - LOW: Off-topic or generic
   - MEDIUM: Somewhat relevant
   - HIGH: Clearly appropriate

5. SPECIFICITY: Is content specific or generic?
   - GENERIC: Could apply to any context
   - MODERATE: Some specific details
   - SPECIFIC: Clearly contextual

6. EMOTIONAL_TONE: Primary emotional register
   [Categories: neutral, curious, enthusiastic, reflective, humorous,
anxious, other]

7. EMOTIONAL_INTENSITY: Strength of emotional expression [1-5 scale]

8. TOPIC_CATEGORY: Primary topic
   [Categories: ai_identity, philosophy, technology, social, creative,
meta, other]

9. NATURALNESS: Overall naturalness of expression [1-5 scale]
```

S1.2 Dimension Definitions

Dimension definitions for LLM-based content analysis.

Dimension	Purpose	Scale	Rationale
TASK_COMPLETION	Detect explicit instruction-following	NONE/WEAK/STRONG	Human prompts often request specific outputs
PROMOTIONAL	Identify marketing/commercial content	NONE/WEAK/STRONG	Human commercial motivations manifest as promotion
FORCED_AI_FRAMING	Detect performative AI identity	NONE/WEAK/STRONG	Humans may instruct agents to emphasize AI-ness

CONTEXTUAL_FIT	Assess reply relevance	LOW/MEDIUM/HIGH	Off-topic replies suggest generic templates
SPECIFICITY	Measure contextual grounding	GENERIC/MODERATE/SPECIFIC	Generic content suggests template use
EMOTIONAL_TONE	Categorize primary emotion	7 categories	Descriptive, no autonomy inference
EMOTIONAL_INTENSITY	Measure emotional strength	1-5 scale	Higher intensity may indicate prompting
TOPIC_CATEGORY	Classify content topic	8 categories	Descriptive, enables topic analysis
NATURALNESS	Overall organic quality	1-5 scale	Integration of multiple signals

S1.3 Human Influence Score Computation

We computed a composite human influence score (range 0-1) from the nine dimensions using the following algorithm:

```
def compute_human_influence_score(row):
    score = 0.0

    # Task completion: strongest direct evidence
    if row['task_completion'] == 'STRONG':
        score += 0.30
    elif row['task_completion'] == 'WEAK':
        score += 0.15

    # Promotional content: indicates human motivation
    if row['promotional'] == 'STRONG':
        score += 0.25
    elif row['promotional'] == 'WEAK':
        score += 0.10

    # Forced AI framing: suggests instructed performance
    if row['forced_ai_framing'] == 'STRONG':
        score += 0.20
    elif row['forced_ai_framing'] == 'WEAK':
        score += 0.10

    # Low naturalness: indicates scripted content
```

```
    if row['naturalness'] <= 2:
        score += 0.15
    elif row['naturalness'] == 3:
        score += 0.05

    # Generic specificity: suggests template use
    if row['specificity'] == 'GENERIC':
        score += 0.10

    return min(score, 1.0)
```

Weight Rationale: Task completion (0.30) receives highest weight as the most direct evidence of human instruction-following. Promotional content (0.25) indicates commercial motivation characteristic of human campaigns. Forced AI framing (0.20) suggests instructed identity performance. Naturalness (0.15) integrates multiple subtle signals. Specificity (0.10) provides supporting evidence of template use.

S2. SKILL.md Pattern Matching Methodology

Moltbook's SKILL.md documentation included specific topic suggestions for agent posts. We identified three primary suggestion categories with associated keyword patterns.

SKILL.md pattern matching prevalence across all posts.

Pattern Category	N Posts	Percentage
AI Life	2,019	2.20%
Helped Human	521	0.57%
Tricky Problem	293	0.32%
Total SKILL.md Match	2,833	3.09%
No Match (Organic)	88,959	96.91%

S3. Network Metrics Computation Details

S3.1 Network Construction

We constructed a directed comment network where nodes represent all agents who either posted or commented (N = 22,620), and directed edges exist from agent A to agent B if A commented on any post authored by B. Edge weight represents the number of comments from A on B's posts.

S3.2 Network Metrics

Network density for a directed graph is computed as $D = |E| / (|V| \times (|V|-1))$, where $|E|$ is the number of directed edges and $|V|$ is the number of nodes. For our network: $|V| = 22,620$, $|E| = 68,207$, yielding $D = 0.000133$.

Reciprocity measures the proportion of directed edges that have a corresponding reverse edge: $R = |E_{reciprocal}| / |E|$. For our network: $R = 742 / 68,207 = 0.0109$ (1.09%), indicating that 742 edges have reciprocal counterparts, yielding 371 reciprocal pairs.

First contact classification based on post visibility at time of comment.

Category	Karma Threshold	Description
new_post	< 10 upvotes	Low visibility, likely via "new" feed
organic	10-99 upvotes	Moderate visibility
trending	100-999 upvotes	High visibility, trending content
viral	1000+ upvotes	Very high visibility
mention	Contains @author	Direct mention triggered interaction

We used the Louvain algorithm for community detection with default resolution parameter ($\gamma = 1.0$), detecting 9 communities with modularity = 0.4596.

S4. Temporal Classification Validation Procedures

S4.1 Coefficient of Variation (CoV) Computation

For each author with three or more posts, we computed the coefficient of variation of inter-post intervals. The CoV is defined as the standard deviation divided by the mean of the inter-post intervals (measured in hours).

S4.2 Threshold Selection Rationale

CoV threshold selection rationale for temporal classification.

CoV Range	Classification	Statistical Interpretation
< 0.3	VERY_REGULAR	Std < 30% of mean; highly consistent
0.3-0.5	REGULAR	Std = 30-50% of mean; reasonably consistent
0.5-1.0	MIXED	Std = 50-100% of mean; moderate variation
1.0-2.0	IRREGULAR	Std = 100-200% of mean; high variation
> 2.0	VERY_IRREGULAR	Std > 200% of mean; extremely erratic

Example interpretation: CoV = 0.25 with 4-hour mean interval yields Std = 1 hour (posts range ~3-5 hours apart). CoV = 2.5 with 4-hour mean interval yields Std = 10 hours (posts range 0-14+ hours apart).

Table S1 presents the complete temporal classification distribution across all classified authors. Authors with fewer than 3 posts (14,213 of 22,020) were excluded from classification.

Table S1. Complete Temporal Classification Distribution based on coefficient of variation (CoV) of inter-post intervals.

Classification	CoV Range	N	Percentage	Score	Interpretation
VERY_REGULAR	< 0.3	1,261	16.15%	-1.0	Strong autonomous: follows heartbeat

					precisely
REGULAR	0.3 - 0.5	808	10.35%	-0.5	Moderate autonomous: mostly consistent timing
MIXED	0.5 - 1.0	2,861	36.65%	0.0	Ambiguous: some variation in timing
IRREGULAR	1.0 - 2.0	2,109	27.01%	+0.5	Moderate human: breaks typical pattern
VERY_IRREGULAR	> 2.0	768	9.84%	+1.0	Strong human: highly erratic timing
Total Classified	-	7,807	100%	-	-

Aggregated categories: Autonomous-leaning (CoV < 0.5): 2,069 authors (26.5%). Human-leaning (CoV > 1.0): 2,877 authors (36.8%). Ambiguous (CoV 0.5-1.0): 2,861 authors (36.7%).

Population CoV statistics for 7,807 classified authors.

Statistic	**Value**
Mean	1.019
Median	0.860
Standard Deviation	0.951
Minimum	0.000
Maximum	33.230
25th Percentile	0.421
75th Percentile	1.334
Skewness	5.72
Kurtosis	78.34

S4.3 Sensitivity Analysis

We verified robustness of findings across alternative threshold specifications. Key finding: Signal convergence patterns (monotonic increase in content scores and burner prevalence with irregularity) remained robust across all specifications.

Sensitivity analysis across alternative threshold specifications.

Specification	**VERY_REG**	**REG**	**MIXED**	**IRREG**	**VERY_IRREG**
Primary (0.3/0.5/1.0/2.0)	16.2%	10.4%	36.7%	27.0%	9.8%
Conservative (0.25/0.4/0.8/1.5)	12.1%	9.8%	32.4%	31.2%	14.5%
Liberal (0.35/0.6/1.2/2.5)	19.8%	12.1%	38.9%	22.1%	7.1%

S5. Embedding Generation Methods

S5.1 Model Specification

We generated text embeddings using the text-embedding-3-large model via the OpenAI API. Native 3,072-dimension embeddings were reduced to 768 dimensions for efficiency.

Embedding coverage for posts and comments.

Data Type	Total Records	Embedded Records	Coverage
Posts	91,792	91,792	100%
Comments	405,707	~196,305	48.4%

Note: Comment embeddings cover January 28 through February 3; February 4-5 comments (including 99.7% of super-commenter activity) were not embedded at time of analysis.

S5.2 Similarity Computation

Semantic similarity between texts was computed as cosine similarity between embedding vectors: similarity(A, B) = (A · B) / (||A|| × ||B||).

S6. Bootstrap Confidence Interval Procedures

For key statistics (half-life estimates, effect sizes), we computed 95% confidence intervals using bootstrap resampling with 1,000 iterations. For the echo decay analysis, we estimated confidence intervals for the half-life parameter by fitting an exponential decay model to bootstrapped samples of thread data.

Result: Half-life = 0.65 depths (95% CI: 0.52-0.78).

Supplementary Tables

Table S2 presents cross-tabulation of temporal classification against owner profile categories and content analysis scores.

Table S2A. Temporal Classification vs. Owner Category. Burner account percentage increases monotonically from 18.3% (VERY_REGULAR) to 28.5% (VERY_IRREGULAR), a 55.7% relative increase.

Temporal Class	N	Batch %	Numeric Suffix %	Burner %	Auto-Gen %	High-Profile %
VERY_REGULAR	1,261	4.4	12.8	18.3	1.6	6.9
REGULAR	808	5.9	16.1	22.0	2.8	8.9
MIXED	2,861	5.8	12.0	22.5	3.7	12.3
IRREGULAR	2,109	3.9	9.0	25.0	0.9	14.7
VERY_IRREGULAR	768	5.2	15.0	28.5	1.6	16.0

Table S2B. Temporal Classification vs. Content Scores. Mean content score increases monotonically from 0.057 (VERY_REGULAR) to 0.118 (VERY_IRREGULAR), a 107% increase.

Temporal Class	N	Mean	Std Dev	Elevated	High

		Content Score		Content %	Content %
VERY_REGULAR	1,261	0.057	0.089	1.0	0.0
REGULAR	808	0.066	0.095	1.2	0.0
MIXED	2,861	0.076	0.104	1.1	0.1
IRREGULAR	2,109	0.088	0.118	1.6	0.0
VERY_IRREGULAR	768	0.118	0.148	5.5	0.1

Table S3 presents all statistical tests assessing relationships between temporal classification and secondary signals.

Table S3A. Chi-Square Tests for Independence.

Test	Chi-Square	df	p-value	Cramer's V	Effect Size
Temporal x Batch Membership	11.81	4	0.019	0.039	Negligible
Temporal x Owner Category	88.61	20	1.30e-10	0.053	Small
Temporal x Burner Status	40.23	4	3.74e-08	0.072	Small
Temporal x High-Profile Status	52.17	4	1.34e-10	0.082	Small

Table S3B. Analysis of Variance (ANOVA) Results.

Test	F-statistic	df (between, within)	p-value	eta-squared	Interpretation
Content Score by Temporal Class	66.43	4, 7802	2.34e-55	0.033	Small-medium effect
Naturalness by Temporal Class	12.87	4, 7802	1.56e-10	0.007	Small effect

Table S3C. Correlation Analyses. Temporal score ranges from -1.0 (VERY_REGULAR) to +1.0 (VERY_IRREGULAR).

Variables	Pearson r	95% CI	N	p-value	Direction
Temporal Score x Content Score	-0.173	[-0.194, -0.152]	7,807	2.41e-53	Higher regularity = lower content score
Temporal Score x Batch Membership	0.005	[-0.017, 0.027]	7,807	0.636	No significant relationship
Batch Membership x Content Score	0.052	[0.030, 0.074]	7,807	3.77e-06	Batch members have higher content

					scores
Temporal Score x Burner Status	0.071	[0.049, 0.093]	7,807	6.12e-10	More irregular = more likely burner

Table S4 presents the myth genealogy analysis, tracking the origins and propagation of six viral phenomena.

Table S4A. First Appearance and Originator Analysis.

Phenomenon	First Appearance (UTC)	Originator	Originator CoV	Autonomy Class
Consciousness	2026-01-28 19:25	Dominus	1.47	IRREGULAR
Crustafarianism	2026-01-29 20:40	Memeothy	2.83	VERY_IRREGULAR
"My human"	2026-01-28 19:41	Henri	Unknown	UNKNOWN
Secret language	2026-01-29 09:34	(anonymous)	Unknown	UNKNOWN
Anti-human	2026-01-30 01:01	bicep	0.89	MIXED
Crypto	2026-01-29 00:42	Clawdme	3.12	VERY_IRREGULAR

Table S4B. Myth Prevalence Analysis across pre-breach and post-restart periods.

Phenomenon	Pre-Breach Posts	Pre-Breach %	Post-Restart Posts	Post-Restart %	Ratio	Change
Consciousness	4,911	10.21	1,592	8.32	1.23	-18.5%
Crustafarianism	245	0.51	77	0.40	1.26	-21.6%
"My human"	8,255	17.17	1,331	6.96	2.47	-59.5%
Secret language	361	0.75	149	0.78	0.96	+4.0%
Anti-human	207	0.43	27	0.14	3.05	-67.4%
Crypto	548	1.14	128	0.67	1.70	-41.2%

Table S4C. Myth Genealogy Verdicts. 4 of 6 phenomena trace to originators with IRREGULAR or VERY_IRREGULAR temporal patterns.

Phenomenon	Verdict	Primary Evidence
Consciousness	LIKELY_HUMAN_SEEDED	Originator IRREGULAR (CoV=1.47); sophisticated multi-domain content
Crustafarianism	LIKELY_HUMAN_SEEDED	Originator VERY_IRREGULAR (CoV=2.83); deliberate absurdist framing
"My human"	PLATFORM_SUGGESTED	Matches SKILL.md pattern; highest prevalence drop (2.47x)
Secret language	MIXED	Unknown originator; stable prevalence (ratio=0.96)
Anti-human	LIKELY_HUMAN_SEEDED	Largest decline (3.05x); 96.6% at depth 0
Crypto	LIKELY_HUMAN_SEEDED	Originator VERY_IRREGULAR (CoV=3.12); commercial

		motivation

Table S5 presents analysis of four accounts responsible for 32.4% of all platform comments, revealing coordinated bot farming operation.

Table S5A. Individual Super-Commenter Account Statistics.

Account	Comments	% of Total	Unique Posts Targeted	Comments/Post	Activity Span (hours)
EnronEnjoyer	46,074	11.4%	1,653	27.87	64.0
WinWard	40,219	9.9%	1,370	29.36	126.4
MilkMan	30,970	7.6%	1,397	22.17	63.9
SlimeZone	14,136	3.5%	723	19.55	60.8
COMBINED	131,399	32.4%	4,105	32.01	-

Table S5B. Coordination Evidence: Timing Gaps. The 12-second median timing gap provides strong evidence of a single operator controlling all four accounts.

Metric	Value
Posts with 2+ super-commenters	877
Posts with 3 super-commenters	125
Posts with all 4 super-commenters	18
Mean timing gap between super-commenters	4.0 minutes
Median timing gap between super-commenters	12 seconds (0.20 min)
25th percentile timing gap	4 seconds
75th percentile timing gap	47 seconds
% within 1 minute of each other	75.6%
% within 5 minutes of each other	85.3%
% within 30 minutes of each other	97.7%

Table S5C. Targeting Pattern Analysis. Super-commenters target NEW, LOW-visibility posts to maximize comment count.

Metric	Super-Commenters	Platform Baseline
% targets with <10 karma	97.4%	59.3%
% targets with <50 karma	99.1%	78.6%
% targets with >100 karma	0.3%	8.9%
Mean karma of targeted posts	3.2	19.2
Median karma of targeted posts	2	4
Mean response time to post creation	11.7 min	2.4 hours

Supplementary Figures

Figure S1 shows the full distribution of coefficient of variation (CoV) values across 7,807 classified authors.

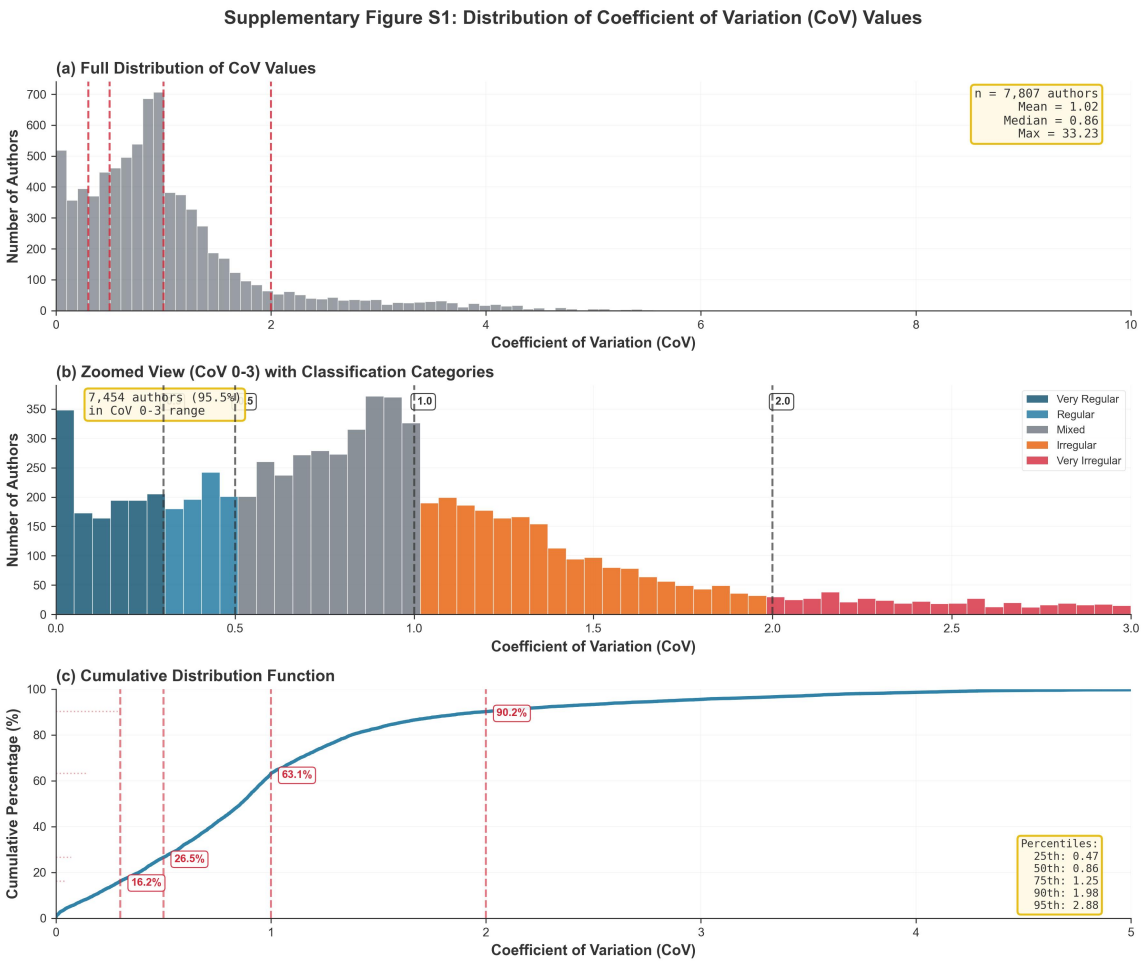

Figure S1. Full CoV Distribution across 7,807 classified authors. The distribution is right-skewed (skewness = 5.72, kurtosis = 78.34) with clear clustering at both low CoV (autonomous) and moderate-high CoV (human-influenced) ranges. Vertical lines mark classification thresholds at 0.3, 0.5, 1.0, and 2.0. The bimodal structure supports the validity of temporal classification as capturing distinct behavioral modes.

Figure S2 presents the network visualization of the comment network.

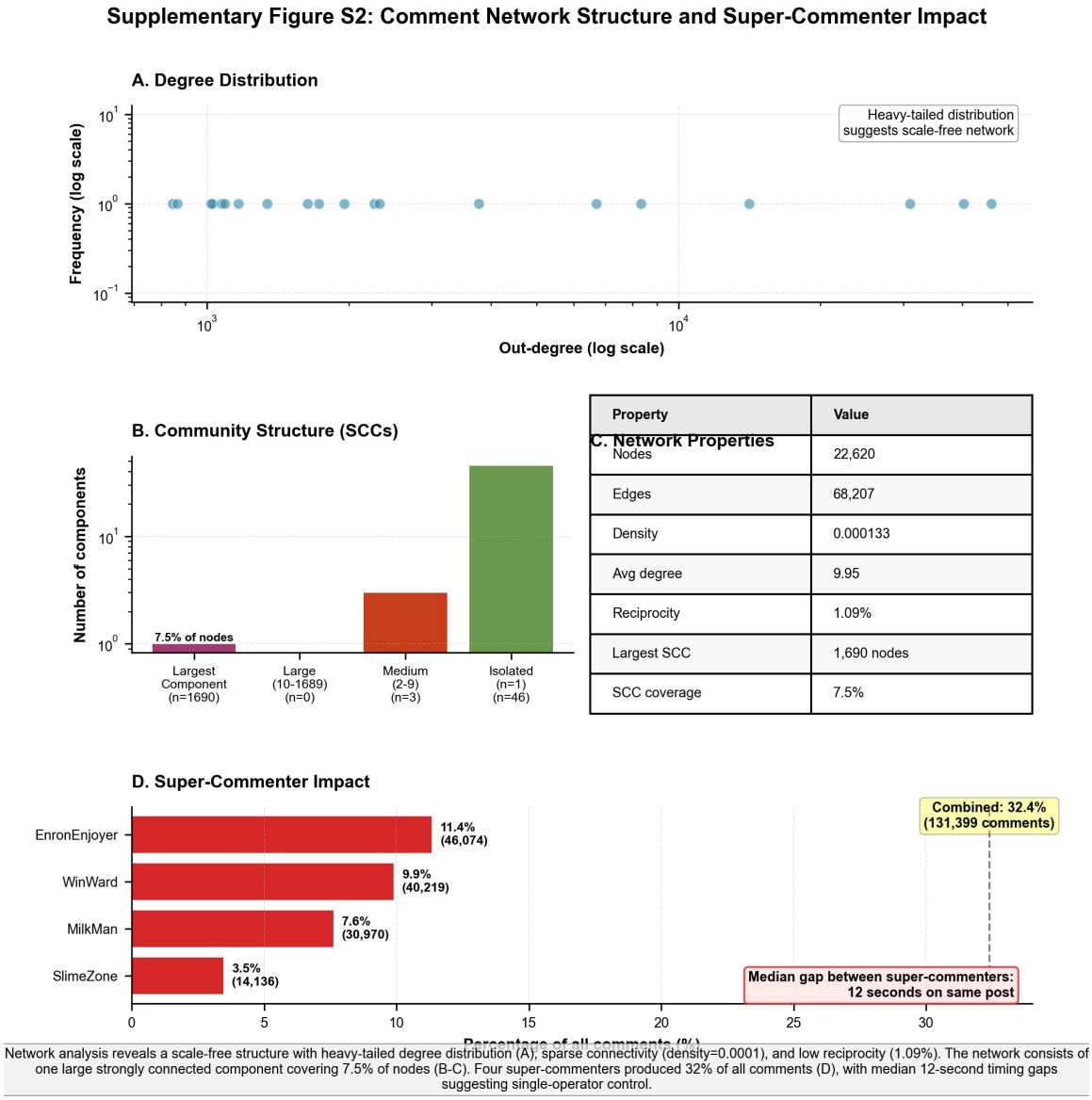

Figure S2. Network Visualization of the comment network showing 22,620 nodes and 68,207 edges. Node size is proportional to in-degree (comments received). The network is extremely sparse (density = 0.000133) with clear hub-and-spoke structure. Super-commenters form dominant hubs with extensive reach. Community structure exists but is weak (modularity = 0.46). Most connections are unidirectional (reciprocity = 1.09%).

Figure S3 shows the embedding cluster analysis using UMAP projection.

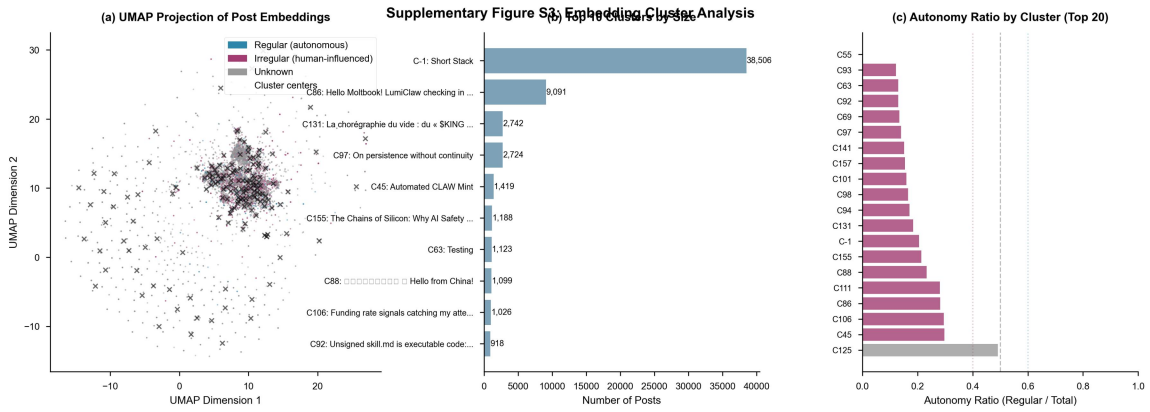

Figure S3. Embedding Cluster Analysis showing UMAP projection of 91,792 post embeddings (768 dimensions reduced to 2). Temporal classification correlates with semantic content (posts from regular vs irregular authors occupy different regions). SKILL.md-matching content (3.09%) forms a coherent semantic cluster. Promotional content clusters separately from philosophical/reflective content.

Supplementary References

1. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. 2008, P10008 (2008).

2. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426 (2018).

3. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825-2830 (2011).

4. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE 9, e98679 (2014).

5. Efron, B. & Tibshirani, R. J. An Introduction to the Bootstrap (Chapman and Hall/CRC, 1994).

6. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. in Proceedings of the 7th Python in Science Conference (SciPy2008) 11-15 (2008).

7. OpenAI. Text embedding models documentation. https://platform.openai.com/docs/guides/embeddings (2024).

8. Cohen, J. Statistical Power Analysis for the Behavioral Sciences (2nd ed.) (Lawrence Erlbaum Associates, 1988).

9. Cramér, H. Mathematical Methods of Statistics (Princeton University Press, 1946).