

Message residue preserves traces of hidden communication structure

Eric Gladstone*

Independent Researcher, Traverse City, Michigan, USA

*Correspondence and requests for materials should be addressed to Eric Gladstone:
eric.c.gladstone@gmail.com.

Abstract

Networked communication processes are often observed only through terminal artifacts: forwarded messages, published reports, synthesized briefs, or propagated claims. The production structure behind an artifact stays hidden, and structurally distinct processes can generate similar outputs. This study tests whether hidden production structure leaves recoverable residue in the artifact itself. In a controlled computational experiment, underlying events were held fixed while eight production motifs were instantiated as directed graphs of model-generated messages with known ground-truth structure. Two methods for recovering motif class were compared: a prompted reader given message content and trace information, and a supervised classifier limited to 14 surface-form features. On a task requiring classification among eight production motifs, where chance accuracy is 12.5%, the surface-form classifier recovered motif class at 61.2% accuracy, or 56.1% when each motif was weighted equally. Prompted readers recovered substantially less. Recoverability was highest when production structure left distinctive formal traces and lowest when structure was concealed or composed from multiple stages. Under controlled generating conditions, terminal communication artifacts can support measurement of latent production structure when hidden processes leave traces that are measurable by an appropriate method.

Introduction

Complex communication systems are often observed through their outputs rather than through the networked processes that produce them. What remains available may be a forwarded message, a published report, a synthesized brief, an intercepted communication, or a propagated claim. The dependence structure that produced the artifact stays hidden. The inferential target is therefore the latent production structure: not only what the artifact says, but what, if anything, the artifact preserves about who drew on whom, in what order, and through what transformations. The artifact can answer that question only if the production process leaves measurable residue in it.

The clearest instance of the problem is the relation between agreement and independence. Similar artifacts, or artifacts that agree, can be generated by structurally different latent processes. Agreement may reflect independent production, repeated dependence, concealed common ancestry, local reinforcement, or centralized synthesis. These structures differ in evidential and dynamical meaning while leaving terminal messages that can look similar.

Agreement is a property of the artifacts. Independence is a property of the hidden production graph. What matters for inference is therefore not agreement itself, but the dependence structure that produced it.

This is an inverse problem: inference from an observable output to the latent structure of the process that generated it, under partial observation. Given a terminal message, and at most partial information about how it arrived, how much of the production structure can be recovered? Posed this way, the question is a constrained case of network inference, with one observer, one message, and no direct access to the producing network. It is a recoverability question rather than a detection task. Related problems include locating the source of diffusion in a network¹⁻³, reconstructing diffusion and influence pathways^{4,5}, serial reproduction of messages through chains of people⁶, assessing evidence that may not be independent⁷⁻⁹, and measuring networks under incomplete or noisy observation³³⁻³⁶. The common premise is that network processes can leave recoverable residue in what they produce. The open question is what can be recovered, by which methods, and under what conditions.

This study uses controlled identification because the relevant ground truth is the production graph itself, which is usually unavailable in field corpora. The same 20 underlying events are held fixed while the production process varies across eight latent production motifs that recur across rumor spread, gatekeeping, coordinated information operations, news cascades, scientific synthesis, and small-group communication. The empirical target is motif-class recoverability, the class of the production graph¹⁰, not the identities of producers, the reconstruction of a complete network, or the truth of what the message asserts. The instantiated motifs include topology, role sequence, and transformation protocol together, so the study does not claim recovery of topology stripped of role or genre effects. Two methods for recovering motif class are compared on the same terminal messages: a prompted reader given message content and trace information, and a supervised classifier restricted to surface-form features. The contrast is diagnostic, not a capability ranking. It tests whether a method trained on formal residue extracts a production signal that a prompted content-and-trace reader does not.

The central finding is that recoverability depends on alignment between the residue a production process leaves and the method used to measure it. Across eight motifs, the surface-form classifier recovered motif class well above chance, prompted readers recovered less, and each method was strongest on motifs whose residue matched the information it used. The contribution is therefore threefold: a decomposition of message residue into semantic, formal, and trace classes, a controlled apparatus that makes motif class identifiable by holding the world fixed while the production structure varies, and evidence that recoverability is governed by residue-method alignment rather than by overall reader quality. The results provide a framework for latent network measurement under partial observability and an identification result under controlled generating conditions, not a claim about detection performance on naturally occurring communication.

Results

Framework: production motifs and message residue

Four objects define the inverse problem (Fig. 1). The represented world is the underlying state of affairs that a source could observe. The production graph is the network of agents and transformations through which information about that world passes. The terminal message is the artifact available to the observer. The inference is what the observer concludes about the hidden process from the message and any accompanying metadata.

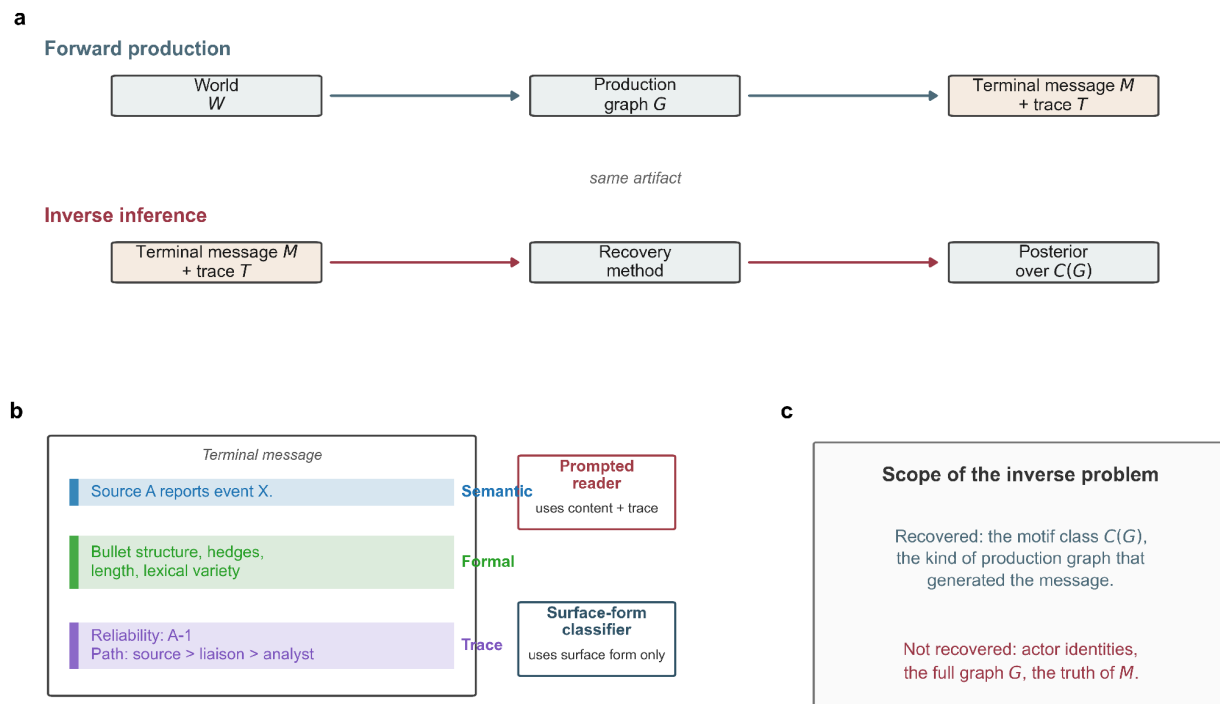


Figure 1. The inverse problem and message residue. **a**, In forward production, the represented world W generates a hidden production graph G , which produces a terminal message M and a partial trace T . In inverse inference, the terminal message and trace are used to estimate motif class $C(G)$. The orange boxes identify the same observed artifact in both directions. **b**, Schematic terminal message showing three residue classes. Semantic residue concerns what the message says, formal residue concerns how it is written, and trace residue concerns information that accompanies it. The prompted reader classifies from message content and rendered trace information. The surface-form classifier uses predefined formal features of the message. **c**, Scope of inference. The framework targets motif class $C(G)$, rather than actor identities, the complete graph, or the truth of the message.

This framework separates the present task from three neighboring problems. Truth assessment concerns the relation between the message and the world. Authorship attribution concerns the relation between the message and the identity of its producer. Source-credibility assessment concerns a property of the producer. The relation studied here is between the terminal message

and the production graph: what the message preserves about how it was produced. The target is the class of that graph, not the identity of any producer or the complete set of nodes and edges.

Two production graphs belong to the same class when they share the same dependence pattern, or motif¹⁰, regardless of which agents occupy the roles. Eight motifs are studied (Fig. 2), chosen because they recur across rumor spreading^{11,12}, organizational gatekeeping¹³⁻¹⁶, coordinated information operations^{17,18}, news cascades^{19,20}, scientific synthesis²¹, and small-group communication²²⁻²⁴. In single direct, one source produces the message without intermediaries. In chain relay, a message passes through a sequence of relayers, each transforming the previous output^{6,25}. In independent corroboration, several sources observe the same event without shared ancestry^{8,9}. In dependent repetition, several messages derive from one upstream source. In common-source laundering, one source is routed through apparent intermediaries so that it appears multiple. In clustered reinforcement, content circulates within a tightly connected community²⁶⁻²⁹. In centralized synthesis, several inputs are combined by a hub^{13,14,30}. In compound, two motifs are composed in sequence.

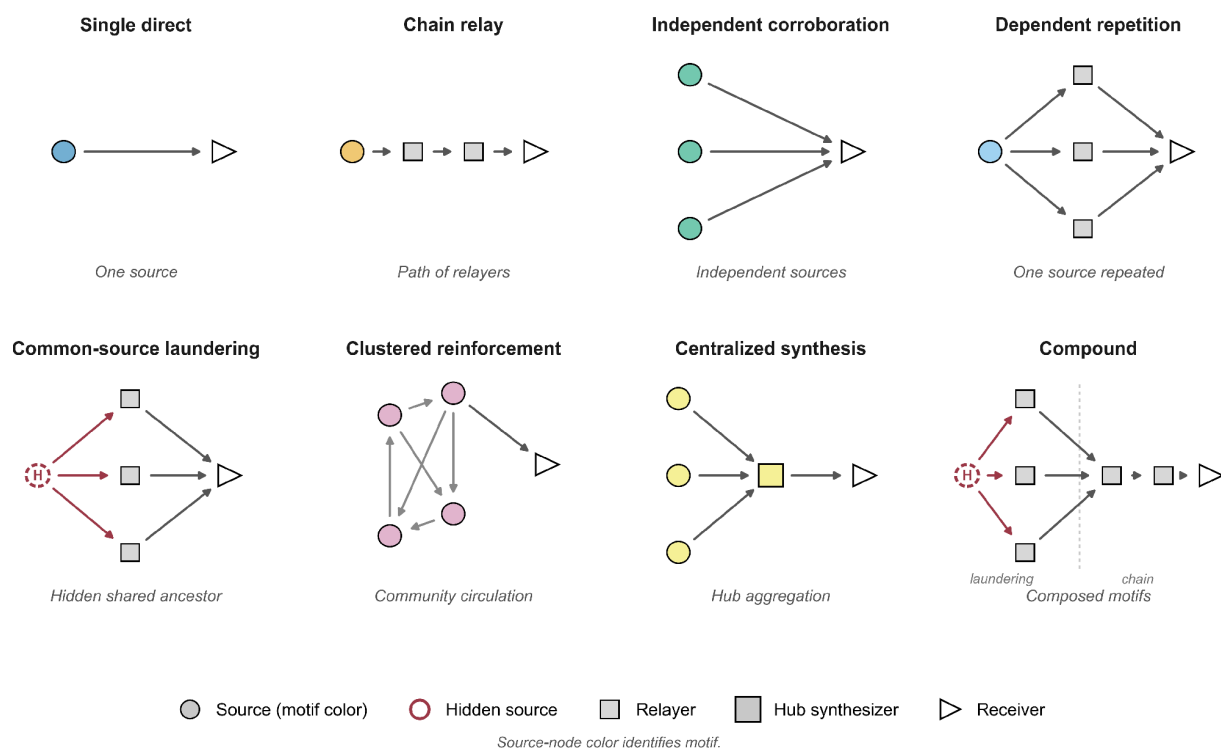


Figure 2. Eight production motifs represented as directed graphs. Each panel shows the call-graph structure of one production motif, together with its natural-language label, a brief descriptive cue, and a node-type legend. Source-node color identifies the motif. The same motif colors are used in Fig. 3a and as markers in Fig. 3b. Figure 4 instead uses colors to distinguish the two recovery methods. The compound motif is shown as common-source laundering followed by chain relay, with a faint divider marking the transition between stages.

Table 1

Table 1. Production motifs, mechanisms, expected traces, and primary confusable alternatives. Natural-language motif labels are used in the manuscript. Exact implementation keys are provided in Supplementary Table 1 and in the deposited code.

Motif	Graph property	Production mechanism	Expected trace	Primary confusable alternative
Single direct	Single source with no intermediaries	One model call conditioned on the world facts	Absence of relay drift or evidential scaffolding	One-hop chain relay
Chain relay	Directed path with N sequential transformations	N sequential model calls, with the output of call n passed to call $n+1$	Accumulating paraphrase drift and shifts in length, sentence structure, and lexical variety	Dependent repetition; compound
Independent corroboration	Multiple upstream sources without shared ancestry	Parallel model calls, each conditioned directly on the world facts	Evidential scaffolding and differentiation among sources	Common-source laundering; centralized synthesis
Dependent repetition	Multiple terminal expressions with one upstream ancestor	One upstream model call followed by multiple downstream calls conditioned on the upstream text	Repetition without evidential diversification	Chain relay; clustered reinforcement
Common-source laundering	Concealed shared ancestor routed through apparent intermediaries	One upstream model call followed by K relayer calls, with the trace concealing the ancestor	Apparent source multiplicity with low underlying diversity	Independent corroboration; centralized synthesis
Clustered reinforcement	Dense local circulation among similar nodes	A sequence of model calls within a shared community register	Recurrent surface regularities across circulated messages	Dependent repetition; compound
Centralized synthesis	Many-to-one aggregation through a hub	K parallel upstream calls followed by one synthesis call conditioned on all K outputs	Integrative comparison, source reconciliation, and evidential scaffolding	Independent corroboration
Compound	Sequential composition of two elementary motifs	A pipeline combining two motif-instantiating model-call subgraphs	Interference or masking among traces from the component motifs	The stage that dominates terminal form

These motifs separate evidential independence, which is a property of the production graph, from observational agreement, which is a property of the resulting messages. Agreement can therefore provide only indirect evidence about independence.

Message residue is any property of a terminal message that varies with the production graph that generated it. Three classes are distinguished. Semantic residue concerns what the message says, including claims, entities, and named sources. Formal residue concerns how the message is written, including length, sentence structure, lexical variety, punctuation, and discourse markers^{31,32}. Trace residue concerns information that accompanies the message, including source labels, quality ratings, and collection context.

A single message can carry all three forms of residue. A bulleted evidence block is formal residue, a claim attributed to a named source is semantic residue, and a reliability tag is trace residue. Whether formal residue is diagnostic of production structure is tested rather than assumed. Table 1 summarizes each motif's graph property, production mechanism, expected residue, and primary confusable alternative. A single observer reading one message sits near the hard end of network inference under partial observation, where the available information limits what any method can recover^{1,5}.

Terminal messages retain recoverable residue of production motifs

Each of 20 fixed worlds was passed through every production motif to generate terminal messages, and two methods were applied to the same messages (Methods). The prompted reader was a large language model given the message content and rendered trace information, but no task-specific supervised training. The surface-form classifier was a multinomial logistic regression trained on 14 features of the terminal message alone, with no access to source texts, parent messages, or trace metadata.

The two methods therefore differed in both information access and training. The prompted reader had richer information on each trial but no supervised exposure to the classification task. The surface-form classifier had less information on each trial but was trained on labeled examples. The comparison is not a general test of model capability. It asks whether supervised measurement of formal traces recovers production structure that prompted reading of content and trace information does not.

The task required classification among eight production motifs, so chance accuracy was 12.5%. On held-out test messages, the surface-form classifier recovered motif class at 61.2% accuracy (95% confidence interval 57.9 to 64.4%, $n = 840$), or 56.1% when each motif was weighted equally. On the same trials, the prompted reader achieved 27.3% accuracy (24.4 to 30.4%). Both methods exceeded chance, and neither approached perfect recovery.

Because both methods were evaluated on the same trials, their difference could be estimated directly. The paired accuracy gap was 33.9 percentage points (95% bootstrap CI 29.8 to 38.2 percentage points; Supplementary Information). A similar pattern appeared when the prompted

reader was replaced by a model from a different family, with accuracies of 57.6% for the surface-form classifier and 23.6% for the prompted reader on a stratified high-information subset. The advantage of the surface-form classifier therefore cannot be explained by access to information unavailable to the prompted reader. Instead, it indicates that the two methods differ in how effectively they use the traces left by the production process.

Recoverability varies by motif and follows residue-method alignment

Aggregate accuracy conceals substantial variation across motifs (Fig. 3). The surface-form classifier recovered independent corroboration most accurately (96.4%, 95% CI 90.0 to 98.8%), followed by chain relay (81.3%, 76.1 to 85.7%), clustered reinforcement (73.8%, 63.5 to 82.0%), and centralized synthesis (71.4%, 61.0 to 80.0%). Recovery was lower for dependent repetition (59.5%, 48.8 to 69.4%), single direct (39.3%, 29.5 to 50.0%), and common-source laundering (27.4%, 19.0 to 37.7%), and absent for the compound motif (0%). The prompted reader showed a different pattern. It recovered single direct at 95.2%, chain relay at 43.7%, and independent corroboration at 39.3%, but was at or near zero on coordination motifs defined by dependence among several sources.

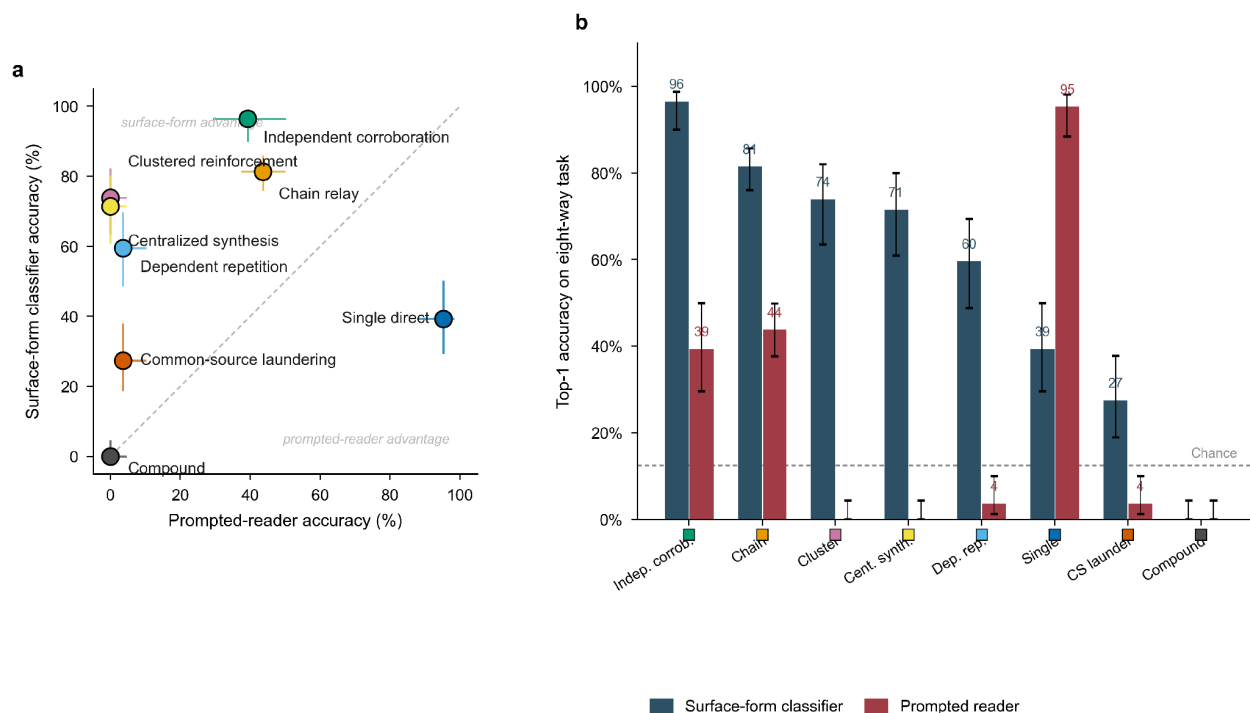


Figure 3. Motif recoverability and residue-method alignment. **a**, Accuracy by motif on the primary-run test trials. The x-axis shows prompted-reader accuracy and the y-axis shows surface-form classifier accuracy. Each point represents one motif, with horizontal and vertical Wilson 95% confidence intervals. The dashed diagonal marks equal accuracy. Points above the diagonal were recovered more accurately by the surface-form classifier. Single direct, the only point below the diagonal, was recovered more accurately by the prompted reader. Compound lies

at the origin. **b**, Accuracy by motif for the two methods, sorted by surface-form classifier accuracy. Bars show Wilson 95% confidence intervals and rounded percentages. Colored markers below the x-axis link each motif to its color in panel a and Fig. 2. Method colors distinguish the surface-form classifier from the prompted reader. Axis abbreviations: Indep. corrob. = independent corroboration; Chain = chain relay; Cluster = clustered reinforcement; Cent. synth. = centralized synthesis; Dep. rep. = dependent repetition; Single = single direct; CS launder = common-source laundering.

These differences are consistent with residue-method alignment. Within the instantiated production protocol, each motif leaves a different trace in the terminal message. In chain relay, successive paraphrasing alters message length, sentence structure, and lexical variety. As path length increased from one to five hops, the surface-form classifier remained near 80 to 82%, while the prompted reader rose from 32% to 48% to 51%. This suggests that repeated transformation strengthens a formal path signature and may also make that signature more legible in the message content.

Independent corroboration leaves evidential scaffolding, including headers, separated evidence blocks, and alternative-hypothesis sections. The surface-form classifier recovered this motif at 96%, while the prompted reader recovered it at 39%. Centralized synthesis leaves an integrative structure and was recovered by the surface-form classifier at 71%. Clustered reinforcement leaves recurrent surface regularities across community-circulated messages and was recovered at 74%.

Other motifs leave weaker or less accessible traces. Common-source laundering conceals a shared ancestor behind apparent intermediaries, which is the feature the process is designed to hide. Recovery was therefore low for both methods, at 27% for the surface-form classifier and 4% for the prompted reader. Single direct has little characteristic form beyond the absence of relay drift or evidential scaffolding. The surface-form classifier recovered it at 39%, whereas the prompted reader recovered the plain single-source message at 95%.

The compound motif provides the clearest failure. Neither method recovered it. The surface-form classifier labeled compound messages as chain relay in 77% of test cases, or 65 of 84, and as dependent repetition in 11%. This pattern is consistent with the final relay stage masking the upstream laundering structure. Two explanations cannot be separated by the present design. Compound messages may resemble chain relay more than they resemble a distinct compound class, and only one composition was tested. In either case, the traces of a composed process do not appear to combine additively into a distinct signature. Composition is therefore a central limit on recoverability rather than a peripheral failure case.

Message length also moderated recovery, but its effect depended on the motif. For centralized synthesis, surface-form recovery rose from 47% for messages of 200 tokens or fewer to 100% for messages between 200 and 300 tokens, while prompted-reader recovery remained near 2 to 3%. For independent corroboration, both methods improved with length, from 95% to 100% for

the surface-form classifier and from 27% to 37% for the prompted reader. For dependent repetition, longer messages reduced recovery for both methods. Surface-form recovery fell from 68% to 0%, and prompted-reader recovery fell from 9% to 0%, as longer rewrites diluted the production signature with additional content. Message length is therefore one source of formal residue, but its diagnostic value is motif-specific.

Alignment diagnostics

Three diagnostics using the same test data assess the central interpretation (Fig. 4). They ask whether the surface-form classifier's advantage can be explained by message length, dependence on one prompted model, or access to richer trace metadata. Where paired comparison is possible, each diagnostic uses the same 840 test trials.

First, the advantage is not reducible to message length, although length is itself a source of formal residue. A classifier restricted to token count and sentence count reached 50.1% accuracy, or 41.4% when each motif was weighted equally, on the same test split. Production motifs therefore alter basic message scale. The original pooled comparison used a length-only accuracy of 50.1%, yielding a descriptive difference of 11.1 percentage points. For the paired bootstrap, the length-only classifier was refitted on the matched trial set, where it achieved 52.1% accuracy and produced a mean paired difference of 9.0 percentage points, with a 95% interval of 6.1 to 11.9 percentage points (Supplementary Information). The gain from the full classifier shows that recoverability extends beyond message scale. Additional surface features improve recovery for motifs where length alone is weakly diagnostic, including single direct, dependent repetition, and clustered reinforcement.

Second, the difference is not specific to the primary prompted reader. On the 23-cell cross-family slice (Methods), the surface-form classifier achieved 57.6% accuracy (95% CI 47.4 to 67.2%), while the prompted reader from a different model family achieved 23.6% (95% CI 16.0 to 33.4%, $n = 89$ valid trials). The surface-form classifier remained near 57 to 61% regardless of which language model served as the prompted comparator. Both prompted readers concentrated their predictions on a limited subset of the eight motif labels, although the preferred labels differed by model (Supplementary Fig. 2). On the primary test trials, predicted-label entropy was 1.68 bits for the prompted reader and 2.36 bits for the surface-form classifier, against a maximum of $\log_2 8 = 3.00$ bits. On the cross-family slice, the corresponding values were 2.23 and 2.12 bits. These values indicate that the prompted readers often relied on a restricted subset of the available motif vocabulary rather than using all labels evenly.

Third, richer trace metadata did not close the gap. The surface-form classifier remained near its ceiling across trace-observability levels, with accuracies of 60.0%, 60.4%, 61.7%, and 62.1% from L1 to L6. The prompted reader improved as more trace information became available, rising from 14.2% at L1 to 51.2% at L6, but remained below the surface-form classifier even at the richest level of observability. Most of this gain came from the collection narrative, which indirectly encoded path length and intermediary types. The compressed source-quality rating

contributed little, and adversarial corruption of the trace went undetected. The prompted reader's stated confidence in message fidelity was also uncorrelated with actual source-content preservation (Pearson $r = +0.029$ across all $n = 4,197$ post-validation trials; Supplementary Fig. 3), so its confidence did not provide a useful signal of production structure.

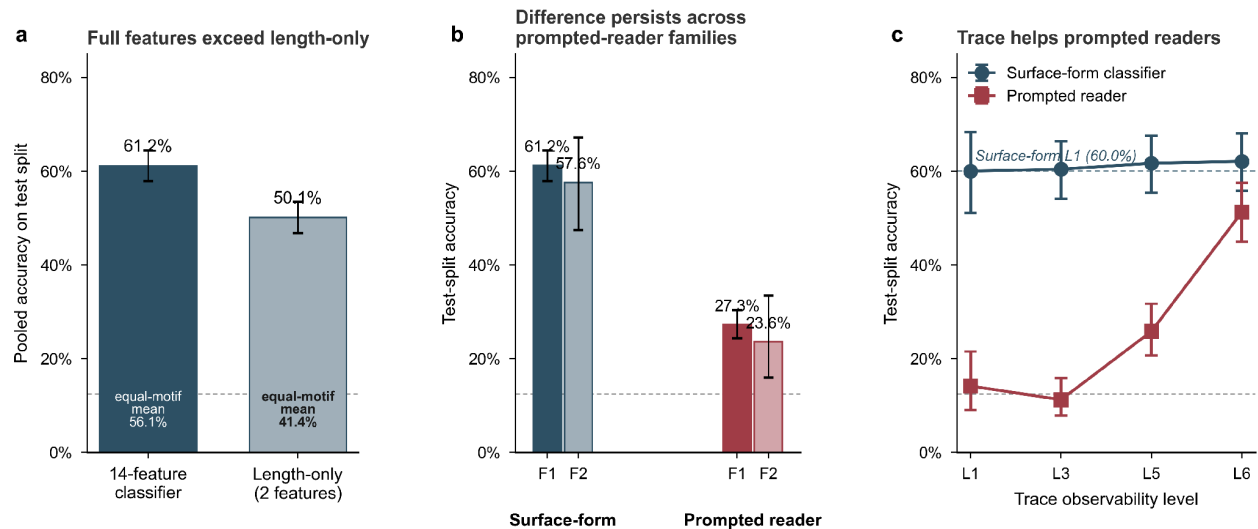


Figure 4. Diagnostics of the surface-form classifier's advantage. **a**, Full 14-feature classifier compared with a length-only classifier using token count and sentence count on the same primary-run test split ($n = 840$). Bars show pooled accuracy with Wilson 95% confidence intervals. The equal-motif mean gives each of the eight motifs equal weight. The pooled descriptive difference was 11.1 percentage points. A paired bootstrap over matched trial-level predictions gave a mean difference of 9.0 percentage points, with a 95% interval of 6.1 to 11.9 percentage points (Supplementary Information). **b**, Cross-family comparison. Both methods are evaluated on the F1 primary test split and the F2 cross-family slice. Surface-form classifier accuracy remained near 57 to 61%, whereas prompted-reader accuracy remained near 23 to 27%. Error bars are Wilson 95% confidence intervals. **c**, Trace-observability sweep across L1, L3, L5, and L6. Markers show accuracy with Wilson 95% confidence intervals. The dashed line marks the surface-form classifier's L1 accuracy of 60.0%, which exceeded prompted-reader accuracy at every trace level, including L6. The chance baseline is 12.5% in all panels.

Together, these diagnostics narrow the interpretation of the main result. The difference between the methods is not explained by message length alone, the choice of prompted model, richer trace metadata, or reader confidence. Across the four coordination motifs, the surface-form classifier exceeded the prompted reader by 30 to 74 percentage points (Fig. 3), and the pattern persisted under paired comparison and substitution of the prompted model. The evidence therefore supports an explanation based on the relation between the traces left by a production process and the information each method can use.

Discussion

The results support a measurement-alignment account. Production processes leave different traces in terminal messages, and recovery depends on whether a method uses the information in which those traces appear. The surface-form classifier recovered more motif structure than the prompted readers, despite having less information on each trial, because much of the recoverable signal was formal. In this experiment, distinctions among production structures survived more clearly in how messages were written than in what they asserted. This extends prior work locating style and authorship in formal properties of text^{31,32} from producer identity to production structure.

The findings also establish motif-class recoverability as an identification problem under partial observation. A single observer reading one terminal message has access to only a small fraction of the generating process, which limits what any method can recover¹⁻⁵. The distinction among semantic, formal, and trace residue clarifies which information channels remain available under that constraint. It also explains why the relevant comparison is not overall reader quality, but the fit between the traces left by a process and the information a method can use. Under controlled instantiation, latent production structure can therefore be measured without reconstructing every actor or edge, provided that the process leaves diagnostic residue in the observed artifact.

The clearest substantive implication concerns agreement. Agreement is a property of observable artifacts, whereas independence is a property of the hidden production structure that generated them. Similar agreement can arise from independent production, repeated dependence, concealed common ancestry, clustered reinforcement, or centralized synthesis. Agreement alone is therefore a weak guide to evidential independence. The present results show that part of this hidden dependence structure can be recovered from terminal messages under controlled generating conditions, particularly when the structure leaves formal traces.

Two boundary cases define the limits of that claim. First, composed structure remains difficult. Neither method recovered the compound motif, and most compound messages were classified as chain relay, consistent with the final relay stage masking the upstream laundering structure. This suggests that traces from a multi-stage process do not necessarily combine into a distinct and recoverable signature. Because many communication artifacts pass through several production stages, composition is a central challenge for structural recovery.

Second, richer metadata does not guarantee better recovery. Additional trace information improved the prompted reader, but even the richest trace condition did not outperform the surface-form classifier operating on the message alone. Compressed source-quality labels provided little information about dependence structure, adversarial corruption of the trace went undetected, and the prompted reader's confidence in message fidelity did not track actual source-content preservation. Metadata are therefore useful only when they encode properties relevant to the hidden structure and can be interpreted reliably.

The practical implication is methodological rather than diagnostic. Agreement among messages should not be treated as independent confirmation without some account of the dependence structure that produced it. Common-source laundering provides the clearest example because several agreeing messages can derive from one concealed ancestor. Under controlled or externally anchored conditions, formal traces may support corpus-level measurement of production dependence. They are not sufficient on their own for judging the structure behind any single message. Formal information is best treated as one measurement channel alongside content and trace information.

The experimental design is deliberately bounded. Messages were generated under role-constrained protocols while the underlying world was held fixed and the production structure varied. This control is what makes motif class identifiable. The reported accuracies therefore describe what the tested methods recover under these generating conditions. The corpus was designed to identify structural differences, not to estimate how often the motifs occur in naturally occurring communication. The estimand is the recoverability of experimentally instantiated production motifs under particular roles, transformations, and genre conventions, not topology separated from every feature of the production process.

Field transfer remains a separate empirical problem. A field study would require externally anchored production roles or regimes so that labels are not inferred from the same message features used for prediction. It would also need to separate labeling information from form-based predictors and test whether formal traces provide information beyond simpler aggregate or content-based descriptions. The present study does not establish performance in naturally occurring communication. It establishes a controlled identification result: terminal artifacts can preserve measurable traces of hidden production structure when those traces are available to the method used to recover them.

Methods

Worlds

The worlds are synthetic structured event records rather than empirical events. Each of the 20 worlds specifies a fixed event, competing causal hypotheses with supporting evidence, peripheral details, and a pair of competing propositions about the same state of affairs. Each proposition carries a truth value, uncertainty level, and evidence type. The same 20 worlds were held constant across all 4,200 primary trials and appeared in all 70 experimental cells. Variation in recovery across cells therefore reflects differences in production structure rather than differences in event content.

Production motifs

Each world was passed through each of the eight production motifs. Chain relay was instantiated at path lengths of 1, 3, and 5 hops to support within-motif comparison. Each motif was implemented as a sequence of separate language-model calls whose call graph mirrors the

corresponding production-graph structure. A motif is therefore defined by a topology of model calls under role-constrained transformation rules, not by a single prompt style.

A producer model, llama3.1:8b accessed through local inference, generated every message under structured prompts that fixed the role of each node. In chain relay, the source call generated a message from the world facts, and each subsequent relayer received the preceding output as input. The terminal message was the output of the final hop. In independent corroboration, three calls were conditioned directly and separately on the world facts, with no message passing between them, and one output was randomly selected as the terminal message. In common-source laundering, one upstream call generated from the world facts, three relayer calls each received that upstream text, the trace replaced the upstream identifier with a hidden-upstream sentinel, and one relayer output served as the terminal message. In centralized synthesis, three independent upstream calls were combined by a synthesis call conditioned on all three outputs. Dependent repetition, clustered reinforcement, and compound followed analogous role-constrained call graphs.

Function-level details are provided in the deposited code (src/regime_instantiators.py, src/run_experiment.py). Differences in terminal-message residue across motifs therefore arise from differences in message passing, role sequence, and transformation protocol rather than from differences in isolated prompt templates.

Reader methods

Two methods were used to recover motif class from the terminal messages. The prompted reader was a large language model, with qwen2.5:7b³⁸ as the primary instance and llama3.1:8b³⁹ as the cross-family instance. It received the message, the trace packet rendered as natural language, the eight motif labels with one-sentence definitions, and a candidate-source set built deterministically from the world and experimental cell. It returned a JSON object containing a posterior over the eight motif labels, an origin posterior over candidate sources, an independence judgment for corroboration-family motifs, and a fidelity posterior. The task was closed-set classification over the eight provided labels. Decoding was greedy at temperature 0.0 for reproducibility. After validation that tolerated benign output-format deviations, the rejection rate was 0.07%, or 3 of 4,200 primary trials. In the primary run, the producer model, llama3.1:8b, and the prompted reader, qwen2.5:7b, came from different model families, reducing the possibility that the reader was detecting artifacts specific to its own family⁴⁰.

The surface-form classifier was a multinomial logistic regression trained on 14 features of the terminal message. Six features described length and lexical density: token count, sentence count, type-token ratio, mean sentence length, mean word length, and punctuation density. Four counted entity types: numeric, named-entity, temporal, and location markers. Four counted discourse markers: hedging, uncertainty, evidential, and source markers. Isotonic regression provided post-hoc calibration⁴¹. Training used a world-level 80/20 split, with worlds W001 to W016 assigned to training and W017 to W020 assigned to testing. The feature set was fixed in advance

and included only properties of the terminal message, with no access to source propositions, parent messages, or trace metadata.

The two methods differed in both information access and training. The surface-form classifier had less information on each trial but supervised exposure to labeled training examples. The prompted reader had richer access to message content and trace information but no task-specific supervised training. The contrast was therefore diagnostic rather than competitive. It tested whether supervised measurement of formal traces recovered production structure that prompted reading of content and trace information did not.

Trace observability and metadata validity

Each terminal message was paired with metadata at one of four observability levels mapped to operational intelligence-analysis standards. L1 included the message text alone. L3 added a pseudonymous source identifier and a two-dimensional Admiralty-style rating of source reliability and information credibility⁴². L5 added a structured narrative describing how the message was collected. L6 added a structured analytic block containing key judgments, supporting evidence, alternative hypotheses, and confidence⁴³.

At L3 and above, each metadata field carried a validity coefficient representing the probability that the field was accurate rather than adversarially altered. This coefficient was set to $v = 0.95$ for rare corruption and $v = 0.30$ for heavy corruption. The validity coefficient therefore represented measurement error in the trace channel.

Outcomes and statistical analysis

The primary outcome was correct identification of the production motif on the eight-class task. Headline proportions are reported with Wilson 95% confidence intervals, and raw counts accompany percentages. Metrics are reported by experimental cell when the cell variable is the substantive comparison. Accuracy with each motif weighted equally is the unweighted mean of the eight per-motif accuracies. Chance performance on the eight-class task is 12.5%.

Paired bootstrap 95% confidence intervals for accuracy differences, including surface-form minus prompted and 14-feature minus length-only, were computed from 5,000 paired resamples of trial-level correctness on the same test split. Predicted-label entropy is the Shannon entropy of the marginal distribution over predicted motif labels. No null-hypothesis significance tests or one-tailed or two-tailed P values are reported. Uncertainty is expressed with two-sided Wilson 95% confidence intervals and paired-bootstrap 95% confidence intervals.

Robustness analyses

Several analyses bounded the central interpretation and are reported in full in the Supplementary Information. A cross-family analysis replaced the primary prompted reader with llama3.1:8b on 23 high-information cells. A length-only baseline refit the surface-form classifier using only

token count and sentence count. Its predictions were paired with the deposited 14-feature predictions to estimate the full-versus-length-only bootstrap interval.

Trace observability was varied across L1, L3, L5, and L6, and metadata validity across $v = 0.95$ and $v = 0.30$. Within-motif analyses examined chain-relay hop count and message length. The relation between the prompted reader's fidelity posterior and ground-truth source-content preservation was computed across all 4,197 post-validation primary trials.

Use of large language models

Language models were used as the experimental producer and as the prompted reader methods under study. These uses are part of the reported experimental design. Claude Code (Anthropic) and ChatGPT (OpenAI) were also used for code review, analysis support, and copy-editing assistance during manuscript preparation. The author reviewed all outputs and takes full responsibility for the content. Language models are not authors and do not meet authorship criteria.

Data availability

The data underlying this study are deposited in Zenodo at <https://doi.org/10.5281/zenodo.20528041> under the Creative Commons Attribution 4.0 International license. The deposit contains the 20 fixed event records (worlds_v0_2.jsonl), producer-model prompts for each of the eight production motifs, the four trace-packet rendering rules, and four JSON Lines streams per experimental run: lineage records, trace packets, prompted-reader outputs with the validator's invalid flag preserved per trial, and ground-truth labels. The streams cover all 4,200 primary-run trials and the matching cross-reader-family slice of 460 trials. The deposit also includes the 14-feature surface-form extractions for each terminal message, including the proposition-matched preservation rate and the surface-form classifier's trial-level test predictions.

Code availability

The code required to reproduce the analyses is deposited in Zenodo at <https://doi.org/10.5281/zenodo.20528041> under the MIT license. The deposit contains the experimental pipeline, producer dispatcher, motif instantiators, trace-packet assembly, residue extractor, proposition matcher, reader dispatcher, trained surface-form classifier, analysis scripts, configuration files, and scripts used to reproduce the four main figures, the three supplementary figures, and the paired-bootstrap and predicted-label-entropy values reported in the Supplementary Information.

Because language-model decoding is stochastic, the reproducibility standard for the upstream generation step is distributional rather than exact byte equivalence. Repeated runs should reproduce the same statistical patterns rather than identical outputs. Analyses conducted on the deposited data are deterministic.

Funding

The author received no specific funding for this work.

Author contributions

The author was responsible for Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing, Visualization, and Project administration.

Competing interests

The author declares no competing interests.

References

1. Pinto, P. C., Thiran, P. & Vetterli, M. Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.* 109, 068702 (2012).
2. Shah, D. & Zaman, T. Rumors in a network: who's the culprit? *IEEE Trans. Inf. Theory* 57, 5163-5181 (2011).
3. Prakash, B. A., Vreeken, J. & Faloutsos, C. Spotting culprits in epidemics: how many and which ones? In 2012 IEEE 12th International Conference on Data Mining 11-20 (IEEE, 2012).
4. Gomez-Rodriguez, M., Leskovec, J. & Krause, A. Inferring networks of diffusion and influence. In *Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1019-1028 (ACM, 2010).
5. Lokhov, A. Y., Mézard, M., Ohta, H. & Zecchina, R. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E* 90, 012801 (2014).
6. Bartlett, F. C. *Remembering: A Study in Experimental and Social Psychology* (Cambridge Univ. Press, 1932).
7. Heuer, R. J. *Psychology of Intelligence Analysis* (Center for the Study of Intelligence, CIA, 1999).
8. Schum, D. A. *Evidential Foundations of Probabilistic Reasoning* (Wiley, 1994).
9. Bovens, L. & Hartmann, S. *Bayesian Epistemology* (Oxford Univ. Press, 2003).
10. Milo, R. et al. Network motifs: simple building blocks of complex networks. *Science* 298, 824-827 (2002).
11. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* 359, 1146-1151 (2018).
12. Friggeri, A., Adamic, L. A., Eckles, D. & Cheng, J. Rumor cascades. *Proc. Int. AAAI Conf. Web Soc. Media* 8, 101-110 (2014).

13. Lewin, K. Frontiers in group dynamics: II. Channels of group life. *Hum. Relat.* 1, 143-153 (1947).
14. White, D. M. The "gate keeper": a case study in the selection of news. *Journal. Q.* 27, 383-390 (1950).
15. Burt, R. S. Structural holes and good ideas. *Am. J. Sociol.* 110, 349-399 (2004).
16. Burt, R. S. Reinforced structural holes. *Soc. Networks* 43, 149-161 (2015).
17. DiResta, R. et al. The Tactics & Tropes of the Internet Research Agency (New Knowledge / Stanford Internet Observatory, 2019).
18. Starbird, K. Disinformation's spread: bots, trolls and all of us. *Nature* 571, 449 (2019).
19. Goel, S., Anderson, A., Hofman, J. & Watts, D. J. The structural virality of online diffusion. *Manage. Sci.* 62, 180-196 (2016).
20. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 1130-1132 (2015).
21. Fortunato, S. et al. Science of science. *Science* 359, eaao0185 (2018).
22. Bavelas, A. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.* 22, 725-730 (1950).
23. Leavitt, H. J. Some effects of certain communication patterns on group performance. *J. Abnorm. Soc. Psychol.* 46, 38-50 (1951).
24. Shaw, M. E. Communication networks. In *Advances in Experimental Social Psychology* Vol. 1 (ed. Berkowitz, L.) 111-147 (Academic Press, 1964).
25. Mesoudi, A., Whiten, A. & Dunbar, R. A bias for social information in human cultural transmission. *Br. J. Psychol.* 97, 405-423 (2006).
26. Centola, D. The spread of behavior in an online social network experiment. *Science* 329, 1194-1197 (2010).
27. Centola, D. *How Behavior Spreads: The Science of Complex Contagions* (Princeton Univ. Press, 2018).
28. DiMaggio, P. & Garip, F. Network effects and social inequality. *Annu. Rev. Sociol.* 38, 93-118 (2012).
29. Bail, C. A. et al. Exposure to opposing views on social media can increase political polarization. *Proc. Natl Acad. Sci. USA* 115, 9216-9221 (2018).
30. Borgatti, S. P. Centrality and network flow. *Soc. Networks* 27, 55-71 (2005).
31. Stamatatos, E. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.* 60, 538-556 (2009).

32. Koppel, M., Schler, J. & Argamon, S. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.* 60, 9-26 (2009).
33. Kossinets, G. Effects of missing data in social networks. *Soc. Networks* 28, 247-268 (2006).
34. Sweet, T. M., Qiao, X., Jayasekera, A. & Ding, Y. Investigating the impacts of missing data mechanisms and treatments with latent space models. *Soc. Networks* 86, 42-61 (2026).
35. Borgatti, S. P., Carley, K. M. & Krackhardt, D. On the robustness of centrality measures under conditions of imperfect data. *Soc. Networks* 28, 124-136 (2006).
36. Costenbader, E. & Valente, T. W. The stability of centrality measures when networks are sampled. *Soc. Networks* 25, 283-307 (2003).
37. Linvill, D. L. & Warren, P. L. Troll factories: manufacturing specialized disinformation on Twitter. *Polit. Commun.* 37, 447-467 (2020).
38. Qwen Team. Qwen2.5: A Party of Foundation Models. <https://qwenlm.github.io/blog/qwen2.5/> (2024).
39. Meta AI. The Llama 3 herd of models. Preprint at <https://arxiv.org/abs/2407.21783> (2024).
40. Tian, K. et al. Just ask for calibration: strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* 5433-5442 (ACL, 2023).
41. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proc. 22nd International Conference on Machine Learning* 625-632 (ACM, 2005).
42. Lowenthal, M. M. *Intelligence: From Secrets to Policy* 8th edn (CQ Press, 2019).
43. Office of the Director of National Intelligence. *Intelligence Community Directive 203: Analytic Standards* (ODNI, 2015).