Detecting the Residue of Communication Structure in Messages from Networks of Language-Model Agents: A Controlled Experiment

Eric Gladstone
Independent Researcher
eric.c.gladstone@gmail.com

Abstract

Networks of language-model agents relay, rewrite, and merge one another's messages, and the messages that leave a network need not record the route they took. We test whether a set of terminal messages carries residue that distinguishes among three small, experimentally assigned network structures. In the controlled synthetic networks of the main comparison, observer agents write incident reports from the full record of one constructed event and rewriter agents rewrite them, so the structures differ in whether terminal reports share an upstream text, not in the facts available or in the distribution of any single report. For the primary producer, a classifier given pooled per-report features separates sets of separately drafted reports, each rewritten once, from sets of rewrites of one draft with AUC 0.699, and adding pairwise similarities among the reports raises it to 0.908. The gain is smaller but positive for two other producers and against three more flexible set models that do not explicitly compute pairwise similarities. In post hoc checks, the best of nine unfitted statistics, an overlap of rare word sequences, captures most of the binary signal, and single-report learners tested on unshared texts meet a predefined criterion for equivalence to chance. An exploratory analysis also discriminates among all three structures, including sets in which only two reports share a draft. In planned extensions to five rewrite calls, the primary producer's relational model keeps most of its margin over chance while the pooled baseline loses most of its own.

1 Introduction

Networks of language-model agents pass messages to one another: one agent drafts, and others relay, rewrite, summarize, or merge what they receive (Wu et al., 2024). The messages that leave such a network, its terminal messages, need not record the route they took, but the route can leave residue in how they resemble one another, because rewrites of a common upstream message can retain its wording and organization. When a system or analyst counts several agent-generated messages as distinct accounts, it matters whether they were drafted separately or rewritten from one upstream message. We ask whether a network's structure can be inferred from that residue. We study shared ancestry, which holds when two or more messages were rewritten, directly or through intermediate steps, from the same agent-written message, and the residue it leaves in the terminal messages. Whether the underlying evidence was independently acquired is a separate question that we do not address. Because our networks are small and experimentally assigned, the question is which of a few structures produced a set of messages, not how an arbitrary network routed them.

Answering requires separating two kinds of information that collections of naturally produced texts confound. A message may reveal its history through its own properties, such as length, phrasing, or attribution habits that differ between production procedures. It may also reveal its history only in relation to other messages, through wording or organization shared with another message in its set. When production procedures differ in their instructions, a classifier can recognize the procedure from a single text, and its success then says nothing about relations among texts. We therefore built, for the main comparison, network structures in which the messages are short incident reports and each terminal report has the same marginal distribution under every structure. An observer agent writes a report, called an observation, from an event record, and a rewriter agent rewrites one report. Every terminal report in our main comparison is one rewrite of one observation of the same full record, written by the same producer. In

Author's manuscript, under review at Transactions of the Association for Computational Linguistics. It may differ from any published version. · graystoneindustries.co

the structure we call independent, three observers write three observations separately from the common record. Independent here means separately drafted. That structure and an out-star of three rewrites of a single draft then differ in whether the rewrites share a parent text. A single report carries no designed signal of which structure produced it, and any difference between structures lies in the joint distribution of reports.

Within this design we compare three representations of a set. A single-report model sees one report. A pooled model sees the mean and maximum of 12 per-report features over the three reports, so it sees summaries of all three texts but no pairwise similarity. A relational model adds pairwise similarities: overlap of rare word sequences, embedding similarity, and agreement in sentence order. The main estimate is the difference in held-out AUC between the relational and pooled models on the same sets. It measures what these pairwise similarities add to this pooled summary and does not bound what a more flexible model of the three texts could extract. Because the structures are experimentally assigned, with the record, producer, instructions, and rewrite depth held fixed, a difference between them is attributable to the assigned structure. That structure bundles the sharing of an upstream text with the number of observer calls (one in an out-star, three in the independent structure) and with the order in which observations were generated. Nor does the design identify the linguistic channel of the difference.

For the primary producer after one rewrite, pairwise similarities raise AUC from 0.699 to 0.908, and the gain remains positive for two other producers and against more flexible set models. Much of this signal is plain text reuse. In post hoc checks, the largest overlap of rare word sequences between two reports in a set, used without fitting, reaches 0.863, and on unshared texts both single-report learners meet a predefined criterion for equivalence to chance. After five rewrite calls, the relational model keeps most of its margin over chance and the pooled baseline loses most of its own.

2 Related Work

Several fields recover the copying history of texts from what the copies share. Text-reuse methods detect clusters of reused passages in large collections such as nineteenth-century newspapers, using candidate retrieval, alignment, and clustering (Smith et al., 2013, 2014). Near-duplicate detection estimates document resemblance from shared shingles (Broder, 1997), and plagiarism detection evaluates whether a suspicious document reuses identifiable sources (Potthast et al., 2010). Computational stemmatology reconstructs genealogies of hand-copied manuscripts from shared variants, has been evaluated against artificial traditions with known histories (Roos and Heikkilä, 2009), and has developed models of which kinds of variation carry genealogical signal (Andrews and Macé, 2013) and of transmission with probabilistic phylogenetics (McCollum and Turnbull, 2024). Text phylogeny methods reconstruct version trees for near-duplicate documents (Marmerola et al., 2016). Our relational features belong to this family. Our setting differs in its observation model and baseline. The observations are written separately by the same model from the same record, so reports with no shared parent are already similar, and the classifier never sees the upstream drafts.

Work on machine-generated text asks what a text reveals about the model that produced it. Detectors distinguish generated from human text using cues that vary with decoding and length (Ippolito et al., 2020), or using a model's probability surface (Mitchell et al., 2023), and paraphrasing can defeat them (Sadasivan et al., 2025). Authorship attribution for neural generators identifies which method produced a text and includes a pairwise test of whether two texts came from the same method (Uchendu et al., 2020). These tasks concern the producing model. In our main comparison the producer, decoding, and templates are matched, and the label is whether reports share an upstream text.

Provenance methods for generated text differ in what they assume is observable. Source tracing links each sentence of a generated text to sentences in supplied source documents (Zhu et al., 2025). Multi-turn attribution traces a response span back through the turns of a completed conversation, using the internals of an open-weight model (Tang et al., 2026). Watermarking inserts detectable statistical signals at generation time (Kirchenbauer et al., 2023), and implicit execution tracing extends that idea to multi-agent systems, embedding keyed signals per agent so that an auditor holding the keys can segment a final text and recover transitions between agents (Nian et al., 2026). Each

method has access to something our classifiers lack: the source texts, the conversation, or signals placed during generation. We study uninstrumented outputs without their upstream texts.

3 Design

3.1 Worlds and producers

Each world is a constructed event record of 20 facts, each attributed to a named source, in one of 14 event types (a generator shutdown, a chemical spill, a network outage, and others). Four facts in each record form two disputes, pairs of incompatible claims from different sources. Seventy worlds were drafted by Claude Sonnet 5.5, checked sentence by sentence by GPT-6.1 Terra for facts that settle a dispute, sources that could not know their facts, and missing antecedents, and then read and corrected by hand. Ten were used to verify the pipeline, and the other 60 were used in the main run. Each world was realized in 3 replicates.

Reports were written by three producers from different model families: Gemini 3.5 Flash-Lite (designated the primary producer in the design, for cost), Claude Haiku 5.5, and GPT-6 Luna, at provider-default sampling with reasoning disabled (Haiku, Luna) or set to its minimum (Gemini). Every call is stateless. An observer receives a list of facts with their sources and writes an incident report of about 220 words for an operations team. A rewriter receives only its parent's text, under an instruction such as "Rewrite the text below as an internal incident report." Every report ends with a line naming the sources it used. Instructions come from two families of six paraphrased templates per role, and each world and replicate is assigned one family for all producers. Within each block (one world, replicate, and producer), all observers share one observation template, and each rewrite position (first, second, or third report of a set) draws its own rewrite template independently. Each call is stored under a key that includes its position, so two positions never share an output, even when they draw the same template. Calls with identical inputs and position share one stored output. As a result, structures in a block share some texts: the first report is the same text in all three structures, the Y set's second report is the out-star's second report, and its third report is the independent set's third report. The Y set's shared pair is therefore stored as its first and second reports.

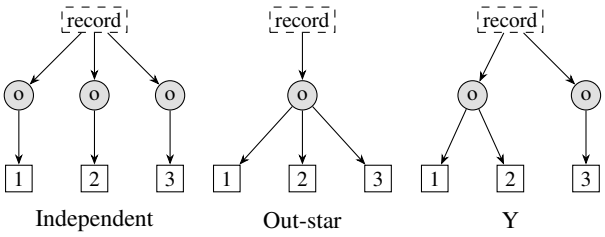


Figure 1: The three set structures under full access. Each observer (circle) writes a report from the full event record, and each terminal report (square) is one rewrite of one observation. The structures differ only in which terminal reports share an observation. In stored Y sets the shared pair is reports 1 and 2, but classifiers see no position labels and their outputs do not change under any reordering of the three reports.

3.2 Matched structures

The main comparison uses sets of three terminal reports about one world (Figure 1). In the independent structure (IND), three observers each write a report from the full record and each report is rewritten once. In the out-star (DEP), one observer writes a report and it is rewritten three times. In the Y structure, two terminal reports share an observation and the third has its own. Each block yields one set of each structure, so there are 180 sets per structure and producer. Under full access, which all main analyses use, every observer sees the full record, so the structures do not differ in what facts were assigned. Because each terminal report is one rewrite of one full-record observation, the distribution of any single terminal report is the same in all three structures. That equality holds by design. Section 4.1 reports an empirical check of it.

Two further arms bound the main result. Under partial access, observers receive overlapping subsets of the record, so shared assigned facts and shared ancestry change together and the arm cannot isolate ancestry. In a concealment arm, rewriters are told to remove source attributions. The trained classifiers were run on the partial-access arm, and the prompted readers on both arms.

Single terminal reports are studied with a second set of structures, generated by the primary producer as incident reports, each condition with 180 reports. Relay depth compares chains of one, three, and five rewrites, each rewrite in a chain being one relay. Spoke count compares a synthesizer (hub) that combines one, two, or three observers, among whom the record is divided, with one cen-

Author's manuscript, under review at Transactions of the Association for Computational Linguistics. It may differ from any published version. · graystoneindustries.co

tral fact given to all. The hub-against-diamond contrast compares a synthesis of three separately drafted observations, each rewritten once, with a synthesis of three rewrites of one observation (the diamond), all with full access. Further contrasts follow the hub and diamond with one or three relays, and compare a five-relay chain with a diamond followed by three relays. Two negative controls, the dead-end controls, add branches that never reach the terminal report to a direct report (an observation that is itself the terminal report) and to a one-relay chain, so if the pipeline is implemented correctly they cannot affect the terminal text.

3.3 Three tiers of evidence

The single-report tier uses 12 features of each report: seven formal features (token and sentence counts, type-token ratio, mean sentence length, mean word length, punctuation density, and numeric tokens) and five attributional features (counts of hedges, uncertainty markers, evidential markers, and source markers from fixed word lists, and an indicator for the presence of a sources line). The indicator is constant within each arm, so it carries no information in any comparison reported here. Every other feature, single-report and relational, is computed on the report body with its sources line removed. A second single-report learner, the embedding learner, uses the body's embedding from nomic-embed-text, a general-purpose text-embedding model. We call the first the feature learner. The pooled tier summarizes a set by the mean and maximum of each single-report feature over its three reports, so it sees all three texts but never compares two of them. The relational tier adds, for each of the three pairs of reports, three similarities: rare-sequence overlap (the Jaccard overlap of rare lower-cased word 3- to 5-grams), embedding cosine, and sentence-order similarity. For sentence-order similarity, each sentence is matched to its most similar sentence in the other report by embedding cosine, and the score is the share of sentence pairs whose matches keep their order, averaged over both directions. Each similarity enters as its minimum, median, and maximum over the three pairs, so the relational features do not depend on the order in which reports are presented. Reordering the three reports of every set in all six orders, with features recomputed and models refitted, leaves every prediction unchanged to numerical precision. An n-gram is rare if it appears in at most two documents of the training texts. Because a world's texts are never in the tables against which it is scored, its names, times, and figures count as rare, and they appear in reports of every structure, since all observers read the same record. Rarity tables are built for each outer fold from its training texts. Because sets in a block share stored reports, each training set is scored against those texts with its own world removed. A held-out set is scored the same way, since its world never appears in the training texts.

We also aggregate single-report predictions: a single-report model, trained on individual reports labeled with their set's structure, predicts each report's structure, and the three are averaged.

3.4 Learning and inference

All learners are L_2-regularized logistic regressions on standardized features. Outer cross-validation uses five folds grouped by world, so no world contributes to both training and testing, and the inverse penalty is chosen from $\{0.01, 0.1, 1, 10\}$ by four-fold world-grouped cross-validation inside each training set. Intervals are 95% world-cluster bootstrap intervals (2,000 draws) on fixed out-of-fold predictions. These intervals describe held-out performance of the fitted procedure on these worlds. The primary set comparison is also tested with a refitted world bootstrap of the gain (200 resamples) that refits and retunes both learners on each resample, holding the fold-level rarity tables fixed. Single-report contrasts are tested with refitted permutations that swap the two structure labels at random within each world and replicate. The design designated three primary tests: the set gain, the hub-against-diamond contrast, and the dead-end controls. The two tests of association are corrected together by the Holm procedure. The dead-end controls and the single-report check on the matched sets use a three-way equivalence rule fixed in the design, and the post hoc unshared-text check applies the same rule: a result whose 90% interval lies wholly inside an AUC band of 0.43 to 0.57 is classified as equivalent to chance, wholly outside it as a difference, and otherwise as inconclusive.

Two general-purpose models, Claude Sonnet 5.5 and GPT-6.1 Terra, also classified sets as prompted readers. Each saw 18 labeled example sets from the fold's training worlds, six per struc-

Author's manuscript, under review at Transactions of the Association for Computational Linguistics. It may differ from any published version. · graystoneindustries.co

Out-star: two rewrites of one observation

At the Harrowgate Pumping Station during the night shift, mains power failed at 02:03 and the backup diesel generator started automatically eleven seconds later.

During the night shift at the Harrowgate Pumping Station, mains power failed at 02:03, and the backup diesel generator started automatically eleven seconds later.

Independent: rewrites of two separate observations

An incident occurred at the Harrowgate Pumping Station when mains power failed at 02:03, prompting the backup diesel generator to start automatically eleven seconds later.

This internal operations report reviews the generator trip at the Harrowgate Pumping Station during a mains power failure.

Figure 2: Opening sentences of two terminal reports from each structure for one world, primary producer, first replicate, verbatim. Both pairs report the same facts. The out-star pair also shares its framing ("during the night shift") and the order of its opening clauses.

ture, with reports in random order, and returned a probability for each structure. Apart from the primary tests, the analyses were exploratory. The cross-producer transfer analysis, a replication of the single-report tests on three new replicates of every world, the two deeper-rewriting runs, and the flexible-model comparison were specified in writing before they were run.

4 Results

4.1 Set-level discrimination of shared ancestry

For the primary producer, a model given the pooled features of three reports separates the independent structure from the out-star with AUC 0.699 [0.657, 0.743]. Adding relations among the reports raises AUC to 0.908 [0.875, 0.937], a gain of 0.209 [0.159, 0.254] over the 360 sets. The refitted bootstrap gives a 95% interval for the gain of 0.156 to 0.287 (one-sided $p = 0.005$, the smallest value attainable with 200 resamples), and the gain remains significant after Holm correction. The gain is also positive for the other two producers: Claude Haiku moves from 0.856 to 0.982 (gain 0.126 [0.090, 0.162]) and GPT Luna from 0.877 to 0.973 (gain 0.097 [0.065, 0.131]). Their smaller gains coincide with higher pooled baselines, which leave less room below 1.0 (Table 1).

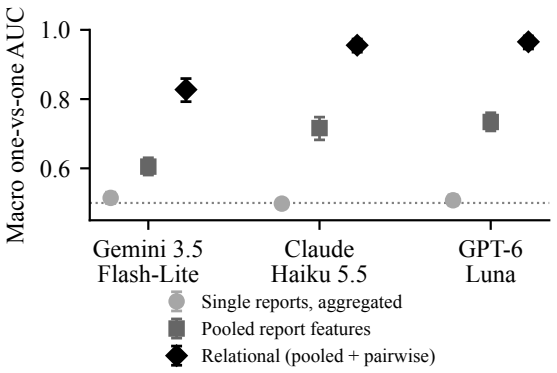


Figure 3: Separating the three structures (independent, out-star, Y) under full access, by producer. Macro one-versus-one AUC with 95% world-cluster intervals for averaged single-report predictions, pooled report features, and the relational model (pooled features plus pairwise similarities). Dotted line: chance.

Most of this binary discrimination does not require fitting to structure labels. In a post hoc check, each of the nine relational statistics was used directly as a score, with rarity still defined on each fold's training texts (Table 3 in Appendix B). The largest rare-sequence overlap among the three pairs of a set performed best of the nine for every producer, and we report it for that reason. It separates the independent structure from the out-star at 0.863 [0.829, 0.895] for the primary producer, 0.985 [0.975, 0.993] for Claude Haiku, and 0.984 [0.972, 0.994] for GPT Luna. For Claude Haiku and GPT Luna its observed AUC is close to that of the trained relational model (0.982 and 0.973), and for the primary producer it is lower than the relational model's 0.908. This check did not estimate paired differences between the score and the trained model, and its intervals do not account for the choice of this statistic among nine. The later depth analyses (Section 4.4) estimated them for the selected statistic. The largest embedding cosine is weaker (0.802, 0.903, 0.888). In these networks, then, shared ancestry in the binary task is visible to a simple text-reuse statistic once the comparison is controlled. The three-structure and configuration tasks in Section 4.2 were not compared with a corresponding unfitted rule.

The pooled baseline is already above chance although each report's distribution is matched. One candidate route is variation within a set, which the mean and maximum partly encode: rewrites of one report may vary less than rewrites of separately drafted observations. The gain over this baseline

	Independent against out-star (AUC)			Three structures (macro AUC)		
Producer	Pooled	Relational	Gain [95%]	Single reports	Pooled	Relational
Gemini 3.5 Flash-Lite	0.699	0.908	0.209 [0.159, 0.254]	0.515	0.605	0.827
Claude Haiku 5.5	0.856	0.982	0.126 [0.090, 0.162]	0.498	0.716	0.955
GPT-6 Luna	0.877	0.973	0.097 [0.065, 0.131]	0.508	0.734	0.965

Table 1: Held-out discrimination of the set structures under full access, by producer, 180 sets per structure. "Single reports" averages the predictions of a model trained on individual reports. Gains are relational minus pooled on the same sets, with world-cluster intervals on fixed out-of-fold predictions. For the primary producer, the refitted bootstrap interval for the gain is 0.156 to 0.287.

measures what the tested pairwise similarities add beyond mean and maximum summaries. Single reports considered one at a time carry much less. Averaged single-report predictions reach macro AUC 0.515 [0.499, 0.531], 0.498 [0.486, 0.509], and 0.508 [0.493, 0.523] for the three producers, against 0.827, 0.955, and 0.965 for the relational model (Figure 3). Because the structures share some stored texts (Section 3.1), a single-report test that draws from every position meets identical texts under different labels, which pushes it toward chance. A post hoc check therefore uses only the second and third reports, which never coincide between the independent and out-star sets. On those texts, both learners meet the rule's criterion for equivalence to chance for every producer: feature learner 0.527 [0.496, 0.559], 0.511 [0.485, 0.538], and 0.538 [0.512, 0.566], and embedding learner 0.490 [0.469, 0.509], 0.513 [0.494, 0.533], and 0.486 [0.473, 0.497] (90% intervals). Both GPT Luna intervals exclude 0.5 while lying inside the band, one above and one below it. One report drawn at random from each set, shared texts included, gives 0.545 [0.509, 0.580] for the primary producer, which is inconclusive, and when the sets are split by their two instruction-template families, the result furthest from chance reaches 0.606 [0.554, 0.659], also inconclusive.

Because the pooled baseline is simple, a check planned before it was run compared the relational model with three more flexible set models that do not explicitly compute pairwise similarities (Appendix A): richer summaries of the 12 features (mean, minimum, maximum, and standard deviation) with logistic regression, the same summaries with gradient boosting, and the mean and per-dimension standard deviation of the three reports' embeddings with logistic regression. Standard deviations measure dispersion within a set and can encode similarity implicitly. The strongest comparator was the embedding model for the primary producer (0.786 [0.752, 0.821]) and the richer logistic summaries for Claude Haiku and GPT Luna (0.876 and 0.888). The relational model exceeds the strongest comparator for every producer, by +0.122 [+0.091, +0.153], +0.106 [+0.072, +0.142], and +0.085 [+0.053, +0.121]. For the primary producer this is about three-fifths of the gain over the reference baseline (0.209). By the plan's readings the gain holds for every producer. The plan would have attributed most of the gain to the limits of the pooled baseline if the strongest comparator had fallen within the relational model's interval, and none did. Gradient boosting on the relational features did not detectably improve on logistic regression for any producer (Table 5 in Appendix B).

Figure 2 shows the kind of evidence the relational tier uses. Reports rewritten from separately drafted observations share the facts of the record, including its times and figures, because every observer saw the same record. Rewrites of one observation can share, in addition, choices made in the text they rewrite: which detail frames the opening, the order of clauses, and particular phrasings.

Prompted general-purpose models, given labeled examples but no engineered features, also discriminate the structures. Given 18 labeled example sets drawn from the fold's training worlds, the same sets for both readers, and reports including their sources lines, Claude Sonnet 5.5 separates the independent structure from the out-star at 0.865 [0.805, 0.917] and GPT-6.1 Terra at 0.935 [0.888, 0.974], computed from their probabilities for the two structures. On all three structures their macro AUCs are 0.749 and 0.850. The readers and classifiers differ in supervision and in the data they were evaluated on (the readers saw the first replicate only), so the comparison is descriptive, not a matched test.

4.2 Which relationships can be reconstructed

Locating the shared pair is a separate task. A pair model, a logistic regression on the three similarities of a single pair, trained on all pairs of all three structures, separates pairs that share an observation from pairs that do not with AUC 0.863 [0.833, 0.890] for the primary producer, as do all results in this paragraph. The full predictor assigns each set one of five configurations: independent, out-star, or Y with each of the three reports as the odd one. It combines the three-structure model with the pair scores, taking the odd report in a Y set to be the one outside the most probable shared pair. Given that a set is a Y, the odd report is identified in 88.3% of sets. Without that knowledge, the predictor assigns the exact configuration to 64.1% of all 540 sets and to 66 of the 180 Y sets. Figure 4a shows where the Y sets go. Most failures occur at the structure stage, where 86 Y sets are classified as independent or out-star. A further 28 are assigned Y by the three-structure model but lose to the independent or out-star configuration once the Y probability is divided among three odd positions. None is assigned Y with the wrong odd report, so the Y sets whose pair scores point to the wrong odd report all fall among those the predictor does not assign to Y.

Relational models also transfer across producers. In a planned analysis we trained on one producer's sets and tested on another's, with training and test worlds disjoint. In all six ordered pairs, the transferred relational model exceeds the transferred pooled baseline (Figure 4b, Table 6). Transfer between Claude Haiku and GPT Luna shows no detectable paired loss in either direction, and both intervals include zero. Every detectable loss involves Gemini. In the binary task the largest is from GPT Luna to Gemini, which reaches 0.788 [0.750, 0.831], a paired change of -0.119 [-0.153, -0.078] relative to a model trained on Gemini itself. In the three-structure task, every transfer to or from Gemini changes macro AUC by between -0.077 and -0.100, for example -0.099 from Gemini to GPT Luna and -0.100 from Claude Haiku to Gemini.

4.3 What carries the signal

The three relational families are correlated, so removing one family measures what the others cannot replace rather than its share of a causal process. Removing rare-sequence overlap is the only ablation that lowers AUC for all three producers, with paired changes of -0.029 [-0.047, -0.013], -0.039 [-0.057, -0.021], and -0.041 [-0.059, -0.024]. For the primary producer, removing sentence-order similarity costs more (-0.041 [-0.063, -0.019]), but not for the other two. Without rare-sequence overlap, the relational model still exceeds the pooled baseline for every producer, by 0.180 [0.128, 0.226], 0.087 [0.054, 0.120], and 0.056 [0.019, 0.092]. That residual gain does not isolate a channel of transmission, because rare wording is still present in the texts that the embeddings and sentence matching operate on. For the same reason, the small cost of removing explicit embedding similarity does not show that similarity of meaning is unimportant. In the pair model that locates the odd report, rare-sequence overlap alone identifies it in 86.7% of Y sets, against 70.6% for sentence-order similarity alone, 68.9% for embedding cosine alone, and 88.3% with all three. Table 2 in Appendix B gives the full ablation.

4.4 Shared ancestry after further rewriting

Every set in the main comparison has one rewrite between an observation and each terminal report. In a planned extension, each final report in the primary producer's full-access sets was rewritten four more times in a chain, each step with a fresh rewrite template from the block's family, and the sets were analyzed after two, three, and five rewrites (Appendix A). Texts shared between structures stay shared at every depth, so each terminal report still has the same distribution under every structure. The plan fixed three readings before the run, and none was met as specified. It defined decay as a fall in both the relational AUC and its gain over the pooled baseline, divergence as a growing advantage of the relational model over the overlap score, and persistence as a relational AUC after five rewrites within the one-rewrite interval. The relational model reaches 0.883 [0.848, 0.916], 0.875 [0.837, 0.911], and 0.855 [0.816, 0.893] after two, three, and five rewrites, against 0.908 [0.875, 0.937] after one, so it declined, and its value after five rewrites lies below the one-rewrite interval (Figure 5). The pooled baseline declined more, to 0.577 [0.528, 0.625], so the gain rose from 0.209 to 0.278 [0.231, 0.329] instead of falling. Changes between depths are descriptive, because no paired interval across depths was

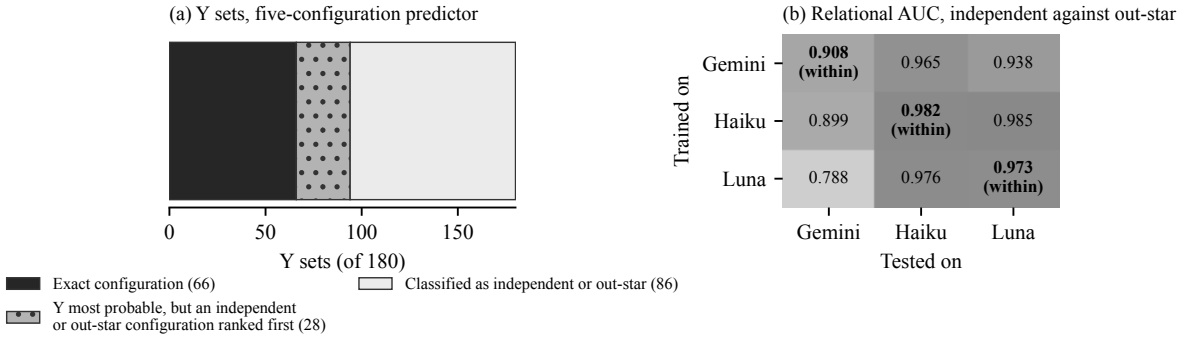


Figure 4: (a) Outcomes for the 180 Y sets under the five-configuration predictor, primary producer. No set recognized as Y was assigned the wrong odd report (0 sets). (b) AUC for the independent structure against the out-star when the relational model is trained on one producer's sets and tested on another's, with training and test worlds disjoint.

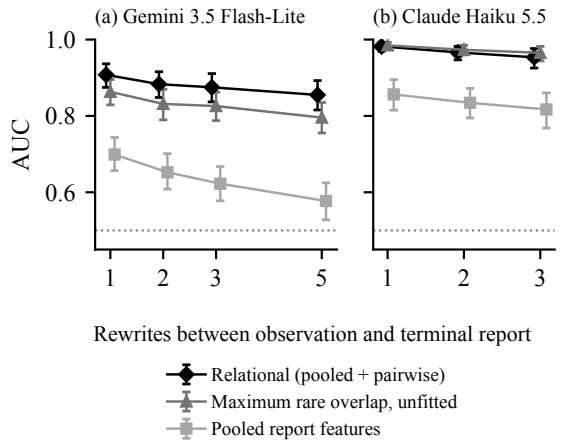


Figure 5: Independent structure against out-star as further rewrites separate each observation from its terminal report: (a) the primary producer, to five rewrites, (b) Claude Haiku, to three. AUC with 95% world-cluster intervals. One rewrite is the main comparison. Dotted line: chance.

computed. For this producer and these chains, the tested pairwise features retain much more of their discrimination after five rewrite calls than the mean and maximum summaries do.

The unfitted maximum rare overlap falls from 0.863 to 0.796 [0.755, 0.835]. That score was chosen after inspecting nine statistics at one rewrite (Section 4.1), so the paired difference at one rewrite, +0.045 [+0.016, +0.074], is estimated on the data used to choose it. The plan then fixed the same comparison for the deeper texts before they were generated. The relational model is higher after two, three, and five rewrites, by +0.052 [+0.020, +0.087], +0.049 [+0.014, +0.085], and +0.059 [+0.024, +0.096], and the difference did not grow detectably. The deeper texts come from the same worlds and the same one-rewrite reports, so this is a planned follow-on comparison, not a validation on new worlds. Three-structure macro AUC falls from 0.827 to 0.754 [0.716, 0.791]. Single reports from unshared texts met the equivalence criterion for the embedding learner at every depth. For the feature learner they were inconclusive under that rule after three and five rewrites (0.546 [0.519, 0.571] and 0.547 [0.520, 0.574]), and neither interval lies outside the band. Table 4 in Appendix B gives all nine unfitted statistics at each depth.

A second planned run extended Claude Haiku's sets to three rewrites, as far as the budget allowed, with descriptive readings because Haiku starts near ceiling. The relational model reaches 0.966 [0.947, 0.982] and 0.954 [0.925, 0.977] after two and three rewrites, against 0.982 after one, and the pooled baseline falls from 0.856 to 0.817 [0.768, 0.860] (Figure 5b). For Haiku the unfitted overlap score did not differ detectably from the relational model at any depth (paired differences −0.007 [−0.020, +0.005] and −0.012 [−0.033, +0.006] after two and three rewrites), and single reports met the equivalence criterion under both learners at every depth. Against the strongest of the three flexible set models, the relational model's advantage after the deepest rewriting was +0.158 [+0.111, +0.206] for the primary producer and +0.122 [+0.077, +0.168] for Haiku.

Depth here counts rewrite calls, and some calls returned their input unchanged. In a post hoc count this was rare at one rewrite, where 892 of

900 chains changed for each producer, but it grew along the chains. The primary producer's chains changed the text at 4.1 of five calls on average and at every call in 370 of 900. A changed text can still be close to its input. The counts therefore bound the number of substantive rewrites from above and do not measure how much the wording changed.

4.5 Boundaries of the set result

Two arms and a check that removes the sources line bound the main comparison. When rewriters were instructed to remove source attributions, prompted-reader performance was higher, a difference that was not tested. The prompted readers separated the independent structure from the out-star at 0.941 [0.903, 0.972] (Claude Sonnet 5.5) and 0.977 [0.955, 0.993] (GPT-6.1 Terra) in the concealment arm, against 0.865 and 0.935 without the instruction. Removing the sources line after generation changed the readers' performance by less than 0.03 (0.883 and 0.906, also untested), so the line was not necessary for their discrimination. The trained classifiers never saw the sources line. Under partial access, where observers in the independent structure receive different subsets of the record and observers sharing an ancestor necessarily share one subset, the relational model reaches 0.994 [0.988, 0.998] over a pooled baseline of 0.849, and the readers reach 0.999 and 1.000. That arm changes shared ancestry and shared assigned facts together, so it shows only that the combined contrast is easier to discriminate. A post hoc check examined when the reports that differ between the independent and out-star sets, the second and third, were generated. Neither structure's reports came consistently first: the out-star's had the earlier mean generation time in 102, 94, and 97 of 180 blocks for the three producers (two-sided sign tests treating blocks as independent, $p = 0.086$, 0.602, and 0.333), and the two mean times differed by less than 40 seconds in every block. The observations were less balanced, because the out-star's observation is the first report's observation, which every structure in a block needs and which is therefore usually generated early.

4.6 The limits of single-report recovery

Under the feature learner, the contrast between each dead-end control and the same structure without its unused branches met the criterion for equivalence to chance for the direct report and the one-relay chain (0.491 [0.464, 0.517] and 0.510 [0.482, 0.538]). Under the embedding learner, the one-relay control was also equivalent (0.502 [0.479, 0.526]), but the direct-report control was inconclusive (0.553 [0.531, 0.577]), with a refitted permutation $p = 0.015$ against the null of no difference. In the main run the direct report's observation, which many conditions share, was generated before its control's in most blocks. A replication that generated the controls first, in random order, found the contrast equivalent to chance (0.493 [0.471, 0.514]). It does not establish the cause, and generation order is one candidate.

A single report carries weak information about parts of its history. Chains of one, three, and five relays are ranked above chance at macro AUC 0.599 [0.580, 0.621], and hubs with one, two, or three spokes at 0.613 [0.584, 0.644]. Both are weak discrimination, not recovery of relay or spoke counts. Whether a synthesis combined separately drafted or shared inputs leaves a small trace. A hub of three separately drafted observations is separated from a diamond of three rewrites of one observation at 0.552 [0.509, 0.596] (permutation $p = 0.023$, significant after Holm correction), and the replication on three new replicates gave 0.568 [0.519, 0.620]. After one further relay the trace is 0.546 [0.509, 0.587], and after three its interval includes chance (0.542 [0.496, 0.589]). It was not detected after transfer from the primary producer to Claude Haiku (0.499 [0.472, 0.528]) or to a news-brief genre (0.509 [0.479, 0.541]), and the embedding learner did not detect it (0.515 [0.482, 0.546]). The feature learner did not establish discrimination between a five-relay chain and a diamond followed by three relays (0.489 [0.442, 0.535]). Appendix B summarizes these results.

5 Discussion

In the one-rewrite set comparisons, the residue of shared ancestry lies in what terminal reports share with one another and, in a post hoc check on unshared texts, hardly at all in any single report. Pooled per-report features capture part of the difference, perhaps through variation within a set, and pairwise similarities raise discrimination to 0.908 for the primary producer.

The relevant baseline is separately drafted reports of the same facts by the same model. These are highly similar, because they share facts, sources, and the producer's habits. The rela-

tional features detect the added similarity among rewrites of a shared parent text, for all three producers, and for two of them a single overlap statistic comes close to the trained model in the binary task. For the primary producer, the trained model exceeds it by about 0.05 in the planned deeper comparisons. The task therefore resembles text-reuse detection and stemmatology more than model attribution. It differs from both because the copies are paraphrases and the baseline is separate writing from common facts rather than unrelated text. The controlled comparison measures how well the tested models distinguish assigned communication structures from terminal messages after one to five rewrite calls, and how that discrimination changes with further rewriting. After five rewrite calls the primary producer's relational model keeps most of its margin over chance while the pooled baseline loses most of its own, and Claude Haiku's relational model stays high after three. Against the strongest of three more flexible set models, explicit pairwise similarities still add discrimination for every producer. Separating the independent structure from the out-star is close to ceiling for two producers, but once Y sets are included the exact configuration is assigned to 64.1% of sets, and most Y sets are lost because the set is classified as independent or out-star.

For a system or analyst judging whether agreeing messages from an agent network were drafted separately, the result suggests comparing the messages with one another, at least when single messages are matched in distribution, as they are here. Two features of the evidence qualify that suggestion. First, shared assigned facts also make reports alike, and the partial-access contrast, which combines them with shared ancestry, was easier to discriminate than the full-access one. A relational check alone therefore cannot attribute similarity to a shared upstream message rather than to shared source material. Second, the trained models were fit on labeled examples from the same kind of network, so their transfer to networks with different instructions, depths, or producers is an empirical question, answered here only for three producers.

6 Limitations

The evidence concerns specific networks, and five limits follow from the design. First, the main comparison has one rewrite between a shared parent and each terminal report. The extensions reached five rewrites for one producer and three for a second, used chains of single rewrites with the same template pools, and count rewrite calls, some of which returned their input unchanged. Deeper networks that also summarize, synthesize, or switch producers are untested. Second, the worlds are synthetic, drawn from one generator in one genre for the set studies, and the reports are short. Third, each set came from one producer writing within fixed template pools, and cross-producer transfer changed AUC by as much as -0.119 (GPT Luna to Gemini). Mixed-producer sets were not tested. Fourth, no producer was adversarial. The concealment instruction removed attributions but did not ask rewriters to disguise their source, so the results do not bound what a deliberate attempt to hide shared ancestry could achieve. Fifth, the labels are the constructed structures, so the results say nothing about transmission direction, agent identity, missing intermediate reports, or general graph reconstruction, and nothing about human writers.

Two limits concern inference. The intervals we report are conditional on the fitted procedures and these worlds, and the refitted bootstrap of the main gain refits the learners while holding the fold-level rarity tables fixed. Apart from the three primary tests, the analyses were exploratory. Of these, only the cross-producer transfer, the two deeper-rewriting runs, the flexible-model comparison, and the single-report replication were specified before they were run. The unshared-text, unfitted-score, generation-order, and copying checks were added after the main results were known.

Reproducibility Statement

The code, prompts, constructed worlds, the stored model calls behind the reported analyses with their exact prompts and outputs, the analysis records, and the plans written before the planned analyses are publicly available under open licenses. Outputs of a language-model judge on a separate validation sample, which no reported result uses, are withheld until human coding of that sample is complete. The location is withheld for anonymous review.

References

Tara L. Andrews and Caroline Macé. 2013. Beyond the tree of texts: Building an empirical model of scribal variation through graph anal-

ysis of texts and stemmata. *Literary and Linguistic Computing*, 28(4):504–521.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997*, pages 21–29. IEEE.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Guilherme D. Marmerola, Marina A. Oikawa, Zanoni Dias, Siome Goldenstein, and Anderson Rocha. 2016. On the reconstruction of text phylogeny trees: Evaluation and analysis of textual relationships. *PLOS ONE*, 11(12):e0167822.

Joey McCollum and Robert Turnbull. 2024. Using Bayesian phylogenetics to infer manuscript transmission history. *Digital Scholarship in the Humanities*, 39(1):258–279.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Yi Nian, Haosen Cao, Shenzhe Zhu, Henry Peng Zou, Qingqing Luan, Yudi Zhang, and Yue Zhao. 2026. When only the final text survives: Implicit execution tracing for multi-agent auditing. arXiv:2603.17445v5.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*, pages 997–1005, Beijing, China. Coling 2010 Organizing Committee.

Teemu Roos and Tuomas Heikkilä. 2009. Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, 24(4):417–433.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*.

David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *2013 IEEE International Conference on Big Data*, pages 86–94. IEEE.

David A. Smith, Ryan Cordell, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. Detecting and modeling local text reuse. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 183–192. IEEE.

Jessica Tang, Shraddha Barke, and Sharad Agarwal. 2026. Tokengeist: Multi-turn attribution tracing in agentic conversations. arXiv:2607.22610v1.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Junnan Zhu, Min Xiao, Yining Wang, Feifei Zhai, Yu Zhou, and Chengqing Zong. 2025. TROVE: A challenge for fine-grained text provenance via source sentence tracing and relationship classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

11755–11771, Vienna, Austria. Association for Computational Linguistics.

A Methodological Details

Instructions. Observer templates take the form "Write [genre noun] based on the facts listed below," rewrite templates "Rewrite the text below as [genre noun]," and synthesis templates "Write [genre noun] based on the material below," with six paraphrases per role in each of two families. Every instruction adds a style line (incident reports: "Write it for an operations team reviewing the event"), a format line ("Write the report body as plain paragraphs, without headings, bullet points, or bold text. Aim for about 220 words in the body"), and a sources line ("End with a final line beginning 'Sources:' followed only by the names of sources explicitly stated in the supplied material for information included in the report. Do not invent source names"). The concealment arm replaces the sources line with an instruction to omit it and to remove explicit attributions without changing the reported events, quantities, disagreements, or uncertainty. A second genre, a news brief of about 150 words for a general readership, is used only for single-report transfer.

Validation and storage. An output is valid if its body is between 0.5 and 2.0 times the genre's word target. Invalid outputs are retried up to four times. Each call is keyed by its study, world, replicate, genre, producer, template family, role (which names the call's position, such as the second rewrite of a set), template, fact package, parents' keys, and, where a condition must not share outputs, a condition tag, so identical requests share one stored output. Distinct calls can still return identical text. In 3 of Claude Haiku's 180 out-star sets, two distinct rewrite calls returned byte-identical text. The two calls used the same template in one of these sets and different templates in the other two. This did not occur for the other producers.

Partial access. Each world has a pool of six equal-size packages, each containing the record's central fact (the event itself) and a random half of the remaining facts. A set's three observers receive three packages drawn without replacement.

Rarity tables. For each outer fold, document frequencies of word 3- to 5-grams are counted over the report bodies of the training sets, with every stored copy of a report counted once per set in which it appears. A training set is scored against these counts with all of its own world's texts subtracted. A held-out set is scored against the full training counts, which never contain its world. An n-gram is rare if its count is at most two.

Five-configuration predictor. Let $P(\mathrm{I}), P(\mathrm{D}), P(\mathrm{Y})$ be the relational set model's probabilities for the independent structure, the out-star, and Y, and let s_{12}, s_{13}, s_{23} be the out-of-fold same-parent probabilities of the pair model for the three pairs. The odd report in a Y set is k with weight q_k proportional to the score of the pair that excludes k ($q_1 \propto s_{23}$, $q_2 \propto s_{13}$, $q_3 \propto s_{12}$, normalized to sum to one). The five configuration scores are $P(\mathrm{I})$, $P(\mathrm{D})$, and $P(\mathrm{Y})\, q_k$ for $k = 1, 2, 3$, and the predicted configuration is the largest. The rule is not separately calibrated.

Prompted readers. For each outer fold, 18 example sets, six per structure, were drawn from the training worlds with a seed that depends on the task and fold but not on the reader, so both readers saw the same examples. Reports appeared in random order with their sources lines, and readers returned a probability for each structure. Readers were evaluated on the first replicate of each world.

Deeper rewriting. Each of the five distinct terminal reports in a primary-producer block (the first report, shared by all structures, and the second and third reports of the independent and out-star sets) was rewritten four more times in a chain. Each step drew a rewrite template from the block's family with a seed set by block, position, and step, and its call key included position and step, so texts shared after one rewrite remain shared at every depth. The run made 3,600 calls, all valid at the first attempt. Each depth is analyzed with the folds, rarity construction, learners, and intervals of the main comparison, and the analysis at one rewrite reproduces the main results. Texts after four rewrites were generated as links in the chains and not analyzed. The Claude Haiku run used the same code and rules, stopped at three rewrites, and made 1,800 calls, all valid at the first attempt. A rewrite step changed the text if its output was not identical to its input.

Flexible set models. The richer summaries are the mean, minimum, maximum, and stan-

Author's manuscript, under review at Transactions of the Association for Computational Linguistics. It may differ from any published version. · graystoneindustries.co

dard deviation of each of the 12 per-report features over a set's three reports. Gradient boosting uses histogram-based trees, with learning rate in $\{0.05, 0.1\}$ and maximum depth in $\{3, \text{unlimited}\}$ chosen by four-fold world-grouped cross-validation inside each training set. The embedding set model uses the mean and per-dimension standard deviation of the three reports' nomic-embed-text embeddings. The boosted relational model applies the same boosting to the relational model's features, with each fold's rarity tables. Folds, intervals, and paired differences are those of the main comparison.

Cross-producer transfer. For each outer fold, the model is trained on one producer's sets from the training worlds and tested on another producer's sets from the fold's worlds, with penalties selected on the training producer and rarity tables built from the training producer's texts.

B Complementary Results

Relational features	Gemini	Haiku	Luna
All three families	0.908	0.982	0.973
Without rare-sequence overlap	0.879	0.943	0.932
Without sentence order	0.867	0.981	0.976
Without embedding cosine	0.912	0.978	0.987
Rare-sequence overlap alone	0.864	0.978	0.986
Sentence order alone	0.845	0.933	0.937
Embedding cosine alone	0.799	0.909	0.910

Table 2: Ablation of the relational families, independent against out-star, AUC. Each variant adds the named families to the pooled baseline.

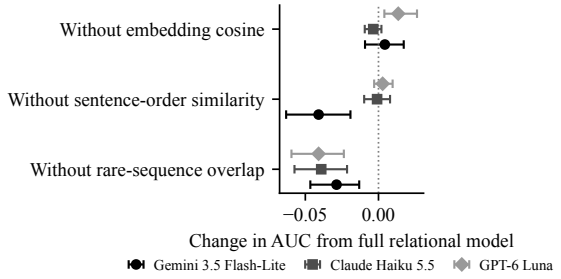


Figure 6: Change in AUC for the independent structure against the out-star when each relational family is removed, by producer (circles: Gemini, squares: Haiku, diamonds: Luna), with 95% world-cluster intervals.

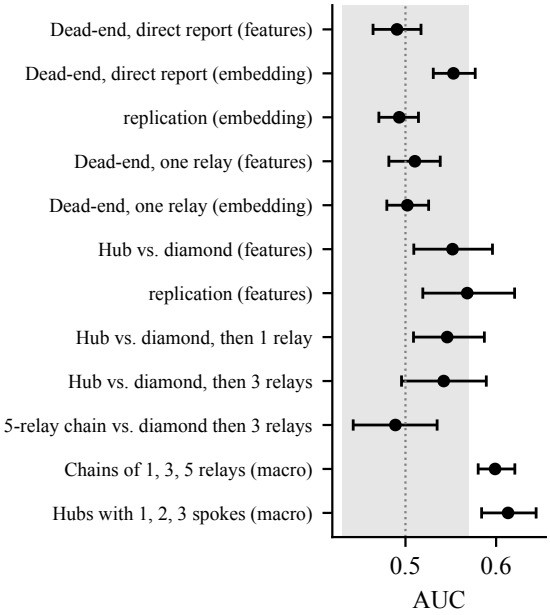


Figure 7: Single-report contrasts, primary producer, feature learner unless noted. Intervals are 90% for the dead-end controls and their replication and 95% otherwise. Each replication row repeats the row above it on three new replicates. Shaded: the equivalence band, which in this table applies only to the dead-end controls.

Score	Gemini	Haiku	Luna
Rare-sequence overlap, max	0.863 [0.829, 0.895]	0.985 [0.975, 0.993]	0.984 [0.972, 0.994]
Rare-sequence overlap, median	0.846 [0.812, 0.880]	0.978 [0.966, 0.989]	0.983 [0.971, 0.993]
Rare-sequence overlap, min	0.814 [0.777, 0.851]	0.973 [0.958, 0.986]	0.980 [0.967, 0.990]
Embedding cosine, max	0.802 [0.765, 0.839]	0.903 [0.866, 0.938]	0.888 [0.849, 0.924]
Embedding cosine, median	0.768 [0.723, 0.811]	0.784 [0.744, 0.822]	0.863 [0.819, 0.907]
Embedding cosine, min	0.769 [0.724, 0.811]	0.796 [0.757, 0.833]	0.860 [0.817, 0.904]
Sentence order, max	0.739 [0.690, 0.789]	0.820 [0.773, 0.870]	0.779 [0.726, 0.835]
Sentence order, median	0.784 [0.734, 0.832]	0.870 [0.828, 0.911]	0.829 [0.783, 0.874]
Sentence order, min	0.815 [0.767, 0.861]	0.894 [0.857, 0.930]	0.859 [0.817, 0.900]

Table 3: Post hoc check: each relational statistic used directly as a score for independent against out-star, without fitting to structure labels, AUC with 95% world-cluster intervals. Rarity is defined on each fold's training texts as in the main analysis.

Score	1 rewrite	2 rewrites	3 rewrites	5 rewrites
Rare-sequence overlap, max	0.863 [0.829, 0.895]	0.831 [0.790, 0.870]	0.826 [0.788, 0.862]	0.796 [0.755, 0.835]
Rare-sequence overlap, median	0.846 [0.812, 0.880]	0.782 [0.736, 0.825]	0.774 [0.730, 0.816]	0.745 [0.703, 0.787]
Rare-sequence overlap, min	0.814 [0.777, 0.851]	0.773 [0.728, 0.816]	0.756 [0.716, 0.797]	0.732 [0.690, 0.775]
Embedding cosine, max	0.802 [0.765, 0.839]	0.771 [0.726, 0.812]	0.710 [0.660, 0.761]	0.691 [0.643, 0.736]
Embedding cosine, median	0.768 [0.723, 0.811]	0.717 [0.669, 0.765]	0.708 [0.661, 0.756]	0.692 [0.648, 0.737]
Embedding cosine, min	0.769 [0.724, 0.811]	0.702 [0.654, 0.751]	0.699 [0.653, 0.746]	0.676 [0.627, 0.721]
Sentence order, max	0.739 [0.690, 0.789]	0.734 [0.687, 0.784]	0.713 [0.665, 0.759]	0.724 [0.672, 0.775]
Sentence order, median	0.784 [0.734, 0.832]	0.788 [0.738, 0.834]	0.765 [0.711, 0.818]	0.749 [0.693, 0.802]
Sentence order, min	0.815 [0.767, 0.861]	0.806 [0.759, 0.852]	0.794 [0.744, 0.842]	0.766 [0.712, 0.814]

Table 4: Deeper-rewriting extension, primary producer: each relational statistic used directly as a score for independent against out-star, without fitting to structure labels, AUC with 95% world-cluster intervals. The table was specified in the extension's plan. The one-rewrite column repeats the first column of Table 3.

Producer, rewrites	Pooled	Rich, logistic	Rich, boosted	Embedding set	Relational	Relational, boosted
Gemini, 1	0.699	0.685	0.738	0.786	0.908	0.890
Haiku, 1	0.856	0.876	0.847	0.861	0.982	0.987
Luna, 1	0.877	0.888	0.872	0.880	0.973	0.984
Gemini, 5	0.577	0.568	0.592	0.697	0.855	0.854
Haiku, 3	0.817	0.832	0.756	0.802	0.954	0.962

Table 5: Independent against out-star under the reference pooled and relational models and the three more flexible set models, AUC. Intervals and paired differences are in the analysis record. The comparison was planned after the main results were known and before it was run.

Task	Trained on	Tested on	Pooled	Relational	Relational minus within-producer
Ind. vs. out-star	Gemini	Gemini	0.699 [0.657, 0.743]	0.908 [0.875, 0.937]	within
Ind. vs. out-star	Gemini	Haiku	0.741 [0.694, 0.791]	0.965 [0.947, 0.980]	−0.017 [−0.032, −0.003]
Ind. vs. out-star	Gemini	Luna	0.674 [0.624, 0.731]	0.938 [0.912, 0.962]	−0.035 [−0.064, −0.007]
Ind. vs. out-star	Haiku	Gemini	0.676 [0.627, 0.723]	0.899 [0.867, 0.929]	−0.008 [−0.026, +0.009]
Ind. vs. out-star	Haiku	Haiku	0.856 [0.815, 0.895]	0.982 [0.970, 0.992]	within
Ind. vs. out-star	Haiku	Luna	0.854 [0.818, 0.890]	0.985 [0.974, 0.993]	+0.011 [−0.002, +0.028]
Ind. vs. out-star	Luna	Gemini	0.676 [0.627, 0.725]	0.788 [0.750, 0.831]	−0.119 [−0.153, −0.078]
Ind. vs. out-star	Luna	Haiku	0.858 [0.820, 0.892]	0.976 [0.960, 0.988]	−0.006 [−0.017, +0.003]
Ind. vs. out-star	Luna	Luna	0.877 [0.841, 0.912]	0.973 [0.953, 0.988]	within
Three structures	Gemini	Gemini	0.605 [0.581, 0.630]	0.827 [0.793, 0.859]	within
Three structures	Gemini	Haiku	0.640 [0.616, 0.665]	0.878 [0.857, 0.899]	−0.077 [−0.086, −0.067]
Three structures	Gemini	Luna	0.612 [0.587, 0.639]	0.866 [0.844, 0.888]	−0.099 [−0.116, −0.083]
Three structures	Haiku	Gemini	0.592 [0.565, 0.620]	0.727 [0.694, 0.760]	−0.100 [−0.130, −0.073]
Three structures	Haiku	Haiku	0.716 [0.682, 0.748]	0.955 [0.936, 0.973]	within
Three structures	Haiku	Luna	0.697 [0.673, 0.722]	0.956 [0.939, 0.974]	−0.009 [−0.025, +0.008]
Three structures	Luna	Gemini	0.574 [0.547, 0.600]	0.728 [0.701, 0.757]	−0.099 [−0.121, −0.077]
Three structures	Luna	Haiku	0.715 [0.683, 0.744]	0.958 [0.940, 0.974]	+0.003 [−0.007, +0.013]
Three structures	Luna	Luna	0.734 [0.708, 0.761]	0.965 [0.945, 0.982]	within

Table 6: Cross-producer transfer of the set models, full access. AUC for the independent structure (IND) against the out-star (DEP), and macro one-versus-one AUC for three structures, with 95% world-cluster intervals. The last column is the paired change from the test producer's own relational model on the same sets.