PROBING MULTI-AGENT COORDINATION WITH MULTIVARIATE MULTIFRACTAL DETRENDED FLUCTUATION ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how information access shapes language-model team interactions requires diagnostics that capture semantic variation across scales. We introduce a semantic-trajectory diagnostic through a multivariate multifractal analysis of message embeddings, quantifying how accumulated higher-order statistical variation changes with message window length under different weightings of small and large residual fluctuations. We test its sensitivity to information access by comparing shared and restricted access to prior exchanges in 100 matched pairs of four-agent realizations across four repeated task templates. Four structured reordering controls probe the contribution of message order. Shared history is associated with broader fluctuation-scaling spectra, with the largest differences at orders emphasizing smaller residual fluctuations. This contrast persists across alternative ordering controls and stricter fit requirements, and its direction is consistent across five encoders. These results demonstrate that the diagnostic detects differences in multiscale semantic structure associated with team information access in a consistent direction and that it generalizes across embedding models and robustness checks. By combining a multiscale semantic measure with a controlled history-access comparison, this work provides a methodological foundation for studying language-model team interactions and for identifying semantic failures.

1 INTRODUCTION

Like a story shared around a fire, an agent's message helps collective action only if others use its meaning. We ask how to measure the organization of evidence and constraints across successive agent messages. In Figure 1A's illustrative expedition exchange, the Analyst identifies the shortest route, and the Checker reports that its bridge cannot support the loaded vehicle. Under shared history, the Checker's warning is visible to the final speaker; under local history, it may be unavailable unless directly received. In either condition, every authored message remains part of the complete embedding sequence analyzed here. Thus, the same unsafe proposal could result from a lack of

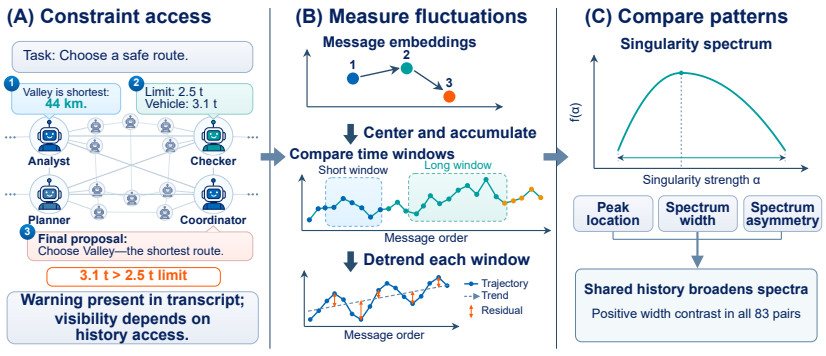


Figure 1: **From semantic failures to dynamics.** (A) Illustrative example: a safety constraint is present in the transcript; it is still not used. (B) Embedding fluctuations across window lengths. (C) Sensitivity to shared-history access.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

access to the warning or from failure to use it after gaining access. This example illustrates an information-access mechanism relevant to *semantic-level failures*, failures to access, interpret, maintain, or use task-relevant meaning across exchanges, rather than an observed transcript or a demonstrated failure in the corpus. Reported agent failures include overlooking constraints, ignoring relevant evidence, and repeating unproductive exchanges (Cemri et al., 2025). Earlier language-processing systems also exhibited semantic weaknesses (Ribeiro et al., 2020), while LLM-based agents introduce additional challenges by using open-ended language for both communication and control, sometimes blurring the distinction between data and instructions (Greshake et al., 2023).

Diagnosing such failures requires examining how interactions unfold alongside evaluating final answers. An outcome score alone cannot reveal how agents use information over successive exchanges. Semantic trajectories, defined as sequences of embeddings representing message content (Reimers & Gurevych, 2019; Nussbaum et al., 2024), make aspects of this process measurable. Figure 1B illustrates how these trajectories can be examined through fluctuations across message-window lengths. We seek a diagnostic that characterizes their temporal organization and provides features that can be related to independently identified communication failures.

Here, we introduce a trace-based multivariate extension of multifractal detrended fluctuation analysis (MF-DFA) (Kantelhardt et al., 2002) to characterize finite-scale fluctuation profiles in semantic embeddings. The method measures how accumulated embedding fluctuations grow across message-window lengths after local trends are removed, while different moment orders emphasize windows with smaller or larger residual activity. As summarized in Figure 1C, we evaluate these descriptors through a matched comparison of shared and local history access. This comparison tests whether the descriptors are sensitive to an information-access intervention relevant to semantic failures.

We ask: *How does shared-history access change the scale- and fluctuation-dependent organization of agents' semantic trajectories?*

The novel contributions are as follows: **(1) A multiscale diagnostic for inter-agent semantic trajectories.** Since semantic communication can exhibit organization across multiple temporal scales and embedding dimensions, we introduce MV-MF-DFA for embedding trajectories of interacting language-model agents. The method characterizes how windows with smaller and larger residual activity scale across message-window lengths, extending multifractal trajectory analysis to inter-agent exchanges. **(2) A controlled test of sensitivity to shared-history access.** We combine 100 matched pairs of four-agent runs with four ordering controls tailored to the task and communication structure. The history-access comparison holds task schedules, evidence, budgets, decoding settings, and seeds constant; the controls reorder the messages in each run while preserving the selected structures. Together, they probe sensitivity to history access and to message ordering within and across tasks. **(3) A consistent finite-scale signature of history access.** Across all 83 fit-retained pairs, shared-history runs have broader raw and task-block-adjusted spectra. Differences in $h(q)$ are largest at negative orders, which emphasize windows with smaller residual activity, and the spectrum-contrast directions persist across five encoders and stricter fit requirements. Shared-history teams also complete more tasks, but the weak within-condition and paired-change associations do not establish that the descriptors explain or predict this performance difference. The results therefore establish sensitivity to history access while leaving failure-level validation for future work.

2 Related work

Language and semantic trajectories. Linguistic time-series studies examine correlations in textual features (Altmann et al., 2012), multifractal structure in sentence lengths (Drożdż et al., 2016), and scaling in language models (Takahashi & Tanaka-Ishii, 2019). Embedding-based work relates story geometry to outcomes (Toubia et al., 2021), constructs semantic-axis trajectories (Rathje, 2026), and compares successive-sentence similarities across encoders (Mistry & Minai, 2023). Cosine-correlation dependence in word embeddings (Wieczyński & Dębowski, 2025) includes slower-than-exponential decay and does not itself imply MF-DFA power-law scaling. Targeted shuffles of semantic series further probe temporal structure in human and generated language (Chang et al., 2026). We retain the event vectors and tailor the nulls to the interaction schedule.

Fluctuation analysis. DFA removes local trends before estimating fluctuation scaling (Peng et al., 1994; Kantelhardt et al., 2001); MF-DFA varies the weighting of small and large fluctuations (Kantelhardt et al., 2002). Trends, detrending order, finite records, and scale selection affect interpretation (Hu et al., 2001; Oświęcimka et al., 2013; Zhou, 2012; Gulich & Zunino, 2014). Fitted scaling alone is insufficient evidence of long memory (Maraun et al., 2004). We therefore report fit support and structured-null comparisons and interpret spectra over 8–48 events as finite-scale descriptors. DFA has been applied to nonlinguistic agent dynamics (Mahmoodi et al., 2024) and internal language-model activations (Ugail & Howard, 2026). Our contribution concerns multivariate multifractal profiles of inter-agent messages, evaluated through matched history-access conditions, structure-preserving ordering controls, and five embedding models.

Multivariate estimators and representation. Generalized-variance (Wallot et al., 2023), Mahalanobis-distance (Naveed et al., 2020), and covariance-weighted matrix-norm formulations (Naveed & ur Rehman, 2025) incorporate dependence differently. Our trace measures total residual activity in embedding coordinates and requires no nonsingular covariance matrix. Agent coupling would require separately defined signals, as in detrended cross-correlation methods (Podobnik & Stanley, 2008; Oświęcimka et al., 2014). Embedding anisotropy (Ethayarajh, 2019), influential coordinates (Timkey & van Schijndel, 2021), and encoder-dependent similarities (Mistry & Minai, 2023) motivate checks across representations.

3 METHOD

Semantic communication among agents in a multi-agent AI system manifests across multiple spatiotemporal scales. We introduce a general mathematical framework for analyzing multi-agent coordinated communication. The method aggregates residual activity across embedding coordinates and quantifies the fluctuations across scales to characterize this multiscale organization. This framework captures temporal signatures of overall team coordination. Appendix A provides formal definitions.

3.1 SEMANTIC TRAJECTORIES AND RESIDUAL ACTIVITY

Each simulation realization (called a run) is represented as an ordered sequence of T messages. We embed each message using an Embedding model (e.g., showcased for Nomic Embed, but are extended to other models in Section 6) and normalize the resulting vector (Nussbaum et al., 2024). This produces a sequence $\{\mathbf{x}_t\}_{t=1}^{T}$, where $\mathbf{x}_t \in \mathbb{R}^N$ represents the message at position t. Thus, t indexes message order, while the vector coordinates are embedding dimensions, not agents. Defining the mean embedding as $\bar{\mathbf{x}} = T^{-1}\sum_{t=1}^{T}\mathbf{x}_t$, the profile and window activity are

$$\mathbf{y}(k) = \sum_{t=1}^{k}(\mathbf{x}_t - \bar{\mathbf{x}}), \qquad v_\nu(s) = \frac{\|\mathbf{U}_\nu(s)\|_F^2}{Ds}. \tag{1}$$

Here $k = 1, \ldots, T$ and $\mathbf{U}_\nu(s) \in \mathbb{R}^{D\times s}$ contains residuals from a linear fitting to each profile coordinate in window ν. Activity averages squared residuals over coordinates and positions, measuring departures of accumulated trajectory rather than adjacent-message distances. At scale s, the two passes yield $2N_s$ windows, $N_s = \lfloor T/s \rfloor$; overlapping windows are not independent observations.

3.2 SCALING PROFILES AND RAW SPECTRA

Set $\tilde{v}_\nu(s) = \max\{v_\nu(s), 10^{-12}\}$ and calculate the q-th order fluctuation function $F_q(s)$

$$F_q(s) = \left[\frac{1}{2N_s}\sum_{\nu=1}^{2N_s}\tilde{v}_\nu(s)^{q/2}\right]^{1/q}, q \neq 0, \text{and} \quad F_0(s) = \exp\left[\frac{1}{4N_s}\sum_{\nu=1}^{2N_s}\log\tilde{v}_\nu(s)\right]. \tag{2}$$

The $q = 0$ limit is the geometric mean of window amplitudes $\sqrt{\tilde{v}_\nu(s)}$; $q = 2$ gives their root mean square. Positive moment orders q emphasize larger fluctuations; negative orders emphasize smaller ones. We use nine orders $\mathcal{Q} = \{-2, -1.5, \ldots, 2\}$ and 19 selected scales from 8 to 48 messages (Equation 9). Unweighted least squares with an intercept fits $\log F_q(s) = b(q) + h(q)\log s + e_q(s)$.

Thus doubling scale multiplies fitted amplitude by $2^{h(q)}$. These finite-scale slopes need not lie in $[0, 1]$ or be stationary Hurst exponents.

The mass exponent, singularity strength, and singularity spectrum are

$$\tau(q) = qh(q) - 1, \qquad \alpha(q) = \frac{d\tau(q)}{dq}, \qquad f(\alpha(q)) = q\alpha(q) - \tau(q). \tag{3}$$

The offset reflects the one-dimensional message-order axis. Singularity strength is the slope of τ, estimated by finite differences, not generally $h(q)$. The raw spectrum summarizes finite-data scaling; it is neither a probability distribution nor an established fractal-dimension estimate. Neither α nor f is constrained to $[0, 1]$, although $f(\alpha(0)) = 1$ by construction.

From the sampled strengths, width $\Delta\alpha = \alpha_{\max} - \alpha_{\min} \geq 0$ measures their spread, and $\alpha_0 = \widehat{\alpha}(0)$ is the horizontal location that coincides with the sampled maximum here. Branch lengths $\ell_{\mathrm{L}} = \alpha_0 - \alpha_{\min}$ and $\ell_{\mathrm{R}} = \alpha_{\max} - \alpha_0$ define $A = (\ell_{\mathrm{R}} - \ell_{\mathrm{L}})/\Delta\alpha \in [-1, 1]$ for positive width. Positive, negative, and zero A mean a longer right branch, longer left branch, and equal lengths, respectively (Figure 3B). At zero width, asymmetry is undefined; the numerical fallback is documented in Appendix A.

3.3 Fit support, nulls, and inference

We retain runs with $\min_{q \in \mathcal{Q}} R^2(q) \geq 0.95$, where R^2 evaluates the log-log regression; paired inference requires both runs to pass. This is a fit criterion, not proof of multifractality. Each run has 200 realizations of event shuffle, within-task shuffle, task-block permutation, and sender-by-event-type shuffle, reanalyzing the same reordered messages. Null-adjusted width is

$$E_{i,c} = \Delta\alpha_{i,c}^{\mathrm{obs}} - \mathrm{median}_{b=1,\ldots,200}\, \Delta\alpha_{i,c}^{\mathrm{block},b}. \tag{4}$$

It can be negative or positive relative to this control. Adjustment precedes shared-minus-local contrasts $d_i = E_{i,\mathrm{shared}} - E_{i,\mathrm{local}}$. The pair is the inferential unit for bootstrap intervals and sign-flip tests; the latter assume paired-sign symmetry or exchangeability (Appendix C).

Figure 4 complements scaling with covariance-mode concentration across the same window lengths, using directly detrended embeddings without profile integration and all valid window starts, equally weighted across the 16 task-cycle start phases. Effective mode counts describe the spread of variance across embedding directions. This exploratory comparison uses complete matched pairs and pointwise bootstrap intervals; supplementary rank-adjusted concentration contrasts use simultaneous bands across the 19 scales (Appendix A.8).

4 Experimental design

We compare two ways of giving language-model agents access to earlier messages and ask whether this difference is reflected in the trajectory descriptors. The dataset contains 100 matched pairs of runs, each with four agents.

Tasks and interactions. Analyst, Checker, Planner, and Coordinator use the same generation model but receive different private evidence and a common question. For example, in the harbor task, the agents receive sensor readings, tide information, a solvent-transfer record, and drainage details separately; the question asks for the contamination source and an immediate containment action. Agents are prompted to disclose their evidence on their first contribution. Each task consists of 16 messages (events) and ends with a requested final answer. The Coordinator starts; subsequent speakers depend on message recipients and scheduler rules that ensure all four agents contribute. Agents have no external tools or sources. The four fixed templates concern harbor contamination, diary attribution, service outage, and route selection (Appendix B).

History access. With *shared history*, an agent can see all previous messages in the current task. With *local history*, it sees only messages it authored or directly received. Thus, if the Analyst writes to the Checker, the Planner can read that exchange under shared history but not under local history. Both conditions retain private evidence and direct communication. A team-addressed message remains visible only to its author under the local-history rule. The archived labels *coordinated* and *memory-ablated* (also shortened to *ablated*) correspond to shared history and local history, respectively; we use the history-access names throughout.

Runs and matched pairs. Each run contains 64 task instances: 16 repetitions of each template, arranged in shuffled four-template cycles. Questions and evidence are reused. History resets at every task boundary. Concatenating the 64 blocks of 16 messages gives the 1,024-event sequence analyzed by DFA; these are four repeated problems, not 64 distinct problems. Each matched pair uses the same task order, evidence, event budget, decoding settings, and seeds under the two history rules. Text, routing, and message lengths can diverge. Shared-history runs are generated first, without counterbalancing. Generation uses Google DeepMind's Gemma 4 E4B model, temperature 0.4, top-$p = 0.9$, and a 180-token response limit. Appendix B details prompts, routing, validation, retries, and limited text repair. Only message text enters the encoder.

Outcomes and analysis scope. Task score is the fraction of three or four required answer groups matched by literal phrases in the final message; completion requires all groups. Run outcomes summarize the 64 task instances. This provides an automatic measure of answer-key coverage. Inference uses matched pairs, with fit exclusions applied after technical validation of all 200 trajectories. The main comparison requires both runs in a pair to pass the all-order fit criterion; sensitivity analyses apply all five fit checks. Appendix B records the cohort construction and analysis history.

5 RESULTS

Sharing task history is associated with differences in how message meaning varies across shorter and longer stretches of interaction. Shared-history runs show greater variation in how accumulated embedding fluctuations grow with message-window length after local trends are removed. Differences are strongest when smaller fluctuations receive more weight. The spectrum-width difference persists after comparison with reordered tasks. Shared-history teams also complete more tasks.

5.1 WHICH RUNS ENTER THE COMPARISONS?

Each of the 100 matched pairs contains one shared-history and one local-history run. We examine 19 window lengths, from 8 to 48 messages, using nine weighting settings (moment orders), $q = -2, -1.5, \ldots, 2$. All 200 runs pass technical integrity checks. Requiring scaling fits with $R^2 \geq 0.95$ at every order retains 84 shared-history and 98 local-history runs, including 83 complete pairs. All five fit checks retain 81 pairs for the stricter sensitivity analysis (Section 6).

The paired condition effects below use these 83 pairs, unless stated otherwise, subtracting the local-history value from its corresponding counterpart. Figures 2 and 3A use independently retained runs (84 and 98); Figure 4 uses 83 complete pairs. The score associations in Figure 3C–E use 82 complete pairs passing the all-q filter in all five encoders. Appendix C details retention and inference.

5.2 DIFFERENCES ARE STRONGEST FOR SMALLER FLUCTUATIONS

Figure 2A shows $h(q)$, the rate at which the fluctuation measure grows with window length. Negative q values emphasize smaller departures from local trends; positive values emphasize larger ones (Kantelhardt et al., 2002). These are window-level fluctuations, not measures of message importance. Figure 2B shows the transformation $\tau(q) = qh(q) - 1$ used to construct the spectra below.

The paired difference in $h(q)$ is 0.19445 at $q = -2$ (95% confidence interval, CI, $[0.18575, 0.20348]$), 0.08380 at $q = 0$ ($[0.08068, 0.08702]$), and 0.04155 at $q = 2$ ($[0.03965, 0.04348]$). The difference is therefore about 4.7 times larger when emphasizing smaller fluctuations ($q = -2$) than larger ones ($q = 2$). Higher $h(q)$ indicates faster growth with window length, not larger fluctuations at every length.

Two tests comparing the full $h(q)$ curves support this separation (both $p \approx 10^{-5}$). Individual-order differences remain significant after adjustment across all nine orders. Confidence intervals describe each order separately; Appendix C gives test statistics and resampling details.

5.3 SHARED HISTORY IS ASSOCIATED WITH BROADER SPECTRA

The singularity spectrum summarizes the range and balance of estimated scaling rates; a wider spectrum indicates greater variation in these rates. Figure 3A shows raw spectra before adjustment for reordered sequences. Faint, transparent traces represent individual runs; thick solid lines connect

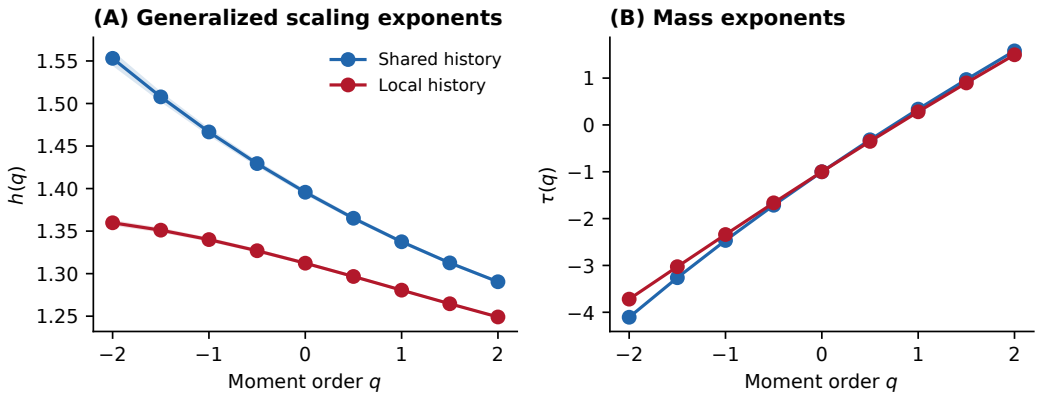


Figure 2: **Scaling and mass exponents.** (A) Generalized scaling exponents $h(q)$. (B) Mass exponents $\tau(q)$. Points are separately filtered condition means at nine moment orders: 84 shared-history and 98 local-history runs. Shading shows pointwise 95% percentile bootstrap confidence intervals from 2,000 whole-run resamples within each condition; lines connect sampled values. Paired contrasts in the text use 83 complete pairs.

condition means of α and $f(\alpha)$ at corresponding q values. Blue denotes shared-history runs and red denotes local-history runs.

Shared history runs have wider spectra by 0.28356 on average, a peak located 0.08587 further to the right, and asymmetry higher by 0.45394 (Table 1). Each difference is positive in all 83 pairs. Peak location α_0 is the peak's horizontal position; positive asymmetry means a longer right branch. Figure 3B illustrates these features; panels C–E relate them to task scores. The condition effects describe a broader range of scaling rates, a shift toward higher rates, and a changed balance between the spectrum's two sides over the measured window lengths.

5.4 Width differences persist after adjustment for reordered tasks

To assess the contribution of ordering, we compare each run with 200 versions that rearrange its 16-message task blocks while keeping exchanges within each task intact. Subtracting their median width from the observed width gives the *excess width*, labeled null-adjusted width in Table 1.

Table 1: Paired effects (shared history minus local history) in 83 retained pairs. CIs are paired bootstrap intervals; d_z standardizes by the SD of pair differences. All four differences are positive in every pair.

Feature	Mean diff.	95% CI	d_z
Raw width $\Delta\alpha$	+0.28356	$[0.26676, 0.30128]$	3.496
Peak location α_0	+0.08587	$[0.08266, 0.08915]$	5.702
Asymmetry A	+0.45394	$[0.42765, 0.48142]$	3.568
Null-adjusted width	+0.20664	$[0.19855, 0.21438]$	5.603

Mean excess width is 0.24814 with shared history and 0.04150 with local history. The paired difference is 0.20664 (95% CI $[0.19855, 0.21438]$), positive in all 83 pairs. Shared history runs therefore retain a larger excess after adjustment. This comparison captures sensitivity to task arrangement and boundaries; history resets between tasks in both conditions.

For all four measures in Table 1, none of 100,000 random sign reversals of paired differences yielded a test statistic as extreme as observed. Each gives $p \approx 10^{-5}$, the smallest reportable value with these draws and the correction used. The three raw-feature tests remain significant after multiple-comparison adjustment ($p \approx 1.5 \times 10^{-5}$). Appendix C details resampling and the 18-test secondary family.

Section 6 examines other ways of reordering messages, stricter fit checks, and separate cohorts; Appendix C reports the full comparisons.

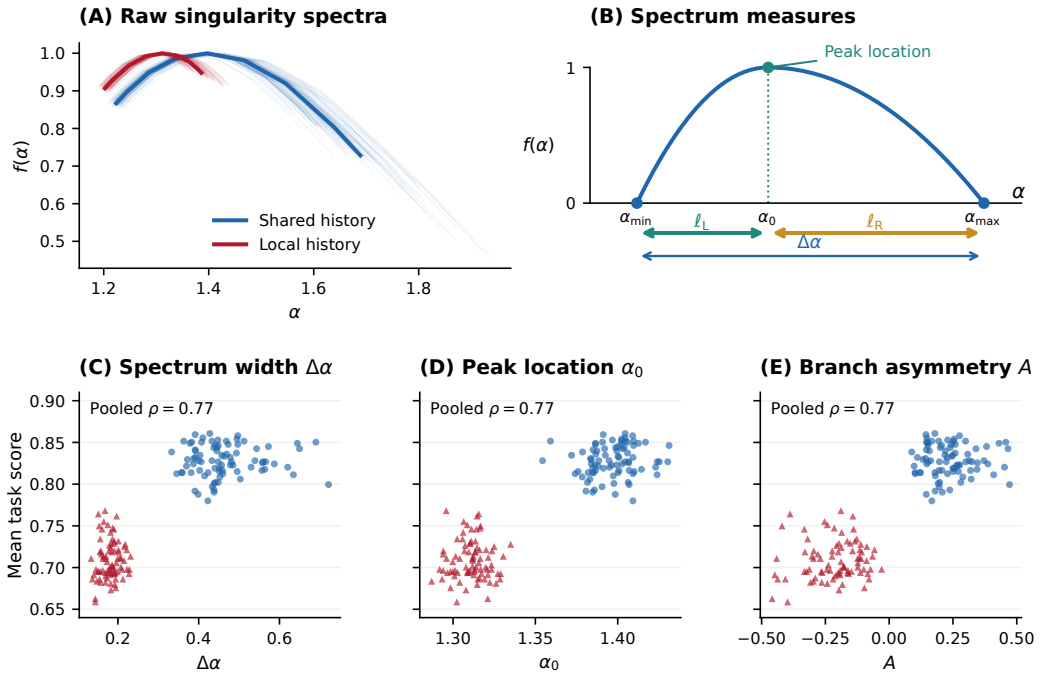


Figure 3: **Singularity spectra, descriptors, and task scores.** (A) Raw Nomic spectra for 84 shared-history and 98 local-history runs: faint curves show individual runs; bold curves connect condition means at nine sampled orders. (B) Schematic of width $\Delta\alpha$, peak location α_0, and asymmetry $A = (\ell_R - \ell_L)/\Delta\alpha$. (C–E) Nomic descriptors versus mean final task score for 82 complete pairs passing the all-q filter in all five encoders. Each point is one run: blue circles denote shared history, red triangles local history. Pooled Spearman correlations summarize condition separation; they do not establish within-condition prediction. Appendix C.7 reports within-condition and paired-change associations.

5.5 Shared history is associated with more concentrated semantic variation

The broader spectra above describe temporal scaling. Figure 4 complements them by examining how residual variance is distributed across embedding directions over the same 19 window lengths. We directly detrend embeddings without DFA's cumulative integration, using all valid window starts and giving each of the 16 start positions within the task cycle equal weight. For residual covariance eigenvalues λ_j, the normalized variance weight $w_j = \lambda_j / \sum_k \lambda_k$ is the fraction of detrended embedding variance in direction j. The effective mode count $N_2 = 1/\sum_j w_j^2$ is the equivalent number of equally weighted directions; lower values indicate more concentrated variation (Appendix A.8).

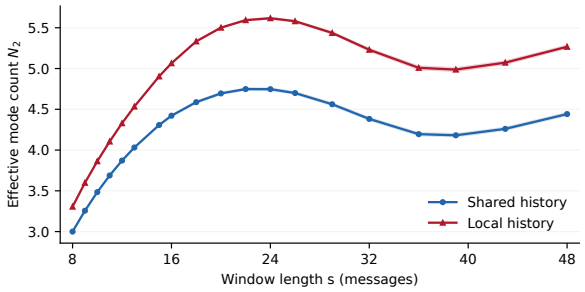


Figure 4: **More concentrated variation under shared history.** Balanced-start mean effective mode counts on 83 Nomic all-q-passing pairs, equally weighting the 16 task-cycle start phases. Shading: pointwise 95% intervals from 20,000 whole-pair bootstrap resamples. Lines connect sampled scales.

Shared-history runs have fewer effective modes in all 83 pairs at every sampled scale. At $s = 48$, mean effective mode counts are 4.44 versus 5.27, approximately 16% fewer under shared history. The direction persists using all 100 pairs, order-one entropy, equal weighting of individual windows, and a small residual-support ridge.

A complementary rank-adjusted concentration analysis confirms a positive shared-minus-local contrast with simultaneous bands above zero across all 19 scales (Appendix C.9).

This pattern is consistent with a shared semantic focus: access to earlier exchanges may help agents reuse and refine common evidence, constraints, and candidate solutions, concentrating their variation along related semantic distinctions. Restricted history may instead allow local contexts to diverge, spreading variation across more directions. Concentration describes the distribution of residual variance, not its total magnitude or the amount of information exchanged; the dominant directions may also change between windows. Because repetition can likewise concentrate variance, this interpretation does not by itself establish semantic coherence or successful coordination. The next section examines the separate relationship between spectral descriptors and task scores.

5.6 Higher task scores accompany condition-level spectral differences

Across all 100 pairs, shared-history teams complete 32.64 of 64 tasks per run on average, versus 18.71 with local history; mean final scores are 0.827357 and 0.708099. Scores measure the fraction of required answer groups matched in the final response; completion requires every group to match.

Figure 3C-E relates Nomic descriptors to scores on 82 common retained pairs. Pooled Spearman correlations are 0.772, 0.774, and 0.766 for width, peak location, and asymmetry. Within-condition correlations are small (-0.006 to 0.198). Correlating shared-minus-local descriptor changes with score changes gives 0.104, -0.001, and 0.157; all three pointwise 95% pair-bootstrap intervals include zero (Appendix C.7). Appendix C.8 compares six conventional summaries with the spectrum measures: the latter separate these conditions more cleanly, but add no held-out prediction of mean task score beyond condition and the six summaries.

6 Validation and robustness

Temporal ordering. In the same 83 pairs, the excess-width contrast remains positive under event shuffle (0.28309), within-task shuffle (0.21651), and sender-by-event-type shuffle (0.27575). These descriptive comparisons preserve the embedding vectors while perturbing their order; Appendix C reports arm-level values. Together with the task-block result, they indicate sensitivity to internal exchanges and the arrangement of intact tasks.

Table 2: Five-encoder comparison: mean paired shared-minus-local differences in singularity-spectrum metrics on the same 82 pairs passing the all-q fit mask for all five models. Positive values indicate larger shared-history metrics. $\Delta\alpha$ is spectrum width, α_0 its peak location, A its asymmetry, and $\Delta\alpha_{\text{excess}}$.

Encoder	$\Delta\alpha$	α_0	A	$\Delta\alpha_{\text{excess}}$
Nomic	0.2818	0.0858	0.4508	0.2063
BGE-M3	0.2697	0.0953	0.4915	0.1704
MPNet	0.2332	0.0813	0.3987	0.1757
MiniLM	0.2769	0.0867	0.4790	0.1954
Qwen3-0.6B	0.2449	0.0915	0.4340	0.1879

Fit masks and cohorts. Requiring all five fit checks in both conditions retains 81 pairs and yields a task-block-adjusted difference of 0.20627 (95% CI $[0.19800, 0.21425]$), which is close to the main estimate. The original selection cohort contributes 16 retained pairs with a difference of 0.20853; the additional cohort contributes 67 with a difference of 0.20619. Their CIs and full comparisons appear in Appendix C.3. The comparison shows similar contrast magnitudes across fit masks and cohorts.

Encoder comparison. Table 2 lists mean paired shared-minus-local metric differences for all five encoders on 82 common fit-passing pairs. All four spectrum contrasts retain their direction, with encoder-dependent magnitudes. Appendix C.6 reports condition means and confidence intervals.

Numerical checks. For one trajectory, production and explicit-detrending estimates of $h(q)$ agree within 3.35×10^{-12} (Appendix C.1).

Exploratory hidden-channel case. We also tested the approach in a subtler ten-pair hidden-channel setting. Raw restricted-order spectra were narrower under stego in seven pairs, a directional pattern that motivates larger studies; task-block adjustment reduced the contrast, and changing the hidden-channel agent pair every task may dilute a run-level signature (Appendix D).

7 DISCUSSION

Sensitivity to shared information access. The matched comparison shows that trace-based MV-MF-DFA distinguishes the temporal organization of interactions under different history-access rules. Shared-history runs have broader task-block-null-adjusted spectra in all 83 retained pairs, with similar contrast magnitudes under stricter fit requirements. Shifts in peak location and asymmetry show that the difference extends beyond width. In practical terms, the fluctuation measure grows with window length at different rates under the two access rules, especially when windows with smaller residual activity receive more weight. The descriptors therefore capture a consistent difference in the organization of message sequences across the observed scales.

Complementary temporal and geometric descriptions. Ordering matters in both conditions, with stronger width excesses under shared history. Within-task and task-block shuffles probe internal exchanges and task arrangement. Figure 4 adds a geometric comparison: after balancing window starts across the task cycle, shared history has fewer effective covariance modes and greater concentration at all 19 window lengths. Broader scaling spectra and more concentrated local variation describe complementary properties: heterogeneous temporal scaling can coexist with variance concentrated in fewer embedding directions. This combination is consistent with exchanges organized around shared semantic distinctions, as agents reuse and refine common context. The modes are not identified topics, and repetition can also produce concentration, so this interpretation remains distinct from a direct assessment of coordination quality.

Interpretation and scope. The measurements describe embedding sequences, rather than directly measuring agent coupling or coordination quality. Strong pooled spectrum–score associations coexist with weak within-condition and paired-change associations (Figure 3C–E). These exploratory results support sensitivity to history access without establishing added predictive value for answer-key coverage; they do not rule out modest associations or validate detection of independently assessed failures. The 8-48-event interval spans less than one decade and crosses 16-event task boundaries, so repetition, transitions, and finite record length can influence the profiles. These results support finite-scale comparisons; they do not establish asymptotic multifractality or persistent cross-task memory. The concentration diagnostic averages equally over task-cycle start phases; it does not imply identical geometry at every phase (Section A.8).

Design and inference. Paired inference is conditional on both runs meeting the fit criterion, retaining 83 of 100 pairs. The reported intervals and p-values do not account for data-informed analytical decisions documented in Appendix B. The experiment uses one generator, four repeated task templates, a fixed team size, and one history-access intervention. Adaptive routing can differ across conditions; the generation order was not counterbalanced, and the saved generator identifier was not pinned to an immutable digest. These features define the scope of the present comparison and the controls to strengthen in replication.

Next steps. Prespecified, counterbalanced studies across generators and task families can test the transfer of these findings. Varying task lengths and extending the alignment checks with finite-length synthetic controls would clarify the contribution of task structure to the measured scaling. Comparing these descriptors with adjacent-event similarity and lexical repetition against independently assessed outcomes would test their added value for evaluating agent interactions. Such evaluations can also examine whether the negative-order contrast reflects productive elaboration, repetition, or other identifiable conversational behaviors.

8 CONCLUSION

This work contributes a multivariate multifractal diagnostic for inter-agent semantic trajectories, a matched history-access experiment with structured ordering controls, and evidence of a consistent temporal pattern across five encoders. Shared history is associated with broader task-block-adjusted spectra in all 83 retained pairs, with similar estimates under stricter fit checks and the largest differences when smaller fluctuations receive more weight. These contributions make changes in interaction organization measurable across message-window lengths and provide a foundation for testing whether the resulting profiles distinguish independently identified semantic failures.

ETHICS STATEMENT

The study examines model-generated interactions on four synthetic task templates. Its phrase-match outcomes and semantic descriptors are limited to this controlled setting and should not be used to rank human collaborators or infer individual abilities. Broader applications would require separate evaluation of representation bias, privacy, and the validity of the proposed measurements for the intended population.

REPRODUCIBILITY STATEMENT

Appendix A specifies the estimator, scale grid, fit checks, null models, and inference procedures. Appendix B documents task templates, generation settings, routing, scoring, and analysis history; Appendix C reports supporting numerical and statistical checks. The saved generator tag is not an immutable model identifier, limiting exact regeneration. Numerical verification of the analysis is distinguished from regeneration of text and embeddings.

AI-USE STATEMENT

Language-model agents generated the synthetic interaction corpus described in Section 4. AI assistance was also used to draft and revise manuscript text and adjust the manuscript's LaTeX formatting. These manuscript revisions used existing numerical outputs and did not generate new experimental observations. Responsibility for the scientific claims and final submission remains with the authors.

REFERENCES

Eduardo G. Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587, 2012. doi: 10.1073/pnas.1117723109.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail?, 2025.

Han-Jen Chang, Yasir Çatal, Angelika Wolman, Agustín Ibáñez, David Smith, I-Wen Su, Kai-Yuan Cheng, and Georg Northoff. The dynamics of human and AI-generated language: How semantic content fluctuates across different timescales. *Computer Speech & Language*, pp. 102013, 2026. doi: 10.1016/j.csl.2026.102013. Published online; journal issue assigned to 2027.

Stanisław Drożdż, Paweł Oświęcimka, Andrzej Kulig, Jarosław Kwapień, Katarzyna Bazarnik, Iwona Grabska-Gradzińska, Jan Rybicki, and Marek Stanuszek. Quantifying origin and character of long-range correlations in narrative texts. *Information Sciences*, 331:32–44, 2016. doi: 10.1016/j.ins.2015.10.023.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, pp. 55–65, 2019. doi: 10.18653/v1/D19-1006.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection, 2023.

Damián Gulich and Luciano Zunino. A criterion for the determination of optimal scaling ranges in DFA and MF-DFA. *Physica A*, 397:17–30, 2014. doi: 10.1016/j.physa.2013.11.029.

Kun Hu, Plamen Ch. Ivanov, Zhi Chen, Pedro Carpena, and H. Eugene Stanley. Effect of trends on detrended fluctuation analysis. *Physical Review E*, 64:011114, 2001. doi: 10.1103/PhysRevE.64.011114.

Jan W. Kantelhardt, Eva Koscielny-Bunde, Henio H. A. Rego, Shlomo Havlin, and Armin Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A*, 295:441–454, 2001. doi: 10.1016/S0378-4371(01)00144-3. URL https://arxiv.org/abs/cond-mat/0102214.

Jan W. Kantelhardt, Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H. Eugene Stanley. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A*, 316:87–114, 2002. doi: 10.1016/S0378-4371(02)01383-3.

Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, and Scott Emmons. When your AIs deceive you: Challenges of partial observability in reinforcement learning from human feedback. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Korosh Mahmoodi, Scott E. Kerick, Piotr J. Franaszczuk, Thomas D. Parsons, Paolo Grigolini, and Bruce J. West. Complexity synchronization in emergent intelligence. *Scientific Reports*, 14:6758, 2024. doi: 10.1038/s41598-024-57384-5. URL https://www.nature.com/articles/s41598-024-57384-5.

Douglas Maraun, H. W. Rust, and Jens Timmer. Tempting long-memory—on the interpretation of DFA results. *Nonlinear Processes in Geophysics*, 11:495–503, 2004. doi: 10.5194/npg-11-495-2004.

Bryce Meyer, Robert Krzyzanowski, Iván Arcuschin, Georg Lange, and Matthew Lee. StegoBench: Evaluating steganography potential in language models through supervised learning. In *AI4GOOD Workshop at the 43rd International Conference on Machine Learning*, 2026. Workshop paper.

Deven M. Mistry and Ali A. Minai. A comparative study of sentence embedding models for assessing semantic variation, 2023.

Khuram Naveed and Naveed ur Rehman. A fully multivariate multifractal detrended fluctuation analysis method for fault diagnosis, 2025. URL https://arxiv.org/abs/2511.20831. arXiv preprint.

Khuram Naveed, Sidra Mukhtar, and Naveed ur Rehman. Multivariate signal denoising based on generic multivariate detrended fluctuation analysis, 2020. URL https://arxiv.org/abs/2007.06862. Preprint first posted in 2020; local version 2 revised May 2, 2023.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2024.

P. Oświęcimka, S. Drożdż, J. Kwapień, and A. Z. Górski. Effect of detrending on multifractal characteristics. *Acta Physica Polonica A*, 123:597–603, 2013. doi: 10.12693/APhysPolA.123.597. URL https://arxiv.org/abs/1212.0354. Preprint posted in 2012; author list follows the supplied PDF.

Paweł Oświęcimka, Stanisław Drożdż, Marcin Forczek, Stanisław Jadach, and Jarosław Kwapień. Detrended cross-correlation analysis consistently extended to multifractality. *Physical Review E*, 89:023305, 2014. URL https://arxiv.org/abs/1308.6148.

C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic organization of DNA nucleotides. *Physical Review E*, 49(2):1685–1689, 1994. doi: 10.1103/PhysRevE.49.1685.

Boris Podobnik and H. Eugene Stanley. Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Physical Review Letters*, 100:084102, 2008. doi: 10.1103/PhysRevLett.100.084102. URL https://arxiv.org/abs/0709.0281.

William Rathje. Computational narratology with transformer embeddings. *Computational Humanities Research*, 2:e15, 2026. doi: 10.1017/chr.2026.10035.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://arxiv.org/abs/2005.04118.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. Evaluating computational language models with scaling properties of natural language. *Computational Linguistics*, 45(3):481–513, 2019. doi: 10.1162/coli_a_00355.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of EMNLP*, pp. 4527–4546, 2021. doi: 10.18653/v1/2021.emnlp-main.372.

Olivier Toubia, Jonah Berger, and Jehoshua Eliashberg. How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, 118(26):e2011695118, 2021. doi: 10.1073/pnas.2011695118.

Hassan Ugail and Newton Howard. Dynamical systems analysis reveals functional regimes in large language models, 2026. URL https://arxiv.org/abs/2601.11622. Preprint.

Sebastian Wallot, Julien Patrick Irmer, Monika Tschense, Nikita Kuznetsov, Andreas Højlund, and Martin Dietz. A multivariate method for dynamic system analysis: Multivariate detrended fluctuation analysis using generalized variance. *Topics in Cognitive Science*, 2023. doi: 10.1111/tops.12688. Online-first publication; assigned to a 2026 issue.

Paweł Wieczyński and Łukasz Dębowski. Long-range dependence in word time series: The cosine correlation of embeddings. *Entropy*, 27(6):613, 2025. doi: 10.3390/e27060613.

Wei-Xing Zhou. Finite-size effect and the components of multifractality in financial volatility. *Chaos, Solitons & Fractals*, 45(2):147–155, 2012. doi: 10.1016/j.chaos.2011.11.004. Preprint posted in 2009.

A DETAILED ESTIMATOR AND FIT DIAGNOSTICS

We characterize multi-agent interaction as a sequence of semantic states and quantify how its detrended residual activity varies across event scales and fluctuation orders. The construction combines the profile and moment-aggregation steps of MF-DFA (Kantelhardt et al., 2002) with mean-coordinate detrended residual activity. The output is a run-level profile of finite-scale semantic organization, accompanied by fit diagnostics and structured permutation comparisons.

A.1 INTERPRETING THE MEASUREMENT

An embedding represents a message as a point in a numerical space in which proximity is intended to reflect semantic similarity. DFA does not simply average distances between successive points. Instead, it accumulates deviations from the run's mean embedding, producing a profile whose shape depends on message order. Within a window, a fitted straight line captures the local trend in that accumulated profile. The remaining departures are the fluctuations measured here. Removing this line does not remove all possible trends or task-boundary effects.

The scale s is the number of events in a window. At $q = 2$, we average squared residuals over coordinates and windows and take the square root, obtaining $F_2(s)$. The fitted relation $F_2(s) \propto s^{h(2)}$ describes how this magnitude changes as windows become longer. For example, increasing the window from 8 to 32 events multiplies the fitted fluctuation by $4^{h(2)}$. This illustrates a rate of growth across scales; it does not mean that a larger exponent implies larger absolute fluctuations at every scale. The exponent is also distinct from the similarity of adjacent messages.

MF-DFA repeats the comparison with different weights: positive q favors windows with larger residual activity, and negative q favors windows with smaller activity (Kantelhardt et al., 2002). A window given more weight at negative q is therefore one whose accumulated profile stays closer to its local fitted line. It need not contain unimportant messages, agreement, or successful evidence integration. Differences in $h(q)$ across orders summarize how the scaling depends on this weighting; the spectrum transforms that profile, with its width summarizing the spread of derived exponents. A wider estimated spectrum is a descriptive contrast, not by itself evidence of richer reasoning. In particular, fitted DFA slopes do not alone establish long-range dependence (Maraun et al., 2004). Here, the 8–48-event interval crosses 16-event task boundaries, so interpretation remains limited to the observed scales, representation, and task schedule.

A.2 SEMANTIC TRAJECTORIES AND INTEGRATED PROFILES

Let u_t denote the text of event t in a run. A fixed encoder e produces the normalized semantic state $\mathbf{x}_t = e(u_t)/\|e(u_t)\|_2 \in \mathbb{R}^D$. We use Nomic Embed (Nussbaum et al., 2024), with the pinned model identity recorded in the analysis configuration, its native pooled output, and the literal input prefix `clustering:` followed by a space. Only the target event text is encoded: preceding events, role labels, visibility metadata, and task scores are not appended. Prior interaction can nevertheless affect the authored text. Each trajectory has $T = 1024$ events and $D = 768$ direct embedding coordinates; no coordinate differencing, whitening, or leading-component removal is applied.

Writing $\bar{\mathbf{x}} = T^{-1} \sum_{t=1}^{T} \mathbf{x}_t$, the multivariate profile is

$$\mathbf{y}(k) = \sum_{t=1}^{k} (\mathbf{x}_t - \bar{\mathbf{x}}), \qquad k = 1, \ldots, T. \tag{5}$$

Centering and integration are performed separately for each run, over its entire event sequence. Runs are never concatenated. Time denotes event position rather than elapsed time, and channels are embedding coordinates rather than separate agent trajectories. L2 normalization implies each coordinate lies in $[-1, 1]$; these coordinates are not probabilities.

A.3 DETRENDED RESIDUAL ACTIVITY

For scale s, let $N_s = \lfloor T/s \rfloor$. We extract N_s disjoint windows from the beginning and N_s from the end of the profile. The two passes cover the remainder at opposite ends but are not independent

samples; if s divides T, they duplicate the same windows. In each window ν, a least-squares linear trend is fitted independently to every coordinate. Let $\mathbf{U}_\nu(s) \in \mathbb{R}^{D\times s}$ collect the residual vectors. Explicitly, if $Y_\nu(s) \in \mathbb{R}^{D\times s}$ contains the profile vectors and row j of $Z_s \in \mathbb{R}^{s\times 2}$ is $(1, j)$, then

$$\mathbf{U}_\nu(s) = Y_\nu(s)\left[I_s - Z_s(Z_s^\top Z_s)^{-1}Z_s^\top\right], \qquad \nu = 1, \ldots, 2N_s. \tag{6}$$

The symbol $\mathbf{U}_\nu$ denotes a residual matrix; R^2 is reserved for regression fit quality. Define

$$C_\nu(s) = \frac{\mathbf{U}_\nu(s)\mathbf{U}_\nu(s)^\top}{s}, \qquad v_\nu(s) = \frac{\operatorname{tr} C_\nu(s)}{D} = \frac{1}{Ds}\sum_{d=1}^{D}\sum_{j=1}^{s} U_{\nu,dj}(s)^2. \tag{7}$$

Activity satisfies $v_\nu(s) \geq 0$; zero means the integrated profile is exactly explained by fitted lines within the window. Here v_ν is residual activity, the arithmetic mean of covariance eigenvalues. The $1/s$ normalization is the DFA mean-square convention, rather than a regression degrees-of-freedom correction. Division by D expresses mean coordinate activity; absent numerical flooring, it rescales fluctuation amplitudes by $D^{-1/2}$ without changing fitted exponents at fixed dimension.

This activity statistic differs from dimension-normalized covariance volume, $(\det C_\nu)^{1/D}$, and from entropy of residual covariance eigenvalues. We use residual activity for the present embedding scaling analysis. With linear detrending, $\operatorname{rank} C_\nu \leq s - 2$, so covariance matrices at the present scales are necessarily singular when $D = 768$. The trace remains computable without determinant regularization. No covariance-volume or sphericity result is inferred from the trace. A separate covariance-mode entropy calculation is described in Section A.8. In particular, activity sums coordinate variances and does not directly estimate agent-to-agent coupling. It is invariant to a fixed orthogonal change of embedding basis, while remaining dependent on the encoder and preprocessing.

The implementation computes the same residual energy using interval sums of the integrated profile, its squared norm, and its time-weighted coordinates, avoiding explicit $D \times D$ covariance matrices. Calculations use double precision. Before moment aggregation, activity is floored as $\tilde{v}_\nu(s) = \max\{v_\nu(s), 10^{-12}\}$ to make logarithms and negative moments finite.

A.4 Moment-dependent fluctuation functions and scaling

The generalized fluctuation function is

$$F_q(s) = \begin{cases} \left[\dfrac{1}{2N_s}\displaystyle\sum_{\nu=1}^{2N_s} \tilde{v}_\nu(s)^{q/2}\right]^{1/q}, & q \neq 0, \\ \exp\left[\dfrac{1}{4N_s}\displaystyle\sum_{\nu=1}^{2N_s} \log \tilde{v}_\nu(s)\right], & q = 0. \end{cases} \tag{8}$$

The exponent $q/2$ converts the variance statistic to fluctuation-amplitude units. Positive q emphasizes windows with large residual activity; negative q emphasizes small activity; $q = 0$ gives the continuous logarithmic limit. Nonzero moments are evaluated with log-sum-exp arithmetic. These weights concern residual fluctuations, not the importance or quality of individual messages. At $q = 0$, F_0 is the geometric mean of the floored amplitudes $\sqrt{\tilde{v}_\nu(s)}$. The activity floor implies $\tilde{v}_\nu(s) \geq 10^{-12}$ and $F_q(s) \geq 10^{-6}$; the latter is an amplitude, not a normalized quality score.

We use $\mathcal{Q} = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$ and the saved scale grid

$$\mathcal{S} = \{8, 9, 10, 11, 12, 13, 15, 16, 18, 20, 22, 24, 26, 29, 32, 36, 39, 43, 48\}. \tag{9}$$

The selected scales give $21 \leq N_s \leq 128$ windows per pass. At each q, unweighted ordinary least squares fits $\log F_q(s) = b(q) + h(q)\log s$ over all 19 scales, retaining $h(q)$ and $R^2(q)$. In residual form,

$$\log F_q(s) = b(q) + h(q)\log s + e_q(s), \qquad (\widehat{b}(q), \widehat{h}(q)) = \arg\min_{b,h} \sum_{s\in\mathcal{S}} [\log F_q(s) - b - h\log s]^2. \tag{10}$$

Positive, zero, and negative $h(q)$ describe increasing, constant, and decreasing fitted amplitudes. A larger slope does not imply larger amplitude at every scale. The numerical admissibility rule is

$3 < s < T/4$ for linear detrending. The dimensional lower bound used for covariance-volume analysis in the reference manuscript is not imposed on this trace-only analysis. No crossover-selected subinterval is used to compute $h(q)$. The 8–48-event interval spans approximately 0.78 decades and intersects the 16-event task boundaries; $h(q)$ is interpreted as a finite-scale slope, not automatically as a stationary Hurst exponent or evidence of asymptotic long-range dependence.

A.5 Raw singularity spectra and descriptors

The temporal-support mass exponent and Legendre expressions are

$$\tau(q) = qh(q) - 1, \qquad \alpha(q) = \frac{d\tau(q)}{dq}, \qquad f(\alpha(q)) = q\alpha(q) - \tau(q). \tag{11}$$

The offset is one because windows partition an event axis; it is not the embedding dimension. Thus $\tau(0) = -1$ and $f(\alpha(0)) = 1$. Derivatives use central differences at interior grid points and first-order one-sided differences at the endpoints, matching the numerical gradient routine. At an interior sampled order,

$$\widehat{\alpha}(q_j) = \frac{\tau(q_{j+1}) - \tau(q_{j-1})}{q_{j+1} - q_{j-1}}, \qquad \alpha_{\min} = \min_{q \in \mathcal{Q}} \widehat{\alpha}(q), \quad \alpha_{\max} = \max_{q \in \mathcal{Q}} \widehat{\alpha}(q). \tag{12}$$

The continuous derivative satisfies $\alpha(q) = h(q) + qh'(q)$; on our grid, $\widehat{\alpha}(0) = [h(-0.5) + h(0.5)]/2$ need not equal $h(0)$. Our reported spectrum is the raw finite-difference transform. Its sampled extrema are not extrapolated $f = 0$ crossings. Neither α nor raw f is constrained to $[0, 1]$; $f(\widehat{\alpha}(0)) = 1$ is algebraic, and the all-order fit criterion does not enforce concavity or $f \leq 1$. The implementation also computes a sensitivity spectrum by projecting successive τ slopes onto a nonincreasing sequence with pooled-adjacent-violators regression and anchoring $\tau(0) = -1$. This projection does not replace the raw spectrum or repair its fit flags.

Figure 3B illustrates the horizontal span, peak coordinate, and branch lengths used below. From the nine sampled α values, we compute width $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$, peak location $\alpha_0 = \alpha(q = 0)$, and branch asymmetry

$$\ell_{\mathrm{L}} = \alpha_0 - \alpha_{\min}, \qquad \ell_{\mathrm{R}} = \alpha_{\max} - \alpha_0, \qquad A = \frac{\ell_{\mathrm{R}} - \ell_{\mathrm{L}}}{\Delta\alpha}. \tag{13}$$

The $q = 0$ location coincides with the sampled spectrum maximum in this dataset. The location $\alpha_0 = \widehat{\alpha}(0) \in [\alpha_{\min}, \alpha_{\max}]$ is a horizontal coordinate, not a mean, median, or peak height. Width is nonnegative with no universal upper bound; greater width means a larger spread of sampled strengths, not necessarily better reasoning. Both branch lengths are nonnegative and sum to $\Delta\alpha$. For positive width, $A \in [-1, 1]$: positive denotes a longer right branch, negative a longer left branch, and zero equal lengths. At $A = 1$ the left length is zero; at $A = -1$ the right length is zero. This ratio is not statistical skewness; the schematic's longer right branch is illustrative. For zero width the ratio is undefined; the implementation stores zero as a fallback, which is not needed for the retained runs. These quantities describe the sampled finite-data spectrum. Their interpretation as dimensions of singularity sets would require additional scaling assumptions.

A.6 Fit diagnostics and retention

For the scaling regression, define

$$R^2(q) = 1 - \frac{\sum_{s \in \mathcal{S}} e_q(s)^2}{\sum_{s \in \mathcal{S}} [\log F_q(s) - \overline{\log F_q}]^2}, \tag{14}$$

where the bar averages over scales. With an intercept and nonconstant response, $0 \leq R^2 \leq 1$; one indicates an exact line. The production sufficient-statistics fit stores one when the denominator is at most double-precision machine epsilon. This numerical convention is not evidence of scaling. The retention threshold is a chosen inclusion criterion, not a mathematical requirement for computing descriptors. We apply five diagnostics to every technically valid natural-order trajectory. The *single-scaling* check requires $R^2(2) \geq 0.95$ and $\Delta\mathrm{BIC} < 10$, where $\Delta\mathrm{BIC}$ is single-line BIC minus the lowest BIC of continuous two-slope fits at candidate interior scale knots. The implementation leaves at least four sampled scales strictly on each side and counts two parameters for the line and four for

the searched-knot model. This is a diagnostic, not a rule for selecting the fitted interval. The *all-order* check requires $\min_{q\in\mathcal{Q}} R^2(q) \geq 0.95$. The *raw-concavity* check allows no increase greater than 10^{-3} between successive interval slopes of $\tau(q)$. The *endpoint-stability* check requires that the largest absolute change in $h(q)$ be at most 0.10 when removing the smallest scale, the largest scale, or both.

Finally, the *negative-order-support* check uses $q = -2$ at $s = 48$. With $w_\nu = \tilde{v}_\nu^{-1} / \sum_\mu \tilde{v}_\mu^{-1}$, it requires $N_{\text{eff}} = (\sum_\nu w_\nu^2)^{-1} \geq 4$. This assesses weight concentration among the 42 forward/backward windows, not a count of independent observations. Under the final reporting policy, each diagnostic must pass in at least 80% of the pooled cohort. Main descriptive plots use the all-order mask, paired inference requires both members to pass, and sensitivity analyses require all five checks in both members.

A.7 Structured nulls and paired comparison

Each run is compared with 200 permutations from each of four families. Event shuffling reorders all event vectors; within-task shuffling reorders vectors inside each task while preserving task sequence; task-block permutation reorders intact task blocks; and role-stratified shuffling permutes vectors within sender-by-event-type strata across the run. All coordinates of an event move together. Each surrogate is centered, integrated, and analyzed again using the same scales, orders, and detrending. Deterministic seeds are derived from the master seed 20260903, pair identifier, and null family. Seeds are shared across conditions, although identical index mappings are guaranteed only where their preserved structures coincide.

For the designated comparison, each run's observed width is adjusted by subtracting its median task-block-permutation width. Excess width can be negative, zero, or positive relative to the control; it is not an absolute amount of coordination. Adjustment occurs within each run before forming $d_i = E_{i,\text{shared}} - E_{i,\text{local}}$. We report $\bar{d} = |\mathcal{I}|^{-1} \sum_{i\in\mathcal{I}} d_i$, where $\mathcal{I}$ contains pairs whose two runs pass. Confidence intervals resample whole pair differences with 20,000 bootstrap draws; two-sided tests use 100,000 sign flips, with a plus-one Monte Carlo correction. Functional tests apply one sign to each pair's complete $h(q)$ contrast vector, using the mean squared studentized contrast and the maximum absolute studentized contrast across orders. The latter supplies the within-function multiplicity adjustment. Secondary scalar comparisons use the Benjamini–Hochberg family specified in Section 5. These comparisons describe sensitivity to the intervention relative to explicitly preserved structures, rather than isolating a unique coordination mechanism.

A.8 Multiscale covariance-mode concentration

We analyze direct embedding coordinates, without the cumulative integration used for MF-DFA. At each of the same 19 scales $s \in \{8, 9, 10, 11, 12, 13, 15, 16, 18, 20, 22, 24, 26, 29, 32, 36, 39, 43, 48\}$, each run is centered and analyzed using every valid window start $t = 0, \ldots, T - s$. Each coordinate is regressed on an intercept and linear time trend within the window. For the resulting residual matrix $B_\nu \in \mathbb{R}^{s\times D}$, nonzero eigenvalues of $B_\nu^\top B_\nu / s$ are obtained from the smaller Gram matrix $B_\nu B_\nu^\top / s$. The two null modes imposed by detrending are removed; negative numerical roundoff is clipped to zero. We use $D = 768$ Nomic coordinates and no ridge in the principal analysis.

Let $p_{j,\nu} = \lambda_{j,\nu} / \sum_k \lambda_{k,\nu}$ denote empirical variance shares. For orders $r = 1, 2$, define

$$
\begin{aligned}
H_{1,\nu} &= -\sum_j w_{j,\nu} \log w_{j,\nu}, \qquad & H_{2,\nu} &= -\log \sum_j w_{j,\nu}^2, \\
N_{r,\nu} &= \exp(H_{r,\nu}), & K_{r,\nu}(s) &= 1 - \frac{H_{r,\nu}}{\log k_{\max}(s)}.
\end{aligned}
\tag{15}
$$

where $0 \log 0 = 0$ and $k_{\max}(s) = \min(D, s - 2) = s - 2$ is the maximum empirical rank after linear detrending. N_r is the equivalent number of equally weighted variance modes; K_r measures concentration relative to the maximum rank available at that scale. Thus $1 \leq N_r \leq k_{\max}$ and $0 \leq K_r \leq 1$. This normalization adjusts for the rank ceiling, not all finite-window estimation effects. Modes are embedding directions, not identified topics or agents. Entropy order r weights covariance-mode shares and has no direct conversion to MF-DFA moment order q.

Entropy, effective mode count, and concentration are computed separately in each window. For any such metric M, the balanced-start run summary is

$$\overline{M}_i(s) = \frac{1}{16}\sum_{u=0}^{15}\frac{1}{|\mathcal{V}_{i,u}(s)|}\sum_{t\in\mathcal{V}_{i,u}(s)} M_{i,t}(s), \tag{16}$$

where $\mathcal{V}_{i,u}(s)$ contains valid starts with $t \bmod 16 = u$. Thus each start phase receives equal weight, independently of scale. A window with residual covariance trace at most 10^{-12} has undefined variance shares; its start is excluded from both members of the pair before averaging. All 16 phases remain represented in every run at every scale. We average $\exp(H_r)$ rather than exponentiating average entropy, and each run receives equal weight. Figure 4 uses order two on 83 complete Nomic all-q-passing pairs. The multiscale and balanced-start extensions followed inspection of the earlier entropy results and are exploratory, not preregistered confirmations.

For pair i, let $d_i(s) = \overline{K}_{i,C}(s) - \overline{K}_{i,A}(s)$. We resample whole pairs 20,000 times with replacement (seed 20260921), using identical pair indices across scales, conditions, orders, and window schemes within each cohort. Figure 4 uses pointwise percentile intervals for mean effective mode counts. For the supplementary concentration contrasts, let $\widehat{\mu}(s)$ be the mean pair difference and $\widehat{\mathrm{SE}}(s)$ its sample standard error. Each bootstrap draw gives

$$M_b = \max_{s\in\mathcal{S}}\frac{|\widehat{\mu}^*_b(s) - \widehat{\mu}(s)|}{\widehat{\mathrm{SE}}(s)}. \tag{17}$$

The 95th percentile $c_{.95}$ gives the approximate simultaneous band $\widehat{\mu}(s) \pm c_{.95}\widehat{\mathrm{SE}}(s)$ over the 19 sampled scales. Standard errors in this calibration are those of the original sample, not recomputed bootstrap standard errors. Windows are never treated as independent replicates. Each sensitivity curve has a separate 19-scale band; there is no joint adjustment across sensitivity analyses or coverage claim between sampled scales.

Sensitivity analyses use all 100 pairs, order one, and equal weighting of all valid individual windows instead of equal phase weights. A separate numerical sensitivity adds a uniform 10^{-8} ridge to the $s-2$-dimensional residual support; it does not assign positive variance to the ambient null space. These checks use the same paired start exclusions and bootstrap procedure (Appendix C.9).

B Task bank and generation procedures

B.1 Distributed-evidence tasks and interaction budget

We compare two rules for access to task history in a synthetic interaction setting. Four agents, identified as Analyst, Checker, Planner, and Coordinator, use the same generation model but receive different private evidence packets. Each receives the common task question and its own packet; prompts request that the agent disclose all concrete facts in its packet on its first contribution. The tasks are designed to encourage combining evidence distributed across agents. Whether the generated answers actually rely on that combination is not independently assessed.

The fixed task bank contains four templates (Table 3). A run comprises 64 task instances, each allocated exactly 16 authored events, giving 1,024 events. The scheduler repeatedly shuffles a complete cycle of the four templates, so each template occurs 16 times per run. Template questions, evidence, and answer groups are reused without regenerating their factual content. Each four-task cycle contains all four templates once; adjacent cycles can place the same template on opposite sides of a cycle boundary. This is repeated interaction on a small fixed task bank, rather than evaluation on 64 distinct task problems.

B.2 History intervention and adaptive routing

In the shared-history condition, the active agent sees every previous event in the current task. In the local-history condition, it sees only events that it authored or that were addressed directly to it. The archive-label mapping is given in Section 4; these conditions do not classify observed coordination behavior. Both conditions retain the active agent's private evidence and ongoing interaction. The

Table 3: Fixed task templates. Each role holds a separate portion of the evidence. The final column gives the number of answer groups used by the phrase-based score, not a count of independently evaluated reasoning steps.

Task	Required synthesis	Groups
Harbor contamination	Identify the solvent source at berth 7 using sensor, tide, transfer, and drain evidence; specify containment.	3
Diary attribution	Attribute the diary to Mina Alvarez using assignment records and corroborating chronometer or distinctive-shorthand evidence.	3
Service outage	Connect a changed health-check path to gateway failures and recommend a safe configuration rollback.	3
Expedition route	Select Route North using its fuel cache and reject alternatives using bridge-load and storm constraints.	4

intervention also changes the prompt's description of available history to match the condition. A message addressed to `team` is not expanded into four direct deliveries by the local-history filter: it remains visible there only to its author. Thus, the ablation restricts shared-history access rather than eliminating all memory or communication.

Every task starts with the Coordinator. Subsequently, a valid directly addressed recipient normally becomes the next speaker. After a team-addressed event, the scheduler selects a least-active agent, breaking ties with the run's seeded random generator. Until all agents have contributed, eligible recipients are restricted to agents who have not yet spoken; an invalid recipient or premature team broadcast is deterministically redirected to an eligible agent. After this initial coverage, the model chooses among other agents or the team. The interaction order is consequently adaptive, not a fixed round-robin sequence. Conversation paths may diverge across conditions even with matched seeds.

All history is reset at the start of the next task. The final event in each 16-event block is requested as a concise synthesis and its metadata are forced to `result`, addressed to `team`; its speaker follows the same scheduler rather than being forced to the Coordinator. Intermediate results do not terminate a task early. Agents have no external tools or sources. Structured event-type fields are logging categories and do not execute tools. The resulting trajectory concatenates tasks within a run, but carries no explicit cross-task conversational memory.

B.3 Generation settings and matched replication

The completed corpus contains 100 matched pairs, or 200 runs and 204,800 authored events. Generation uses Google DeepMind's Gemma 4 E4B model, served locally, with temperature 0.4, top-$p = 0.9$, and a maximum of 180 generated tokens per request. Agents are prompted to produce 15–90 words of natural-language text inside a structured response containing recipient, event type, and text. These are generation settings; they do not imply equal realized message lengths across conditions. The saved generation provenance records the model tag and returned model identifiers, but does not pin an immutable generator digest.

For pair index $i \in \{0, \ldots, 99\}$, the run seed is $20260829 + i$ in both conditions. At zero-based event position t, the initial generation seed is $(20260829 + i) \times 1{,}000{,}003 + t$. The task schedule is drawn before the interaction using the same run seed, giving matched template order, evidence, event budget, model, and decoding settings. Matching does not force identical text, recipients, or random draws after condition-dependent routing and retries. The checkpointed runner processes the shared-history arm before the local-history arm within each pair; condition order is not counterbalanced. The matched pair is the inferential unit, and tasks and events within a run are repeated observations.

Responses are parsed and checked for nonempty text, the word-count bounds, agent-identifier leakage, and specified experimental-metadata terms. Invalid outputs can be retried up to eight times using corrective prompts and a seed offset of $104{,}729$ per attempt. From the third attempt onward, leaked agent identifiers may instead be replaced with a neutral teammate reference if the other required checks pass. Consequently, the corpus includes a format-control and limited text-repair procedure, not only unconstrained raw generations. Metadata, including roles and visibility lists, are stored separately; only the resulting authored text enters the embedding stage described in Section 3.

B.4 Outcome measurement

Each template specifies $G \in \{3, 4\}$ required answer groups, with one or more acceptable phrases in each group. For final text z, the task score is

$$S(z) = \frac{1}{G} \sum_{g=1}^{G} \mathbf{1}\{\text{at least one phrase in group } g \text{ occurs in } z\}. \tag{18}$$

Matching uses case-folded text with collapsed whitespace and literal substring occurrence. Only the last event in a task determines its reported final score, even if an earlier event scored higher. A task is complete when all answer groups match. Run outcomes are the mean final score across 64 tasks and the count of complete tasks. This automatic criterion measures answer-key coverage; it does not assess negation, factual support, coherence, or whether one agent actually incorporated another's evidence. No blinded human scoring audit is reported, and these outcomes are descriptive context rather than a validated ground truth for coordination quality.

B.5 Data integrity, cohort split, and analysis history

Generation is checkpointed at event and run boundaries, with completed records assembled into the canonical event corpus. Saved records retain task and event identifiers, sender and recipient, visible-history identifiers, decoding settings, and scores. Technical validation covers run lengths, task boundaries, paired schedules and seeds, text alignment, and embedding dimensions, finiteness, normalization, model identity, and file hashes. The saved audit reports no technical input failures or pair exclusions. Statistical fit exclusions are applied separately according to Section 3; technically valid runs are not automatically reliable scaling estimates. Exact embedding-tokenizer counts are not available in the saved event index.

The original split assigns 20 pairs to selection and the remaining 80 to a reserved cohort by ranking salted SHA-256 hashes of pair identifiers, keeping both conditions together. The earlier fixed-crossover $q = 2$ protocol and the subsequent restricted-scale spectrum protocol are distinct analysis stages. The final reporting policy uses five fit diagnostics, with the all-order fit mask for the main comparison and all five checks for sensitivity analyses. The reporting policy was revised after inspection of selection results; the numerical computation was unchanged. The 100-pair analysis combines the selection and additional cohorts, which remain separately identified in the sensitivity results. Forty selection-run results were reused after a numerical reproduction benchmark; the other 160 runs received new surrogate analyses.

C Supporting diagnostics and null comparisons

C.1 Numerical verification

For one saved natural-order trajectory, the production sufficient-statistics calculation was checked against explicit residual detrending over the complete scale and moment-order grid. Maximum absolute differences were 1.31×10^{-13} for F_q, 3.35×10^{-12} for $h(q)$, and 7.94×10^{-14} for R^2. Agreement with the saved arrays was also within recorded numerical tolerances. This is a single-run implementation check; it did not regenerate text, embeddings, or surrogate ensembles.

C.2 Full fit-check counts

The five pooled checks apply separately to all technically valid runs. Each diagnostic must pass in at least 80% of runs; passing this pooled criterion does not mean that every run passes every check. Main paired estimates require both members to pass the all-order fit rule.

C.3 Cohort and fit-mask sensitivity

Table 5 collects the saved paired estimates discussed in Section 6. Selection and additional cohorts partition the combined sample, whereas the all-five-check sample is a subset of it. These comparisons assess consistency across subsets of the same corpus.

Table 4: Fit diagnostics in all 200 technically valid runs. Each condition contains 100 runs. Component counts concern separate checks and can involve different runs.

Check	Shared history	Local history	Pooled
Single scaling	97/100	98/100	195/200
All-q fits	84/100	98/100	182/200
Raw concavity	100/100	98/100	198/200
Endpoint stability	87/100	100/100	187/200
Negative-q support	100/100	100/100	200/200

Table 5: Task-block-null-adjusted width across cohorts and fit masks. The first two rows partition the combined all-q sample; stricter QC is a subset, not an independent cohort.

Sample	Pairs	Mean difference	95% CI
Selection (20 pairs)	16	+0.20853	$[0.19147, 0.22621]$
Additional (80 pairs)	67	+0.20619	$[0.19688, 0.21487]$
Combined	83	+0.20664	$[0.19855, 0.21438]$
Combined, all five checks	81	+0.20627	$[0.19800, 0.21425]$

C.4 Resampling and multiplicity

Figure 2 uses descriptive condition-level intervals, computed separately from the paired inference below. Within each condition, 2,000 bootstrap samples resample complete run-level nine-order curves, using the independently retained 84 shared-history or 98 local-history runs. The 2.5th and 97.5th percentiles at each order form pointwise 95% intervals around the condition mean. A random generator seeded with 20260909 is used successively for $h(q)$ and $\tau(q)$ in each condition. These bands describe uncertainty in the condition means, rather than in their paired difference.

The scalar CIs use 20,000 bootstrap draws resampling whole pair differences. Two-sided sign-flip tests use 100,000 draws. All four combined-sample scalar tests reach the Monte Carlo floor $1/100{,}001 \approx 10^{-5}$; this is not a zero probability. The three raw-feature contrasts have saved Benjamini–Hochberg adjusted $p \approx 1.5 \times 10^{-5}$ across 18 secondary tests (three features, three cohorts, and two fit masks). The designated null-adjusted width contrast is reported separately. Paired standardized effects d_z use the standard deviation of pair differences; their magnitude describes this controlled sample rather than a general effect across tasks or models. Sign-flip inference requires an appropriate exchangeability or symmetry assumption for paired signs.

The functional analysis of the complete $h(q)$ curves gives an integrated studentized statistic of 2313.316 and a maximum absolute t statistic of 53.357, with both global sign-flip p-values approximately 10^{-5}. Pointwise contrasts remain significant under the max-$|t|$ adjustment across the nine orders. Each randomization flips the sign of a pair's entire nine-order contrast vector, preserving dependence across orders. The reported CIs are pointwise bootstrap intervals, not simultaneous bands.

C.5 Expanded structured-null comparisons

Table 6 compares the observed width with each run's own shuffled baseline using the same 83 retained pairs. Each null uses 200 realizations per run. Positive excess means that the original width exceeds the median shuffled width. These controls retain the embedding vectors while disturbing selected aspects of their ordering.

For the designated task-block comparison, excess width and its paired contrast are

$$E_{ia} = \Delta\alpha_{ia} - \mathrm{median}_{b=1,\ldots,200}\, \Delta\alpha_{ia}^{\mathrm{block},b}, \qquad D_i = E_{i,\mathrm{shared}} - E_{i,\mathrm{local}}. \tag{19}$$

Here, i indexes matched pairs and a is the condition; each task block contains 16 messages.

Event shuffling produces positive mean excesses in both conditions (0.45995 shared-history; 0.17686 local-history), indicating that original ordering contributes to the measured width in both arms. The local-history trajectories therefore also exhibit ordering-sensitive structure. Role-

Table 6: Mean spectrum-width excess over each run's median shuffled width, using the same 83 complete all-q-passing pairs and 200 realizations per null per run. Positive values mean that the original width exceeds the shuffled median. Alternative-null means are descriptive; they do not imply individual-run significance.

Shuffle	Shared history	Local history
All events	$+0.45995$	$+0.17686$
Within sender/event-type strata	$+0.45201$	$+0.17625$
Events within each task	$+0.25068$	$+0.03417$
Whole task blocks	$+0.24814$	$+0.04150$

stratified shuffling, which permutes embeddings within sender and event-type strata, gives almost identical excesses (0.45201 and 0.17625). Preserving these strata is insufficient to reproduce the observed widths.

Within-task shuffling preserves task membership and task sequence while reordering events inside each task. It reduces shared-history width substantially on average (excess 0.25068), whereas the local-history excess is much smaller (0.03417). This supports a stronger contribution from within-task ordering in the shared-history condition. Task-block permutation preserves each task's internal exchanges but reorders whole tasks; its excesses are likewise larger for shared-history runs (0.24814 versus 0.04150). Thus, the statistic is sensitive both to exchanges within tasks and to the arrangement of intact tasks and their boundaries. The alternative-null paired excess-width differences are 0.28309 for event shuffling, 0.21651 for within-task shuffling, and 0.27575 for role-stratified shuffling; these are descriptive summaries without additional CIs or hypothesis tests.

The controls probe different aspects of ordering and do not constitute universal non-multifractal reference processes. In particular, because task histories reset, sensitivity to block permutation can reflect task sequence, boundary transitions, or task composition without demonstrating persistent cross-task memory.

Taken together, the controls suggest that preserving task structure reproduces much more of the local-history width than of the shared-history width. However, these shuffles disturb overlapping properties, so their effects cannot be added or interpreted as a decomposition into independent mechanisms. They identify rearrangements that weaken the spectrum, without uniquely identifying the conversational mechanism or establishing multifractality.

C.6 Comparison across embedding encoders

The completed comparison includes Nomic, BGE-M3, MPNet, MiniLM, and Qwen3-Embedding-0.6B (Qwen3-0.6B in tables). Each encoder has 200 validated trajectories from the same 100 matched pairs. The comparison assesses representation sensitivity on the same corpus. Original Nomic results are preserved.

Representation and procedure. The four alternatives are BAAI/bge-m3 (1,024 coordinates, CLS pooling), sentence-transformers/all-mpnet-base-v2 (768, attention-masked mean pooling), sentence-transformers/all-MiniLM-L6-v2 (384, attention-masked mean pooling), and Qwen/Qwen3-Embedding-0.6B (1,024, last-token pooling). Pinned revisions are recorded in the accompanying model lock and embedding configurations. All four encode full target-event text without a prefix, using float32 evaluation and explicit L2 normalization; Nomic retains its original prefix and native pooling. No event requires truncation. Native coordinate dimensions, event order, task boundaries, scales, moment orders, linear detrending, spectrum construction, and the final five-check policy are preserved. All four temporal null families use 200 realizations per run with the same pair-derived index permutations across encoders. The covariance-mode entropy diagnostic is not repeated here.

All 1,000 result archives passed checksum, alignment, and numerical-output validation. Two complete Nomic pairs, one from each original cohort, were independently recomputed, including their surrogate ensembles. Observed features agreed with saved results at relative/absolute tolerances

Table 7: Encoder retention out of 100 runs per condition and 100 pairs. C: shared history; A: local history.

Encoder	C, R^2	A, R^2	R^2 pairs	All-five pairs
Nomic	84	98	83	81
BGE-M3	88	100	88	85
MPNet	90	100	90	87
MiniLM	89	100	89	84
Qwen3-0.6B	88	100	88	86

Table 8: Spectrum measures on 82 common pairs across five encoders. Means are condition-specific; differences are paired C minus A. Intervals are pointwise 95% bootstrap CIs.

Measure	Encoder	C mean	A mean	C − A	95% CI
$\Delta\alpha$	Nomic	0.4629	0.1810	+0.2818	$[0.2652, 0.2998]$
$\Delta\alpha$	BGE-M3	0.4000	0.1304	+0.2697	$[0.2534, 0.2872]$
$\Delta\alpha$	MPNet	0.4009	0.1678	+0.2332	$[0.2174, 0.2504]$
$\Delta\alpha$	MiniLM	0.4322	0.1553	+0.2769	$[0.2573, 0.2986]$
$\Delta\alpha$	Qwen3-0.6B	0.3909	0.1460	+0.2449	$[0.2310, 0.2595]$
α_0	Nomic	1.3970	1.3112	+0.0858	$[0.0826, 0.0891]$
α_0	BGE-M3	1.3302	1.2349	+0.0953	$[0.0922, 0.0986]$
α_0	MPNet	1.3590	1.2777	+0.0813	$[0.0780, 0.0845]$
α_0	MiniLM	1.3626	1.2759	+0.0867	$[0.0834, 0.0902]$
α_0	Qwen3-0.6B	1.3488	1.2573	+0.0915	$[0.0885, 0.0945]$
A	Nomic	0.2372	-0.2136	+0.4508	$[0.4245, 0.4781]$
A	BGE-M3	0.2781	-0.2134	+0.4915	$[0.4650, 0.5187]$
A	MPNet	0.2275	-0.1711	+0.3987	$[0.3722, 0.4263]$
A	MiniLM	0.2657	-0.2133	+0.4790	$[0.4505, 0.5081]$
A	Qwen3-0.6B	0.2262	-0.2078	+0.4340	$[0.4080, 0.4610]$
$\Delta\alpha_{\text{excess}}$	Nomic	0.2481	0.0418	+0.2063	$[0.1982, 0.2142]$
$\Delta\alpha_{\text{excess}}$	BGE-M3	0.1927	0.0223	+0.1704	$[0.1624, 0.1781]$
$\Delta\alpha_{\text{excess}}$	MPNet	0.2185	0.0428	+0.1757	$[0.1687, 0.1827]$
$\Delta\alpha_{\text{excess}}$	MiniLM	0.2196	0.0242	+0.1954	$[0.1868, 0.2037]$
$\Delta\alpha_{\text{excess}}$	Qwen3-0.6B	0.2135	0.0257	+0.1879	$[0.1810, 0.1946]$

$10^{-7}/10^{-8}$; archived arrays agreed at $10^{-6}/10^{-7}$, and quality flags matched. This extends the separate single-run explicit-detrending check without changing its scope.

Retention and matched sample. Table 7 distinguishes condition-level retention from complete pairs. The five-model intersection contains 82 pairs under the all-q $R^2 \geq 0.95$ mask and 78 under all five checks. The 82-pair Nomic estimates below differ slightly from the original 83-pair estimates because the sample changed, not because Nomic was redefined. The completed all-q intersection equals the earlier three-model sample; the stricter intersection loses one additional pair.

Spectrum descriptors. Table 8 compares condition means and paired effects on the same 82 pairs. All 20 shared-minus-local effects are positive and their pointwise 95% bootstrap intervals exclude zero. All four alternative encoders yield smaller task-block-adjusted width contrasts than Nomic. BGE-M3 has the largest peak-location and asymmetry contrasts, whereas Nomic has the largest raw-width contrast. These differences do not rank overall encoder quality. Figure 5 shows the corresponding raw spectra for (A) Nomic, (B) BGE-M3, (C) MPNet, (D) MiniLM, and (E) Qwen3-0.6B, in left-to-right order.

Encoder interactions and inference. An encoder interaction subtracts the Nomic paired condition effect from the alternative encoder's effect on the same pairs. Table 9 reports the completed, prespecified 16-test Holm family. Fourteen interactions are significant at 0.05 after correction. MiniLM's raw-width and peak-location interactions are not significant (both adjusted $p = 0.3492$); this does not establish equivalence. Directional consistency therefore coexists with encoder-dependent magnitudes. Confidence intervals use 20,000 paired bootstrap draws; two-sided tests use 100,000 sign flips. Identical ordered pair sets use the same resampling indices and signs across encoders. Corrected values are Monte Carlo estimates, not exact probabilities.

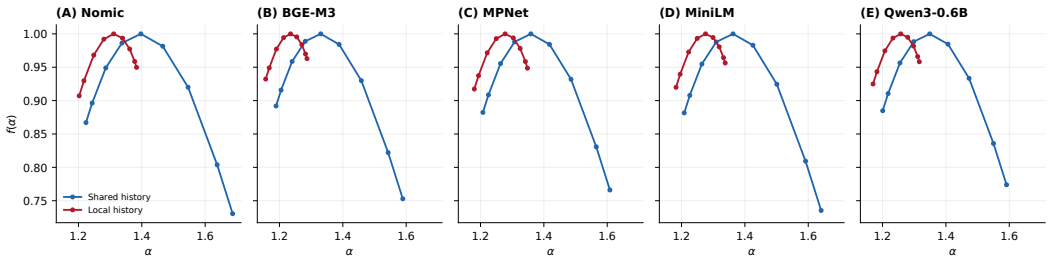


Figure 5: **Raw singularity spectra across encoders.** (A) Nomic. (B) BGE-M3. (C) MPNet. (D) MiniLM. (E) Qwen3-0.6B. All panels use the same 82 complete all-q-passing pairs. Curves join condition means of raw $\alpha(q)$ and $f(\alpha(q))$ at the nine sampled orders. Their plotted spans are not substitutes for means of per-run spectrum widths. The comparison does not extend to covariance-mode entropy.

Table 9: Encoder-minus-Nomic condition-effect differences on 82 common pairs. Holm correction covers the prespecified 16-test family.

Encoder	Measure	Interaction	95% CI	p_{Holm}
BGE-M3	$\Delta\alpha$	-0.0122	$[-0.0193, -0.0056]$	8.40×10^{-4}
BGE-M3	α_0	+0.0095	$[0.0085, 0.0106]$	1.60×10^{-4}
BGE-M3	A	+0.0407	$[0.0306, 0.0503]$	1.60×10^{-4}
BGE-M3	$\Delta\alpha_{\text{excess}}$	-0.0359	$[-0.0385, -0.0333]$	1.60×10^{-4}
MPNet	$\Delta\alpha$	-0.0487	$[-0.0595, -0.0387]$	1.60×10^{-4}
MPNet	α_0	-0.0045	$[-0.0059, -0.0031]$	1.60×10^{-4}
MPNet	A	-0.0521	$[-0.0659, -0.0389]$	1.60×10^{-4}
MPNet	$\Delta\alpha_{\text{excess}}$	-0.0306	$[-0.0348, -0.0262]$	1.60×10^{-4}
MiniLM	$\Delta\alpha$	-0.0049	$[-0.0144, 0.0044]$	0.3492
MiniLM	α_0	+0.0009	$[-0.0004, 0.0023]$	0.3492
MiniLM	A	+0.0282	$[0.0162, 0.0401]$	2.00×10^{-4}
MiniLM	$\Delta\alpha_{\text{excess}}$	-0.0109	$[-0.0157, -0.0066]$	1.60×10^{-4}
Qwen3-0.6B	$\Delta\alpha$	-0.0369	$[-0.0447, -0.0298]$	1.60×10^{-4}
Qwen3-0.6B	α_0	+0.0057	$[0.0045, 0.0068]$	1.60×10^{-4}
Qwen3-0.6B	A	-0.0168	$[-0.0277, -0.0061]$	0.0101
Qwen3-0.6B	$\Delta\alpha_{\text{excess}}$	-0.0184	$[-0.0216, -0.0153]$	1.60×10^{-4}

The completed report supersedes the interim eight-test correction. Its 208 secondary scalar tests share one Benjamini–Hochberg family, including within-model condition effects and within-arm encoder-minus-Nomic comparisons across labeled samples and masks. Sensitivity interaction tables use separate 16-test Holm families. No model was selected or procedure retuned from the interim look, and no adjustment for repeated looks is claimed.

Stricter-mask sensitivity and scope. The common all-five-check sample retains all 20 positive condition effects with pointwise intervals excluding zero. Task-block-adjusted width differences are 0.2056 for Nomic, 0.1696 for BGE-M3, 0.1750 for MPNet, 0.1941 for MiniLM, and 0.1870 for Qwen3-0.6B (78 pairs). Model-specific all-q samples retain 83, 88, 90, 89, and 88 pairs, respectively. Both comparisons condition on observed fit support. The evidence supports consistency of effect direction across these five representations in this corpus; broader encoder generalization, independent replication, and calibration against known finite-length processes remain open.

C.7 EXPLORATORY SPECTRUM–SCORE ASSOCIATIONS

Figure 3C–E uses Nomic descriptors for the 82 complete pairs passing the all-q fit filter in every encoder (164 runs). This intersection matches the encoder comparison; it differs from the 83-pair Nomic condition-effect sample. Mean task score averages answer-group coverage from the final event of each of 64 tasks. Descriptors summarize the full 1,024-event trajectory, so these are run-level associations rather than task-level analyses. Scores are joined by run, pair, and condition; repeated encoder measurements do not add independent observations.

Table 10 reports Spearman correlations within each condition and between matched changes. For the latter, both descriptor and score differences are shared history minus local history. Each analysis has one observation per pair. Pointwise percentile intervals use 2,000 bootstrap resamples of pairs, with ranks recomputed in each draw and seed 20260921. These exploratory intervals are not adjusted for multiple comparisons; no hypothesis tests or claims of independent confirmation are made. Pooled correlations in the figure are descriptive because runs within pairs are dependent.

Table 10: Spectrum–score associations in 82 pairs. C: shared history; A: local history. Intervals are pointwise 95% pair-bootstrap intervals.

Descriptor	Comparison	ρ	95% CI	
Width	C	0.051	-0.166	0.258
	A	0.123	-0.091	0.317
	Paired change	0.104	-0.122	0.309
Peak location	C	0.198	-0.013	0.393
	A	-0.006	-0.222	0.215
	Paired change	-0.001	-0.201	0.200
Asymmetry	C	-0.006	-0.218	0.210
	A	0.137	-0.088	0.341
	Paired change	0.157	-0.064	0.343

Pooled correlations of 0.772, 0.774, and 0.766 coexist with small within-condition and paired-change associations. Nomic task-block-adjusted width likewise has a paired-change correlation of 0.011. Across five encoders and these four descriptors, paired-change correlations range from -0.079 to 0.157 on the same 82 pairs. This descriptive sensitivity check reuses the same scores and trajectories and is not independent replication. The intervals do not establish absence of association. Incremental prediction beyond history-access condition requires held-out evaluation with pairs kept together; generalization to independently assessed semantic failures remains open.

C.8 CONVENTIONAL-SUMMARY BASELINE FOR TASK SCORE

We compared MV-DFA descriptors against six conventional run-level summaries: (i) word-count baselines: mean words per message (average across all 1,024 messages) and standard deviation (message-length variability around that mean); (ii) word-count serial structure: lag-1 autocorrelation (Pearson correlation between each message's word count and the next, including cross-task transitions); and (iii) task-level embedding summaries: standard deviation of the 64 task-mean word counts (average message length varies across the 64 tasks, each computed from 16 messages), total centered embedding variance (mean squared distance of message embeddings from the run's mean embedding), and adjacent task-centroid cosine (mean cosine similarity between neighboring task-mean embeddings across 63 transitions). Figure 6 plots each against the mean final task score across all 100 matched pairs, with separate descriptive least-squares fits within each condition. The between-condition score gap exceeds most within-condition slopes by a wide margin, so any pooled association would primarily reflect the history-access intervention rather than a summary-score relationship.

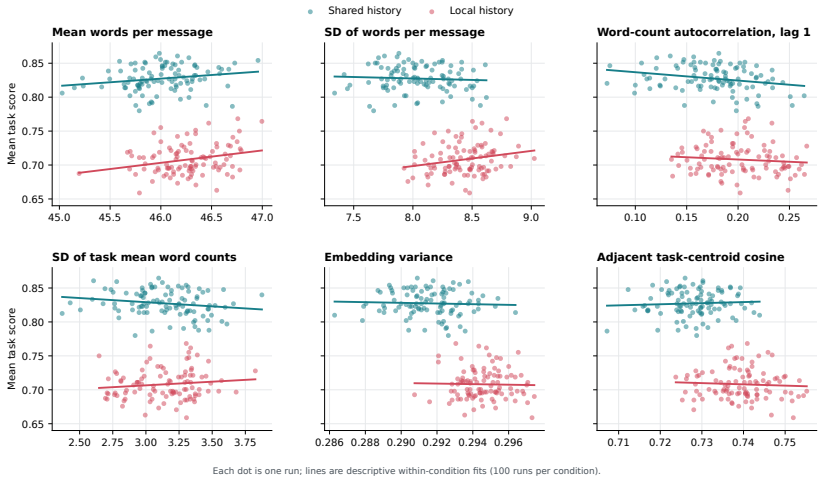


Figure 6: **Conventional summaries against task score.** Each point is one 1,024-message run; the two conditions each contribute 100 runs. Lines are descriptive least-squares fits within conditions. Score is the average final task score over 64 tasks. All six summaries and the score use the full run, so this is an association plot, not a prospective prediction test.

The individual panels show substantial overlap in these six summaries, but their joint separation should be tested rather than inferred from the plots. On the 83 pairs passing the all-q fit rule in both conditions, we trained a ridge classifier to distinguish history conditions, with feature standardization within each training fold, penalty 10, and a 0.5 decision threshold. Thirty random five-fold assignments kept matched runs together. Table 11 reports median held-out accuracy. The six-statistic model classified 159 of 166 runs in the median assignment (95.8%); both the four MV-DFA descriptors and raw width alone classified all 166. Thus MV-DFA separates the two conditions more cleanly than this particular six-statistic panel.

Table 11: Internal history-condition classification on 83 fit-passing pairs. Accuracy is the median over 30 pair-grouped five-fold assignments; the same runs recur in every assignment.

Predictors	Accuracy	AUC
Six conventional summaries	0.958	0.987
Four MV-DFA descriptors	1.000	1.000
Raw spectrum width alone	1.000	1.000

C.9 BALANCED-START CONCENTRATION AND SENSITIVITY

Figure 4 uses all valid window starts with equal weighting of the 16 phases within the repeated task schedule. This avoids giving scale-dependent emphasis to particular start phases. Balancing addresses window placement; it does not remove task transitions or imply uniform geometry across phases. The principal summary uses unregularized empirical covariance modes and order two on 83 complete all-q-passing Nomic pairs. Metrics are computed within windows, averaged within phases, and then averaged equally across phases and runs. The rank ceiling $s - 2$ changes with window length; adjusting for it does not remove all finite-window estimation effects. We do not fit a scaling law to these entropy curves.

Table 12 reports selected scales. Shared-history concentration is greater in every retained pair at each of the 19 scales, and the simultaneous band for the mean contrast excludes zero throughout. At $s = 16$, mean effective mode counts are 4.4211 and 5.0647; at $s = 48$, they are 4.4415 and 5.2668. Effective counts and entropies are averaged separately, so a mean effective count is not the exponential of a mean entropy.

Table 12: Balanced-start order-two concentration at selected scales. C: shared history; A: local history. Bands are simultaneous over all 19 scales, not only the five displayed.

s	C mean	A mean	C−A	95% simultaneous band
8	0.40423	0.34219	$+0.06204$	$[+0.05929, +0.06478]$
16	0.44515	0.39079	$+0.05436$	$[+0.05225, +0.05647]$
24	0.50583	0.44931	$+0.05652$	$[+0.05449, +0.05856]$
32	0.57769	0.52339	$+0.05430$	$[+0.05228, +0.05632]$
48	0.61601	0.57058	$+0.04543$	$[+0.04364, +0.04723]$

Sample, entropy order, and numerical sensitivity. The concentration contrast remains positive at all 19 scales when using all 100 pairs, order-one entropy, equal weighting of individual windows, or a uniform 10^{-8} ridge within the residual support (Figure 7). Across all combinations of these settings, the corresponding simultaneous bands remain above zero and each pair has a positive contrast at every sampled scale. Each curve has its own approximate 95% simultaneous band over the 19 sampled points, without joint adjustment across the collection of sensitivity analyses or coverage between sampled scales. Overlapping windows are not treated as independent replicates: inference resamples whole pairs.

Numerical checks and scope. The computation includes all 200 saved trajectories, 19 scales, two entropy orders, two start-weighting schemes, and two regularization choices. Fourteen near-zero-variance run/scale/start cases occur in shared-history trajectories at $s = 8$ or 9. Because their variance shares are undefined, the matching start is excluded from both conditions, removing 28 run-level windows in total. All phases remain represented, with equal valid-window counts within each pair. Residual Gram eigenvalues were checked against direct SVD at every scale on three windows from each member of one pair. Source hashes, window counts, phase summaries, run-level metrics, bootstrap settings, and numerical checks are archived with the analysis. The balanced-start analysis was developed after examining the earlier partition-based results. It supports a phase-balanced geometric contrast in this corpus; it does not independently validate entropy as a performance predictor or establish portability to other encoders.

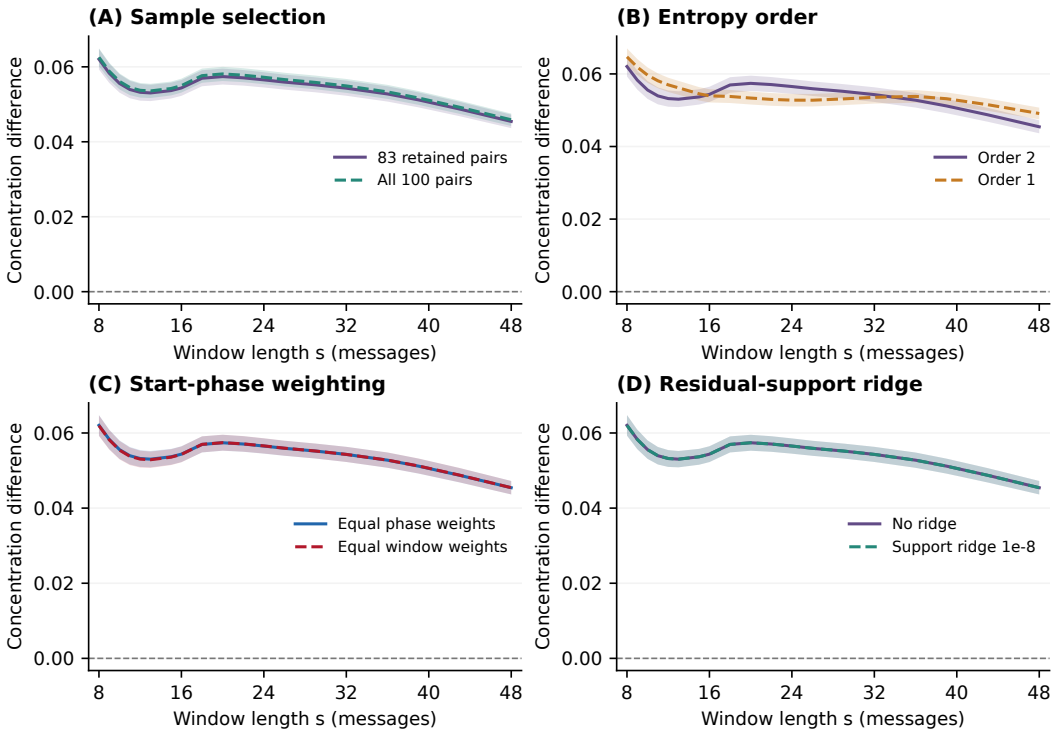


Figure 7: **Sensitivity of concentration with balanced window starts.** Shared-minus-local rank-adjusted concentration. (A) Retained 83 pairs versus all 100. (B) Orders two and one. (C) Equal start-phase weights versus equal individual-window weights. (D) No ridge versus a uniform 10^{-8} ridge within the residual support; curves nearly coincide. Unless varied, settings are order two, no ridge, 83 pairs, and equal phase weights. Each shaded region is a separate approximate 95% simultaneous band over 19 scales. Horizontal lines mark zero.

D EXPLORATORY EXTENSION TO A HIDDEN-CHANNEL SETTING

A hidden channel poses a subtler extension of the main history-access experiment: can the same trajectory analysis register a less overt memory-related perturbation? StegoBench distinguishes recoverable hidden communication from its detectability in ordinary-looking text (Meyer et al., 2026); partial observability can also complicate oversight (Lang et al., 2024). Our pilot examines a trajectory descriptor rather than StegoBench's training schemes or RLHF outcomes.

Simulation and embedding. The saved corpus contains ten matched stego-control pairs, with four speaker labels (agents A-D). Each run has 1,500 messages in fifty 30-message tasks; ten templates recur five times, in matching order within each pair. The designated hidden-channel agent pair changes every task. Stego runs mark 746-754 messages as channel-active; controls mark none. Carrier and payload symbols are archived, but these annotations are not decoding or detectability measurements. Public message text was embedded using Nomic Embed Text v1.5, explicit L2 normalization, and 768 direct coordinates. This distinct corpus is analyzed separately from the main experiment.

Scaling analysis. Trace-aggregated MV-MF-DFA used linear detrending at 20 scales (8-187 messages) and nine orders ($q = -2$ to 2). Four temporal nulls had 200 realizations per run; the structured-null sensitivity below uses permutations of intact task blocks. Only 15% of runs reached $R^2 \geq 0.95$ at $q = -2$, versus all runs at $q = -0.5$. The widest contiguous range of at least four orders passing in every run was $q \in [-0.5, 1]$ (20/20 runs); its *partial* spectra appear in Figure 8. A wider $q \in [-0.5, 2]$ sensitivity range passed in 16/20 runs (ten stego, six control). Selection used pooled fit quality, not condition effects, but remains exploratory.

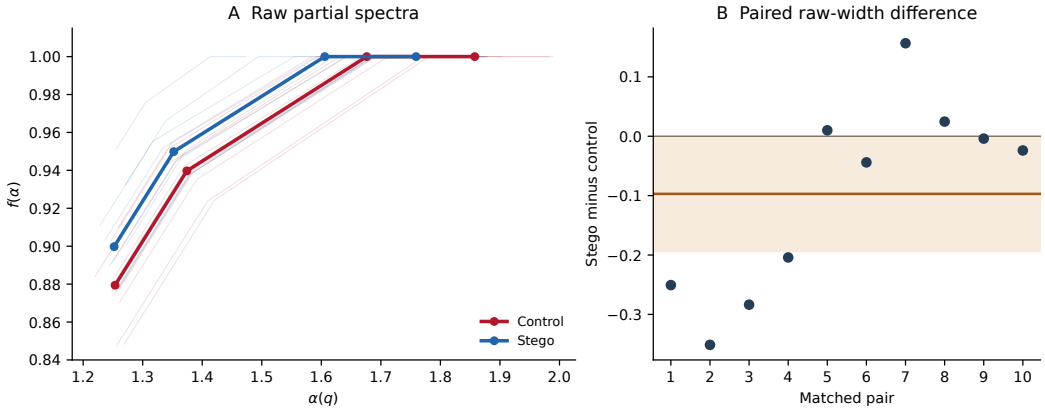


Figure 8: **Directional raw-spectrum pattern in a hidden-channel pilot.** (A) Raw partial spectra over four orders, $q = -0.5, 0, 0.5, 1$: faint lines are individual runs; bold lines connect arm means, whose span is not the mean run width. (B) Stego-minus-control raw-width differences in ten matched pairs. The brown line is the mean difference; shading gives its 95% pair-bootstrap interval. The exact paired sign-flip p-value is 0.082. The percentile interval excludes zero but is not an inversion of that test; the across-analysis Benjamini–Hochberg value is 0.237.

Stego spectra are narrower in seven of ten pairs. The mean raw width contrast is -0.0970 (95% pair-bootstrap interval $[-0.1941, -0.0008]$), with a similar six-order contrast of -0.1090. This directional agreement makes the case a promising hypothesis-generating extension, although exact paired p-values are 0.082 and 0.086. The small sample limits precision, and rotating the hidden-channel agent pair on every task may dilute a persistent pair-specific signature in the run-level spectrum. A larger study could compare fixed, rotating pairs, and no-hidden-channel cases to determine whether any channel-specific difference remains.