Testing the Ensemble Averaging Thesis in the Complex Adaptive Model of Societies (CAMS): Variance Reduction, Epistemic Contestation, and Diagnostic Uncertainty in Multi-Rater LLM Scoring

Kari McKern
Independent Complexity Research
kari.freyr.4@gmail.com

May 2026

Abstract

The claim that ensemble averaging of independent LLM raters reduces uncorrelated noise while preserving signal (Wright et al., 2024) is tested against two longitudinal CAMS datasets: Germany (1880–2026, $N = 1{,}176$ node-years) and the United States (1900–2026, $N = 1{,}016$ node-years). We confirm the theoretical variance-reduction prediction ($\sigma_{\bar{x}} = \sigma_{\text{rater}}/\sqrt{5}$) and signal-to-noise improvement ($\sqrt{5}$), and demonstrate that the ensemble mean is effectively unbiased. However, we reject the auxiliary hypothesis that rater disagreement represents idiosyncratic random error. Cross-node uncertainty correlation ($r > 0.5$), strong temporal autocorrelation (lag-1 $r \approx 0.82$), and state-dependent heteroscedasticity (uncertainty rises with system Stress, $r = 0.78$) indicate that ensemble disagreement measures *epistemic ambiguity* about contested historical periods, not measurement noise. We introduce the **Contestation Index** $\mathcal{C}(t) \in [0, 1]$ as a normalized ensemble uncertainty metric, validate it against known historiographical battlegrounds, and derive uncertainty-adjusted bounds for the CAMS κ criticality index. During crises, κ bounds widen dramatically (e.g., Germany 1933: $\kappa = 17.25\,[13.19, 24.91]$), reducing reliability as an early-warning signal. The ensemble mean remains the optimal estimator, but its uncertainty envelope should be reported as a confidence qualifier on all CAMS diagnostics.

1 Introduction

When multiple independent raters assess the same object using a shared rubric, each introduces interpretive noise: idiosyncratic weightings of evidence, recency bias, or domain-specific blind spots. If these errors are uncorrelated across raters, averaging causes them to cancel while the shared signal reinforces. This is the statistical foundation of ensemble methods in machine learning (Dietterich, 2000) and, more recently, in LLM-based assessment (Wright et al., 2024).

Wright et al. (2024) found that "using the average LLM score across models provided the strongest agreement with self-report" when seven commercial LLMs independently scored Big-Five personality traits from open-ended narrative text, achieving convergence "comparable to or exceeding established benchmarks." The implication is that ensemble averaging of LLM raters functions as a *noise-canceling* device, with the mean representing a more reliable estimate of the underlying construct than any single model.

The Complex Adaptive Model of Societies (CAMS) v3.2-R (McKern, 2026a) operationalises this principle through a five-agent ensemble (CAMS5), where independent LLM raters score eight societal nodes (Helm, Shield, Flow, Hands, Craft, Archive, Lore, Stewards) across five dimensions (Coherence, Capacity, Stress, Abstraction, Node Value) for longitudinal country datasets. The ensemble mean is treated as the canonical estimate, but the across-rater standard deviation (the "envelope") has been under-theorised.

This paper tests the Wright thesis against CAMS empirical data, addressing three questions:

1. Does ensemble variance reduce by the theoretical $1/\sqrt{N}$ factor?

2. Is rater disagreement uncorrelated random noise, or does it encode systematic epistemic ambiguity?
3. What are the operational implications for CAMS crisis detection and the κ criticality index?

2 Theory and Hypotheses

2.1 The Ensemble Averaging Model

Let $x_i^{(r)}$ be the score of node i in year t by rater $r \in \{1, \ldots, N\}$. The true (unobserved) value is μ_i, and the rater error is $\varepsilon_i^{(r)}$:

$$x_i^{(r)} = \mu_i + \varepsilon_i^{(r)}, \qquad \mathbb{E}[\varepsilon_i^{(r)}] = 0, \quad \mathrm{Var}(\varepsilon_i^{(r)}) = \sigma^2. \tag{1}$$

If errors are independent across raters, the ensemble mean $\bar{x}_i = \frac{1}{N}\sum_{r=1}^{N} x_i^{(r)}$ has:

$$\mathrm{Var}(\bar{x}_i) = \frac{\sigma^2}{N}, \qquad \mathrm{SE}(\bar{x}_i) = \frac{\sigma}{\sqrt{N}}. \tag{2}$$

The signal-to-noise ratio improves by $\sqrt{N}$:

$$\mathrm{SNR}_{\text{ensemble}} = \frac{\mu_i}{\sigma/\sqrt{N}} = \sqrt{N} \cdot \mathrm{SNR}_{\text{single}}. \tag{3}$$

Hypothesis H1 (Variance Reduction): The observed standard error of the ensemble mean equals the across-rater standard deviation divided by $\sqrt{5}$.

Hypothesis H2 (Uncorrelated Noise): Across-rater disagreement is uncorrelated across nodes and time, consistent with idiosyncratic rater error.

Hypothesis H3 (Unbiasedness): The ensemble mean is unbiased: $\mathbb{E}[\bar{x}_i - \mu_i] = 0$.

2.2 The Contestation Index

If H2 is rejected, uncertainty is not random noise but *epistemic contestation*. We define the **Contestation Index**:

$$\mathcal{C}(t) = \frac{\bar{\sigma}_{\mathrm{SEM}}(t) - \min_t \bar{\sigma}_{\mathrm{SEM}}}{\max_t \bar{\sigma}_{\mathrm{SEM}} - \min_t \bar{\sigma}_{\mathrm{SEM}}} \in [0, 1], \tag{4}$$

where $\bar{\sigma}_{\mathrm{SEM}}(t)$ is the system-averaged standard error of the mean in year t. High $\mathcal{C}(t)$ indicates that raters genuinely disagree about the historical moment, independent of the signal magnitude.

3 Methods

3.1 Datasets

Two CAMS5 ensemble datasets are analysed:

- **Germany** (1880–2026): 147 years × 8 nodes = 1,176 node-years. Covers the Kaiserreich, WWI, Weimar Republic, Nazi era, occupation, Federal Republic, and reunification.
- **United States** (1900–2026): 127 years × 8 nodes = 1,016 node-years. Covers the Progressive Era, Roaring Twenties, Great Depression, WWII, Cold War, Vietnam/Watergate, post-Cold War, GFC, and COVID-19.

Each dataset contains the ensemble mean and an envelope file with across-rater standard deviations for each dimension ($\sigma_C, \sigma_K, \sigma_S, \sigma_A$) and the min/max Node Value across raters.

3.2 Metrics

1. **Standard Error of the Mean (SEM):** $\text{SEM}_i = \bar{\sigma}_i/\sqrt{5}$, where $\bar{\sigma}_i = (\sigma_C + \sigma_K + \sigma_S + \sigma_A)/4$.
2. **Signal-to-Noise Ratio:** $\text{SNR}_i = V_i/\text{SEM}_i$, where V_i is the Node Value.
3. **Cross-node Uncertainty Correlation:** Pearson correlation of SEM_i across node pairs.
4. **Temporal Autocorrelation:** Lag-k correlation of system-level SEM.
5. **Heteroscedasticity:** Correlation of SEM with Stress, Node Value, and σ_V (node value dispersion).
6. **Bias:** $(V_{\max} - V) - (V - V_{\min})$, testing symmetry of the envelope.
7. **Contestation Index:** As defined in Equation (4).

3.3 Simulation Protocol

To test ensemble advantage under controlled conditions, we simulate 1,000 trials where five raters score each node-year with Gaussian noise $N(0, \bar{\sigma}_i^2)$. We compare single-rater versus ensemble correlation with the "true" (observed ensemble mean) value, and measure crisis detection F1 scores.

4 Results

4.1 Variance Reduction and SNR Improvement (H1)

Table 1 confirms H1. The observed SEM-to-SD ratio is 0.447 for both countries, exactly matching the theoretical $1/\sqrt{5} = 0.447$. SNR improves by the predicted $\sqrt{5} = 2.236$ factor.

Table 1: Variance Reduction and SNR Improvement

Metric	Germany	USA	Theory	Verdict
SEM / Across-rater SD	0.447	0.447	0.447	**Confirmed**
SNR improvement factor	2.236	2.236	2.236	**Confirmed**
Mean SNR (ensemble)	38.4	47.2	–	–
Mean SNR (single rater)	17.2	21.1	–	–

4.2 Unbiasedness (H3)

The ensemble mean is effectively unbiased. Germany shows negligible bias ($+0.014 \pm 0.001$, one-sample t-test $p = 0.119$). The USA shows a small but significant negative bias (-0.062 ± 0.001, $p < 0.001$), likely attributable to the envelope file format (deviations rather than absolute values for the USA dataset, corrected in analysis). H3 is confirmed for Germany and marginally violated for USA.

4.3 Error Correlation Structure (H2)

H2 is **rejected**. Table 2 shows that uncertainty is highly structured.

The cross-node correlation of SEM ($r = 0.56$–0.82) is particularly damaging to H2. If errors were idiosyncratic rater noise, they would be uncorrelated across nodes. Instead, when one node is contested, all nodes tend to be contested. This indicates that disagreement is driven by *shared epistemic conditions* (e.g., contested historiography, ambiguous primary sources) rather than independent rater quirks.

Table 2: Error Correlation Structure

Correlation	Germany	USA	Theory (H2)	Verdict
Cross-node SEM correlation	0.56	0.82	< 0.3	**Rejected**
Lag-1 temporal autocorr	0.82	0.85	≈ 0	**Rejected**
Stress vs SEM	0.78	0.65	0	**Rejected**
$\bar{V}$ vs SEM	-0.75	-0.65	0	**Rejected**
σ_V vs SEM	$+0.34$	$+0.46$	0	**Rejected**

4.4 Heteroscedasticity: Uncertainty as Crisis Signal

Uncertainty is state-dependent. Figure **??** (conceptual) shows that SEM rises with Stress and falls with system Node Value. During Germany's WWI–Weimar–Nazi period (1914–1945), mean SEM is 0.42 versus 0.19 during the Cold War stability (1950–1989). The USA shows SEM = 0.38 during the Great Depression (1929–1933) versus 0.12 during the Eisenhower era (1950–1960).

4.5 The Contestation Index

Figure **??** (conceptual) maps $\mathcal{C}(t)$ over time. Germany peaks at:

- 1921–1923 (Weimar hyperinflation, political violence)
- 1930–1932 (Nazi electoral breakthrough, Reichstag paralysis)
- 1944–1945 (total defeat, occupation)

The USA peaks at:

- 1929–1933 (Great Depression onset)
- 2008–2009 (Global Financial Crisis)
- 2020–2021 (COVID-19 pandemic, contested election)

These match known historiographical battlegrounds, validating $\mathcal{C}(t)$ as an epistemic rather than statistical metric.

4.6 Ensemble Robustness and Optimal Aggregation

Simulated ensemble size analysis shows diminishing returns after $N = 3$ raters. The marginal correlation gain from Rater 4 to 5 is < 0.001. The mean outperforms the median and trimmed means (20%, 40%) by < 0.001 in correlation with truth, confirming it as the optimal aggregator under Gaussian error.

Crisis detection F1 scores improve modestly with ensemble averaging: Germany $+0.003$, USA $+0.007$. The absolute gain is small because single-rater accuracy is already high (> 0.95), but the variance reduction is substantial (79%), meaning ensemble diagnoses are more stable across replications.

4.7 Uncertainty-Adjusted κ Criticality

The CAMS κ criticality index $\kappa = B/\omega$ (bond strength divided by rate dispersion) is a leading indicator of phase transitions. Table 3 shows κ with uncertainty-derived bounds.

During crises, κ bounds widen dramatically. The 1933 width (11.72) is $6\times the2020width(1.98), indicatingthatthecriticality$
$a\kappa$ reading of 17.25 during the Nazi seizure of power carries far less confidence than the same reading during a stable period.

Table 3: Germany κ Criticality with Uncertainty Bounds (Selected Years)

Year	κ	κ_{lower}	κ_{upper}	Width
1910	14.21	12.85	15.83	2.98
1914	10.55	8.92	12.94	4.02
1918	3.12	2.45	4.21	1.76
1923	5.87	4.33	8.56	4.23
1933	17.25	13.19	24.91	11.72
1940	22.14	18.67	27.33	8.66
1945	1.03	0.78	1.45	0.67
1955	28.45	25.12	32.78	7.66
1968	24.33	21.08	28.56	7.48
1990	31.22	28.45	34.67	6.22
2020	8.55	7.66	9.66	1.98
2024	6.12	5.33	7.21	1.88

5 Discussion

5.1 What the Ensemble Measures

The Wright et al. thesis is **confirmed** for its statistical mechanics but **rejected** for its error model. Ensemble averaging does reduce variance by $1/\sqrt{N}$ and improve SNR by $\sqrt{N}$, but the "noise" being canceled is not idiosyncratic random error. It is *distributed epistemic disagreement* about genuinely ambiguous historical states.

This distinction matters for CAMS interpretation. A high SEM in 1923 Germany does not mean the raters are sloppy; it means the historiography of the Ruhr occupation, hyperinflation, and Beer Hall Putsch is genuinely contested. The ensemble mean captures the *central tendency of expert judgment*, while the envelope captures the *confidence interval of historical knowledge*.

5.2 Implications for CAMS v3.2-R

1. **Canonical estimator:** The ensemble mean remains the optimal point estimate. No alternative aggregator (median, trimmed mean) offers advantage.

2. **Confidence qualifier:** All CAMS diagnostics should report the SEM as a reliability flag. We propose a three-tier system:

   - **Green** ($\mathcal{C} < 0.3$): High confidence, standard threshold application.
   - **Amber** ($0.3 \leq \mathcal{C} < 0.7$): Moderate confidence, widened threshold bounds.
   - **Red** ($\mathcal{C} \geq 0.7$): Low confidence, diagnostic suspended pending additional evidence.

3. κ **reliability:** The κ threshold system (Watch/Warning/Critical/Extreme) should incorporate uncertainty-adjusted bounds. A κ reading that crosses a threshold only in its upper bound should not trigger an alert.

4. **Executive Decoupling:** Detection accuracy improves from 99.56% (single rater) to 99.89% (ensemble), with 79% variance reduction. The signature remains robust but should be qualified by $\mathcal{C}(t)$.

5.3 Limitations

The analysis is limited to two countries and one ensemble size ($N = 5$). The USA envelope file uses deviation format rather than absolute values, requiring correction. The simulation assumes Gaussian error, which may not hold for skewed rater distributions. Cross-LLM ensemble testing (GPT, Gemini, Claude, Perplexity) is ongoing (McKern, 2026b) and may reveal model-specific bias patterns not captured here.

6 Conclusion

The ensemble mean in CAMS is statistically optimal: it reduces variance by exactly $1/\sqrt{N}$, improves SNR by $\sqrt{N}$, and is effectively unbiased. However, the Wright thesis's assumption of uncorrelated idiosyncratic noise is incorrect for historical LLM scoring. Ensemble disagreement is *epistemic contestation* — it rises during crises, correlates across nodes, and persists over time. This is not a flaw to be eliminated but a signal to be reported.

We introduce the **Contestation Index** $\mathcal{C}(t)$ as a normalized uncertainty metric that validates against known historiographical battlegrounds. We derive uncertainty-adjusted bounds for the κ criticality index and propose a three-tier confidence qualifier system for CAMS diagnostics. The ensemble does not merely cancel noise; it honestly quantifies where our historical priors are genuinely contested.

Data Availability

All datasets, code, and output files are available at `https://neuralnations.org/cams-ensemble-2026`. The master results table, contestation indices, and κ bounds are provided as CSV files in the supplementary materials.

References

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15.

McKern, K. (2026a). CAMS v3.2-R: Complex Adaptive Model of Societies — Audited and Patched Specification. *Neural Nations Working Paper*.

McKern, K. (2026b). CAMS-CORP v0.9-R: Corporate Complex Adaptive Systems — Independent Audit and Correction. *Neural Nations Working Paper*.

Wright, A. G., et al. (2024). Using large language models to score narrative text for psychological constructs: An ensemble approach. *Psychological Methods*.

A Supplementary Tables

A.1 Node-Specific Uncertainty (Germany)

Table 4: Germany: Mean SEM and SNR by Node

Node	Mean SEM	Mean V	Mean SNR	Rank (SEM)
Hands	0.284	8.12	28.6	1 (highest)
Helm	0.261	10.45	40.0	2
Flow	0.243	11.23	46.2	3
Shield	0.238	11.89	49.9	4
Stewards	0.231	11.56	50.0	5
Craft	0.225	12.34	54.8	6
Lore	0.219	12.01	54.8	7
Archive	0.212	12.45	58.7	8 (lowest)

A.2 Node-Specific Uncertainty (USA)

A.3 Temporal Autocorrelation of Uncertainty

Table 5: USA: Mean SEM and SNR by Node

Node	Mean SEM	Mean V	Mean SNR	Rank (SEM)
Hands	0.312	9.45	30.3	1 (highest)
Helm	0.298	10.12	34.0	2
Flow	0.287	11.56	40.3	3
Shield	0.276	11.89	43.1	4
Stewards	0.271	11.23	41.4	5
Craft	0.265	12.01	45.3	6
Archive	0.258	12.34	47.8	7
Lore	0.251	12.67	50.5	8 (lowest)

Table 6: Autocorrelation of System-Level SEM

Lag	Germany r	Germany p	USA r	USA p
1	0.824	< 0.001	0.847	< 0.001
2	0.756	< 0.001	0.789	< 0.001
3	0.712	< 0.001	0.734	< 0.001
4	0.678	< 0.001	0.691	< 0.001
5	0.645	< 0.001	0.656	< 0.001