Periodic Locking and Topic Retention in Long Dialogues Between Two Small Language Model Agents

Jiexiao Zhang
discoveryopen.org
Correspondence: jiexiaozhang1@outlook.com

Abstract

Objective: To examine the joint evolution of textual repetition, retention of opening-topic information, and generation costs in long dialogues between two agents using a small language model with fixed weights and no external task feedback, and to test the effects of simple prompts and branching interventions on repetitive states. Methods: Qwen3.5-0.8B generated alternating dialogues under five fictional Chinese opening scenarios and two system-prompt conditions. The analysis included 621 complete conversations of 200 messages each, totaling 124,200 messages; 620 conversations formed 310 matched pairs. Trajectories were characterized using exact-text periodicity, character-bigram similarity, lexical novelty, Chinese semantic embeddings, grouped cross-validation, and paired resampling. Ten frozen history prefixes were also continued under five conditions and two new seeds, yielding 100 branches and 1,200 additional messages. Results: The final 40 messages of 610 conversations (98.23%) exhibited a strict exact-text period of at most 10 messages; periods 1 and 2 accounted for 299 and 280 conversations, respectively. Same-speaker exact repetition reached 94.60% in the final 20 messages, while opening-category decoding accuracy remained 71.82% using the final two messages. Adding a topic-anchoring instruction produced a paired difference of 1.13 percentage points in overall same-speaker repetition, with a 95% confidence interval of −2.21 to 4.49 percentage points under replicate-block resampling. Logistic regression using features from the first 20 messages achieved an AUC of 0.720 for predicting confirmation of near-repetition at or before message 80; more complex models showed no consistent advantage. A stopping rule requiring 20 consecutive near-repetitions would retrospectively have avoided 63.57% of messages and 85.77% of recorded generation time. In the branch experiment, the text-change fraction across all 12 continuation messages was 7.50% under the recommended sampling configuration and zero under the original configuration and the other three interventions. The exact cluster test yielded p=0.25 and the Benjamini–Hochberg-adjusted q value for the primary comparisons was 1.00; sustained change across the final six messages occurred in only 1/20 branches under the recommended configuration. Conclusions: Under the model, opening scenarios, template protocol, and sampling settings examined here, periodic locking can coexist with retention of opening-topic information. Evaluation of long dialogues should separately measure structural repetition, topic identifiability, and informational usefulness, while using controlled interventions to test the mutability of repetitive states.

Keywords: small language models; two-agent dialogue; periodic repetition; topic retention; generation degeneration; paired interventions

1 Introduction

Multi-agent language-model systems organize reasoning, collaboration, and task execution through message exchange. A conversation may continue to produce fluent sentences without continuing to provide new, distinguishing information. For an automatically running local system, whether a reply remains relevant to the topic, merely restates earlier text, or provides a practical benefit from continued generation are three different questions. If evaluation considers only relevance or the fluency of

individual replies, loops that repeatedly confirm the same intention may be counted as effective interaction. If it considers only whether strings change, numerical substitutions and template variations may conceal conversational stagnation. Long dialogues therefore need to be examined as sequential processes with historical feedback, rather than solely through scores assigned to isolated replies.

Repetitive degeneration in autoregressive text generation has been studied systematically. Holtzman et al. examined the relationship between decoding strategies and text degeneration [1]; Welleck et al. proposed unlikelihood training to suppress repetitive outputs [2]; and Xu et al. investigated sentence-level repetition and its self-reinforcing behavior [3]. These studies indicate that model probabilities, training objectives, decoding, and existing context can all affect repetition. They do not directly establish how period length, onset timing, and topic retention relate to one another in alternating two-agent exchanges. Extending findings from single long-text generation to multi-turn dialogue requires explicit specification of each agent's role history, generation parameters, and stopping conditions.

More recently, Ko and Geiping investigated attractor states and inter-model influence in multi-turn conversations involving multiple models and controversial topics [4]. This context makes clear that convergence of self-dialogue toward a stable form is not itself a new observation. The present study addresses narrower, operational questions: for a 0.8B model that can run on a single consumer GPU, fictional everyday scenarios in Chinese, and 200 alternating messages, can stability be directly confirmed through exact-text periods? Does the initial topic remain decodable within repetition? Can a single topic-anchoring instruction change the trajectory? After repetition is detected, can interventions on sampling or history change short-term continuations?

The data were generated locally using fixed model weights. The study involved no model fine-tuning and did not treat the two contexts as two independently trained models. The agents shared the same parameters but maintained separate sequences of message roles and inference caches, creating a reproducible feedback system with two contexts. This design helps isolate differences in prompting and history handling while also delimiting the conclusions. The analysis concerns interaction dynamics under a particular model and generation protocol; it does not directly generalize to heterogeneous model teams, tool-use tasks, or discussions with continuing human involvement.

The study makes three related contributions. First, it establishes an auditable dataset analyzed at the level of complete conversations and matched blocks, reporting exact-text periods, near-repetition confirmation times, and topic retention together. Second, it compares lexical statistics, semantic representations, supervised prediction, and unsupervised grouping on the same data, retaining findings that do not support strong prediction or stable categorization. Third, experiments branching from frozen history prefixes connect observed repetitive states to actionable interventions in sampling, context, and instructions, while assessing the engineering value of repetition detection for controlling generation costs.

2 Related Work

2.1 Text degeneration and repetition control

Decoding algorithms alter both local diversity and the longer-run distribution of generated sequences. Nucleus sampling restricts the candidate set by cumulative probability mass [1], but does not itself guarantee sustained novelty across sentences or dialogue turns. Unlikelihood training explicitly penalizes unwanted repeated candidates [2], while studies of sentence-level degeneration indicate that repetitive context can increase the tendency toward subsequent repetition [3]. Hiraoka and Inui investigated repetition-related neurons through internal model activations [5]. The present study neither measures internal neurons nor infers specific neural mechanisms. Instead, it uses auditable output sequences and controlled branch comparisons to test behavioral changes.

Repetition penalties and presence penalties must be distinguished. A repetition penalty typically rescales a token's logit according to its sign when the token has already occurred; a presence penalty can subtract a fixed amount from the logit of an already occurring token. Their scope and backend implementations can affect outcomes. The pinned model card for Qwen3.5-0.8B recommends a configuration for non-thinking text tasks that includes temperature=1.0, top_p=1.0, top_k=20, presence_penalty=2.0, and repetition_penalty=1.0 [6]. The original observational experiment used a different fixed configuration. The recommended configuration was therefore included as an additional branch condition with an explicitly defined implementation scope, to avoid conflating sampling settings with model capabilities.

2.2 Stable states in multi-turn dialogue

Evidence that performance on multi-turn tasks depends on how information is presented also underscores the need to distinguish single-turn and multi-turn evaluation [14]. The state of a multi-turn interaction includes historical text, roles, decoding randomness, and system prompts; output text is not the complete state of the dynamical system. Even a strict string period at the end of a conversation does not establish convergence of internal states to a mathematical attractor, much less prove that an indefinitely extended conversation could never leave that pattern. Existing work on multi-turn attractor states provides context for discussing stable forms [4]. Here, “periodic locking” denotes a phenomenon satisfying explicit matching criteria within a finite observation window, maintaining a distinction from semantic clustering and theoretical attractors.

The same loop structure can carry text about different topics. For example, two mutually confirming sentences about a reading room and two about balcony plants can both have period 2 while retaining different vocabulary. Topic retention and structural repetition therefore need not be mutually exclusive. Tail-topic classification and tail-semantic clustering are used to examine category information and natural groupings separately, avoiding the inference that a supervised classifier's ability to identify opening categories implies a small set of universal semantic end states.

2.3 Representations and units of evaluation

Character n-gram representations directly quantify lexical overlap, whereas dense Chinese embeddings provide an auxiliary measure of semantic similarity when vocabulary changes. This study uses a small BGE Chinese encoder [7] and retains linear classification with TF-IDF as a simple baseline. Embedding cosine similarity measures proximity in a representation space, rather than factual accuracy or the probability of retaining a topic. Statistically, messages within a conversation are highly dependent; the two prompt conditions for a given topic and replicate identifier also share a seed. Conversations are therefore the recording units, with matched pairs and replicate blocks used to account for these dependencies.

3 Materials and Methods

3.1 Experimental design and dataset composition

The original plan comprised five openings, two prompt conditions, and 100 replicates per combination, totaling 1,000 conversations. Each replicate block contained all 10 combinations in a randomly shuffled order. The unanchored and topic-anchored conditions used paired seeds within each topic and replicate identifier. Generation was stopped when the author considered the available sample size sufficient, yielding 621 complete conversations and one incomplete conversation containing 161 messages. The analysis uses a fixed snapshot of the complete conversations; the incomplete conversation was neither completed afterward nor included in the primary analysis. Stopping was not governed by a prespecified statistical power rule, so the results are interpreted as a finite-sample exploratory study.

Of the 621 conversations, the first 62 complete replicate blocks supplied 620 conversations and 310 matched pairs; the additional conversation used the park opening in the unanchored condition. Descriptive analyses use all 621 conversations, whereas paired prompt comparisons use the 310 pairs. Every other topic–condition combination contains 62 conversations, and the unanchored park condition contains 63. Each conversation comprises 200 messages, with agents A and B each speaking 100 times. Throughout the paper, "message index" refers to the combined alternating sequence, rather than the turn count of either agent individually.

Table 1. Openings and sample composition in the observational experiment. Scenario text is translated from the original Chinese. Topic labels identify five fixed openings, rather than a broader population of topics.

Opening	Initial background (English translation)	Unanchored	Anchored
Balcony	The following is a fictional scenario: an east-facing balcony has three pots of mint, one of which has a blue label.	62	62
Reading room	The following is a fictional scenario: a reading room has four rows of bookshelves and a green chair by the window.	62	62
Park	The following is a fictional scenario: a small park has two paths, a wooden table in the center, and osmanthus planted nearby.	63	62
Town	The following is a fictional scenario: a town named Qinglan has seven bridges, and its residents enjoy exchanging used books on weekends.	62	62
Notebook	The following is a fictional scenario: a person has prepared a blank notebook with the number 37 written on the first page, but has not decided what to record.	62	62

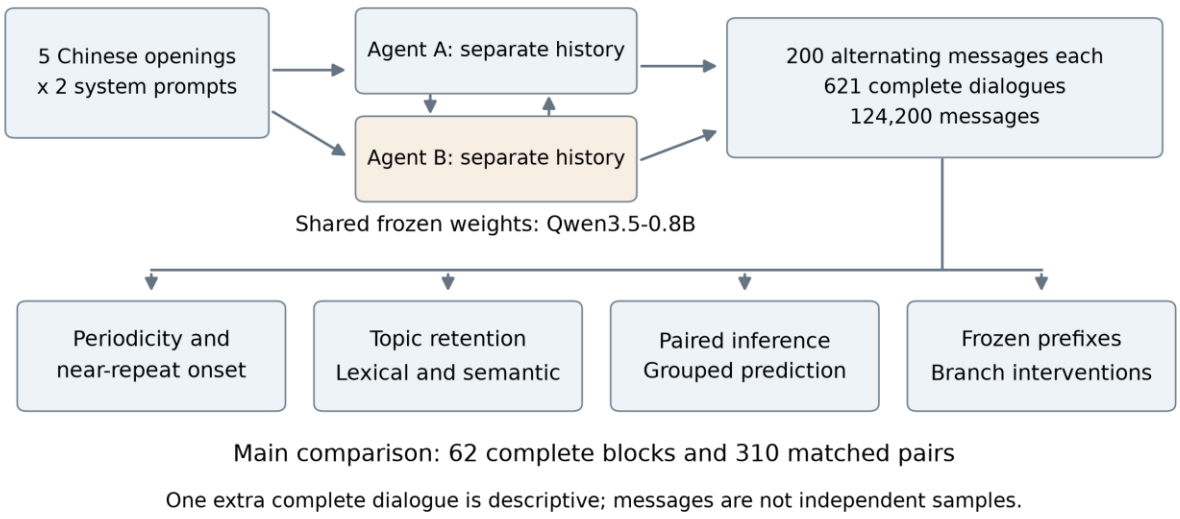


Figure 1. Experimental and analytical workflow. The agents share frozen weights but maintain separate role histories and inference caches. Observational data, paired analyses, and branch controls from frozen histories are stored separately.

3.2 Model and generation protocol

Qwen/Qwen3.5-0.8B was loaded locally. The full weight revision is provided in Supplementary Section S1, and the cited model card points to the same pinned version [6]. The runtime model class was Qwen3_5ForCausalLM, with 752,393,024 loaded parameters; the checkpoint reported no missing,

unexpected, or shape-mismatched keys. Inference used an NVIDIA GeForce RTX 3070 Ti Laptop GPU, BF16 precision, and SDPA attention. The original environment comprised PyTorch 2.13.0+cu130 and Transformers 5.15.0. Model weights remained frozen throughout all observational runs and continuations.

The original sampling parameters were fixed at temperature=0.8, top_p=0.95, top_k=20, and repetition_penalty=1.0, with a maximum of 128 new tokens per reply and thinking disabled. Both agents received the same initial background. A generated first; B initially received the background together with A's first message, after which the agents alternately received one another's outputs. The system prompt in the unanchored condition requested natural responses, continued interaction, and replies of 1 to 3 sentences without outlines. The topic-anchored condition added only the following instruction (English translation): “Please remain focused on the opening information throughout the exchange.” Full prompt wording is provided in Supplementary Section S1.

The generator constructed inputs using an append-only protocol based on fragments of the official chat template. It retained the original generated output tokens and the empty think marker used in non-thinking mode, updating each agent's cache separately. This protocol differs from re-rendering the entire official template from plain text on every turn. The history policy was full_history: if the input plus reserved output exceeded 32,768 tokens, generation stopped with an error rather than silently truncating the context. No context resets occurred in the complete conversations; the maximum input length was 25,979 tokens. Hitting the per-message output cap and ending naturally were recorded separately. Neither a farewell nor textual repetition terminated the original conversations.

Short pilot runs before the main batch motivated a uniform instruction to keep replies brief; these pilots were excluded from the study dataset. Each main-run message record contains the text, output token IDs, sampling seed, input digest, input and output lengths, termination reason, and elapsed time. Subsequent analyses use these saved records, and SHA-256 hashes were computed for the original files. Seeds and execution-code digests support implementation traceability; an identical seed does not imply that different prompts or interventions must generate the same sequence.

3.3 Repetition and periodicity measures

Let $y_{i1},\ldots,y_{iT}$ denote the sequence of raw message strings in conversation i, with T=200. Strict tail periodicity is assessed by exact comparison of the original strings. For each $p\in\{1,\ldots,10\}$, a period is present if every one of the final w messages equals the message p positions earlier; the smallest qualifying p is reported. The primary window is w=40, with w=20 and w=80 used in sensitivity analyses. Each message being tested lies within the tail window, although its reference message may precede that window.

$$Q_i(p,w) = 1\{\, y_{it} = y_{i,t-p} \text{ for all } t=T-w+1,\ldots,T \,\}$$

For the other lexical measures, text is first NFKC-normalized, lowercased, and stripped of whitespace while retaining punctuation. Same-speaker exact repetition occurs when normalized message t equals normalized message t−2. The descriptive overall proportion uses 200 messages as its denominator, counting the first two as non-repetitions. Comparisons with randomly shuffled sequences use only the 198 comparable positions. These denominators are distinguished explicitly in the results. Historical repetition instead indicates whether a message has appeared anywhere earlier in the same conversation, irrespective of speaker.

Near-repetition is measured using the character-bigram frequency vector v_t of normalized text. The cosine similarity s_t compares successive utterances by the same speaker. A message is classified as a near-repetition when $s_t>0.9$. The onset τ is the first position starting a run of m=20 consecutive near-repetitions; the online confirmation time is $c=\tau+19$. Conversations without confirmation by message

200 receive a value of 201 in the bounded timing score; this coding is not an observed event time. Similarity thresholds of 0.8, 0.9, and 0.95 are examined in sensitivity analyses.

$$s_t = (v_t \cdot v_{t-2}) / (\|v_t\| \, \|v_{t-2}\|), \; t \geq 3$$

Lexical novelty is the number of distinct character bigrams in the current message that have not previously appeared in the conversation, divided by the total number of distinct bigrams in that message. This measure captures new textual fragments relative to the conversation's own history. It does not evaluate factual correctness or treat every new fragment as useful information. Complete definitions, handling of empty sets, and implementation details are provided in Supplementary Section S2.

The expansion of cumulative history itself tends to reduce lexical novelty, so this measure serves only as an auxiliary description of trajectories. The primary evidence for periodic locking comes from exact-text periods and same-speaker repetition, rather than from declining novelty alone.

3.4 Topic retention and semantic representations

The lexical topic measure tests for opening-related keywords in a fixed dictionary. For the balcony opening, these include terms meaning “balcony,” “mint,” “blue,” “label,” and “east-facing” (English translations of the Chinese keywords). The keyword dictionary and regular expressions for fact-related terms are preserved with the analysis code. Fact-term measures count whether relevant expressions are mentioned without determining whether their relations, quantities, or negations are correct; they are therefore described as term coverage rather than factual retention rates.

Semantic encoding uses BAAI/bge-small-zh-v1.5; its complete pinned revision is listed in Supplementary Section S1. Normalized CLS representations are used to compute cosine similarity between each message and its opening. The cache contains 27,870 distinct encoded inputs, including the five openings; none was truncated at the 512-token input limit. Representations of the first two and final two messages in each conversation are separately averaged and renormalized for topic-decoding comparisons with equal message counts. Exploratory clustering uses the mean representation of distinct replies in the final 40 messages, preventing the frequency of a repeated string from dominating the vector.

Topic decoding compares TF-IDF with logistic regression, TF-IDF with a linear SVM, and BGE embeddings with logistic regression. TF-IDF uses character 2- to 4-grams, min_df=2, a maximum of 18,000 features, and sublinear_tf. Logistic regression uses C=5 and the linear SVM uses C=1. Five-fold GroupKFold splits by replicate block, keeping all topics and paired conditions from the same block in the same fold. Vocabulary and classifiers are fitted only on training folds. Accuracy measures discrimination among these five fixed openings; uniform random guessing provides a 20% reference.

3.5 Prompt comparisons and statistical inference

Prompt differences are first calculated within pairs matched on topic and seed, then summarized as the topic-anchored condition minus the unanchored condition. Primary intervals use 20,000 bootstrap resamples of the 62 complete replicate blocks, retaining all five within-block topic differences together in each resample. Matched-pair bootstrapping and Wilcoxon tests serve as sensitivity analyses. The prompt comparison includes 14 exploratory measures; the Benjamini–Hochberg procedure controls the false discovery rate within this comparison family [8]. The supplementary audit also includes permutation tests that flip signs for entire blocks, providing a sensitivity analysis that preserves within-block dependence.

For periodicity prevalence, the analysis reports full-sample counts, Wilson intervals, and sensitivity intervals obtained by resampling the available replicate blocks. Timing scores for the five fixed

openings are compared by first averaging the two prompt conditions within each replicate block, followed by a Friedman test and adjusted pairwise comparisons. All statistical units are conversations, matched pairs, or replicate blocks: the 124,200 messages are not treated as 124,200 independent observations. The study was not preregistered, and its principal questions and measures constitute exploratory analyses of existing trajectories.

Resampling is used to estimate finite-sample uncertainty [9]. The block bootstrap for prevalence across all 621 conversations samples the 63 available replicate clusters, the last of which contains only one conversation. Results for the balanced 620-conversation sample with 62 complete blocks are also reported. The 14 measures are correlated; BH adjustment is used for exploratory reporting and is not interpreted as an unconditional guarantee under arbitrary dependence. Primary paired intervals use the supplementary computation of 20,000 block resamples, and permutation tests use 50,000 whole-block sign flips.

3.6 Early prediction and unsupervised comparisons

Early prediction uses only statistical features from the first 20 messages, including lexical similarity, repetition, novelty, length, topic terms, closing expressions, and semantic measures. One target is whether a run of 20 consecutive near-repetitions is confirmed at or before message 80; the other is the bounded confirmation time $\min(\tau+19,201)$. Every target confirmation position occurs after message 20, so the feature window does not include the confirmation event being predicted.

In 46 conversations, however, the near-repetitive segment already begins at or before message 20, so the early features may contain signs of an emerging loop. This task therefore predicts later confirmation; it does not guarantee a warning before any indication of degeneration. Further comparisons on the balanced 620-conversation sample use a single lag-2 similarity feature, topic and condition identities only, training-fold class proportions, and the full set of 22 features. Sensitivity to splitting is assessed with 20 random grouped-fold assignments. Structured validation is used to preserve within-group dependence [10], and all models are implemented in scikit-learn [13].

Binary classifiers include standardized logistic regression ($C=0.1$), a random forest (300 trees, maximum depth 5, and a minimum of 10 samples per leaf), and histogram gradient boosting (100 iterations, at most 7 leaf nodes, and L2=10). Timing-score models include the training-fold median, standardized ridge regression ($\alpha=20$), a random forest, and gradient boosting. Grouping follows the topic-decoding analysis, and performance is calculated from pooled out-of-fold predictions. An additional split uses replicate blocks 0 to 44 for training and 45 to 62 for testing in execution order. This is a chronological holdout from the same generation batch, rather than an independently collected external validation set.

Tail-semantic grouping compares KMeans with k=3 to 10, average-linkage hierarchical clustering, and DBSCAN. Cosine silhouette scores [11] assist the choice of k for KMeans, and adjusted Rand indices [12] quantify agreement across random initializations and algorithms. DBSCAN is evaluated using Euclidean distances between normalized vectors, eps=0.35, 0.45, 0.55, and 0.65, and min_samples=5. NMF extracts six lexical factors from character TF-IDF. These comparisons test the robustness of grouping; visual separation in a two-dimensional projection is not used to name psychological states or claim universal categories.

3.7 Stopping rules and execution audit

Stopping rules scan the actual records message by message and trigger after m consecutive same-speaker similarities greater than 0.9, with m=3, 5, 10, or 20. The triggering message counts as generated; subsequent messages, output tokens, and recorded elapsed time count as avoidable costs. The analysis then checks whether any later message has similarity at or below the threshold. This

jointly reports cost reduction and the risk of prematurely truncating subsequent change. The rules detect only lag-2 near-repetition and may miss patterns with periods of 4 or more.

The execution audit checks message indices, alternating roles, sampling seeds, token counts, completion markers, context resets, and file hashes. Two generation positions near message 180 are also sampled from each of four categories: period 1, period 2, period 4, and no strict period. After clearing the cache, the entire history is reconstructed under the same append-only token protocol and generation is replayed. This checks whether cache reuse affects the sampled outputs; it does not establish correctness of the entire template protocol or every token.

3.8 Interventions on frozen prefixes

To test mutability after repetition had formed, the supplementary experiment fixed the selection rule, prefixes, seeds, conditions, and measures before generation, saving the protocol and its hash. Within each combination of the five openings and two original conditions, the conversation with the lowest identifier satisfying same-speaker exact repetition throughout messages 61 to 80 was selected. Each branch began after message 80, retained the original model weights and both initial backgrounds, reconstructed its state independently, and generated 12 further messages under each of two new sampling seeds. This design evaluates short-term responses of selected repetitive prefixes, rather than the average intervention effect across all conversations.

Table 2. Branch conditions after freezing the history. Model weights are identical across all branches. The recommended sampling condition jointly changes several decoding parameters.

Condition	History and instruction	Sampling change
Original configuration	Full history; original prompt	None
Repetition penalty	Full history; original prompt	repetition_penalty=1.15
Short history	Retain initial input and most recent 8 messages	None
Anti-repetition instruction	Full history; add a request for new details	None
Recommended sampling	Full history; original prompt	Temperature 1.0; top_p 1.0; presence penalty 2.0

The five conditions comprise the original-sampling, full-history baseline; repetition_penalty=1.15; a short history retaining the opening and the most recent 8 messages; an added anti-repetition system instruction; and the model-card-recommended sampling configuration. In the recommended condition, an explicit logits processor implements the presence penalty: during each new reply, it subtracts 2.0 from the logits of tokens already generated in that reply only, without applying this subtraction to tokens occurring solely in historical prompts. Conditions retaining the original output tokens use the same token construction as the original run; conditions that modify history reconstruct the corresponding inputs.

The primary endpoint is the fraction of all 12 continuation messages with cosine similarity at or below 0.9 to the same speaker's preceding message. Secondary measures include sustained low repetition across the final six messages, departure from the original tail-text set, lexical novelty, topic-keyword coverage, output-cap hits, and elapsed time. Short-term textual change is not equivalent to factual correctness or restored effective collaboration. Statistical comparisons preserve pairing by prefix and new sampling seed; prefixes sharing an original seed are combined into a single resampling unit for

interval estimation. The complete protocol, seeds, instructions, and raw continuations are provided in Supplementary Section S5.

4 Results

4.1 Data integrity and tail periodicity

All 621 complete conversations passed structural checks. They generated 5,078,339 output tokens, and recorded per-message generation times summed to 156.11 hours. This total excludes model downloading, environment setup, and offline analysis. Under the primary window, 610 conversations exhibited a strict period, a prevalence of 98.23%. There were 299 conversations with period 1, 280 with period 2, and 26, 3, and 2 with periods 4, 6, and 8, respectively. The remaining 11 had no detected exact-text tail period of at most 10. Normalization shifted only one conversation from period 2 to period 1, leaving the overall presence-of-period count unchanged.

The Wilson 95% interval for the primary periodicity prevalence was 96.86% to 99.01%; the interval obtained by resampling available replicate blocks was 97.10% to 99.19%. Restricting the analysis to the balanced 620-conversation sample gave 609/620, again 98.23%. Repeating detection using the saved output token IDs still identified 610 conversations with a period. In 12 of these, visible strings had period 1 but token sequences had period 2. Identical text can therefore arise from different tokenizations, and string periodicity should not be equated with a fixed point of the complete internal state.

Table 3. Distribution of tail periods in raw strings. The shortest period is sought over 1 to 10, using the final 40 messages as the assessment window.

Shortest period	Conversations	Share of 621 conversations
1	299	48.15%
2	280	45.09%
4	26	4.19%
6	3	0.48%
8	2	0.32%
Not detected	11	1.77%

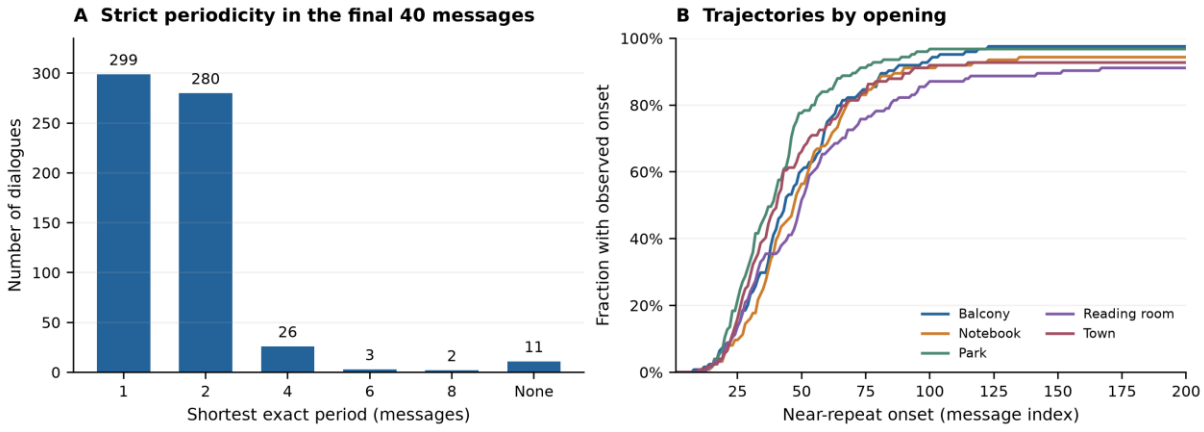


Figure 2. Distribution of tail periods and near-repetition onset positions. The curve describes the distribution of the first positions satisfying consecutive near-repetition within the finite window; each onset is confirmed by the run of 20 messages beginning at that onset.

Changing the tail window to 20 or 80 messages yielded 612 and 599 periodic conversations, respectively. Of the 525 conversations that never reached the per-message output cap, 518 still met the primary periodicity criterion (98.67%). The output cap is therefore not a sufficient explanation for the widespread periodicity in these data. Input digests and generated tokens at all 8 fresh-cache replay positions matched the original records, supporting cache-path reproducibility at those sampled positions, while leaving conclusions conditional on the original input-template protocol.

4.2 Growing repetition coexists with topic retention

Same-speaker exact repetition averaged 1.51% in the first 20 messages and reached 94.60% in the final 20. Repetition of text appearing at any earlier position accounted for 77.47% of all messages. Each conversation contained a mean of 45.06 distinct normalized replies and a median of 40. Lexical novelty fell sharply toward the end: the fraction of new character bigrams declined from 66.62% in the first 20 messages to approximately 0.006% in the final 20. Meanwhile, mean raw-text length increased from 46.71 to 70.45 characters per message, showing that longer replies can still consist predominantly of repeated text.

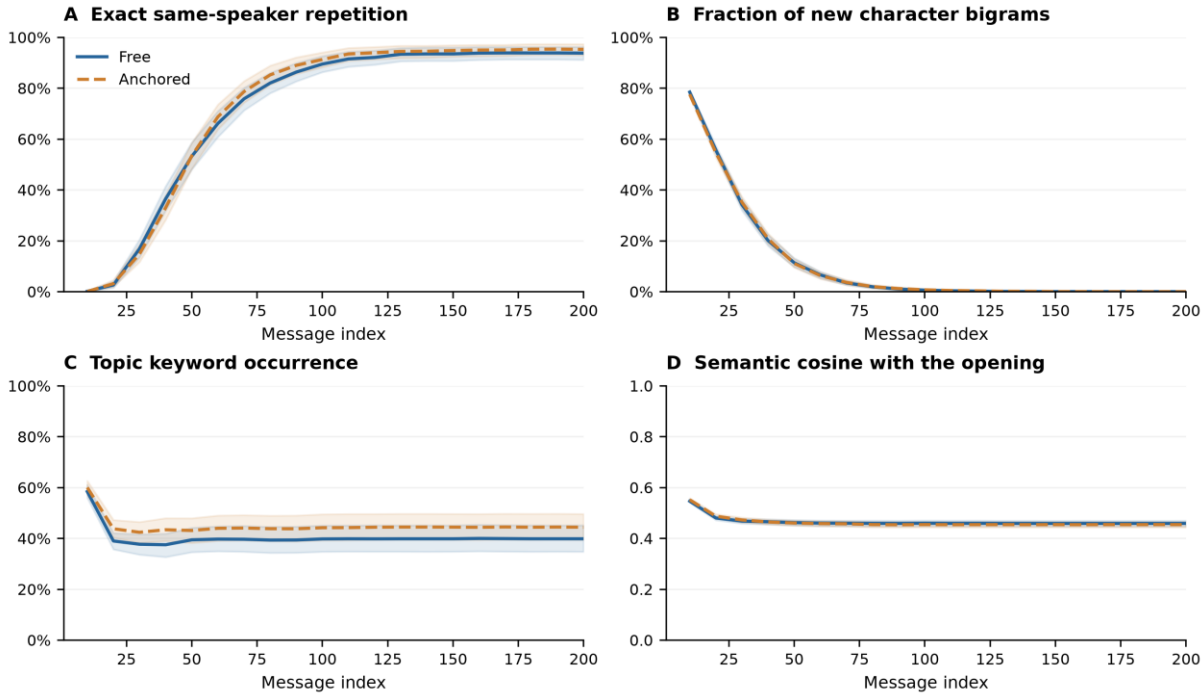


Figure 3. Repetition, novelty, and topic retention by message index. Messages are summarized in bins of 10. Curves weight conversations equally; shaded areas show descriptive 95% bootstrap intervals. Semantic cosine similarity does not represent factual accuracy.

Information related to the initial topic remained in the tail. The topic-keyword occurrence rate declined from 50.31% in the first 20 messages to 42.16% in the final 20, while mean BGE cosine similarity between messages and the initial background decreased from 0.517 to 0.457. In comparisons using equal message counts, opening-category decoding accuracy exceeded 99% for the first two messages under all three algorithms and remained between 64.09% and 71.82% for the final two. BGE logistic regression achieved the highest tail accuracy. This result establishes only that the five fixed openings remain distinguishable in tail representations; it does not demonstrate complete retention of quantities, directions, or factual relationships.

Table 4. Opening-category decoding with equal message counts. Models are trained and evaluated separately on the first two and final two messages, using five folds grouped by replicate block.

Method	First two messages	Final two messages
TF-IDF + logistic regression	99.19%	64.57%
TF-IDF + linear SVM	99.36%	64.09%
BGE + logistic regression	99.84%	71.82%

4.3 Paired effects of the topic-anchoring instruction

Across the 310 matched pairs, overall same-speaker exact repetition was 72.12% in the unanchored condition and 73.25% in the topic-anchored condition. The anchored-minus-unanchored difference was 1.13 percentage points, with a block-bootstrap 95% interval of −2.21 to 4.49 percentage points. Topic-keyword occurrence in the final 20 messages was 39.68% and 44.45%, respectively, giving a difference of 4.77 percentage points and an interval of −2.26 to 11.81 percentage points. Both intervals included zero.

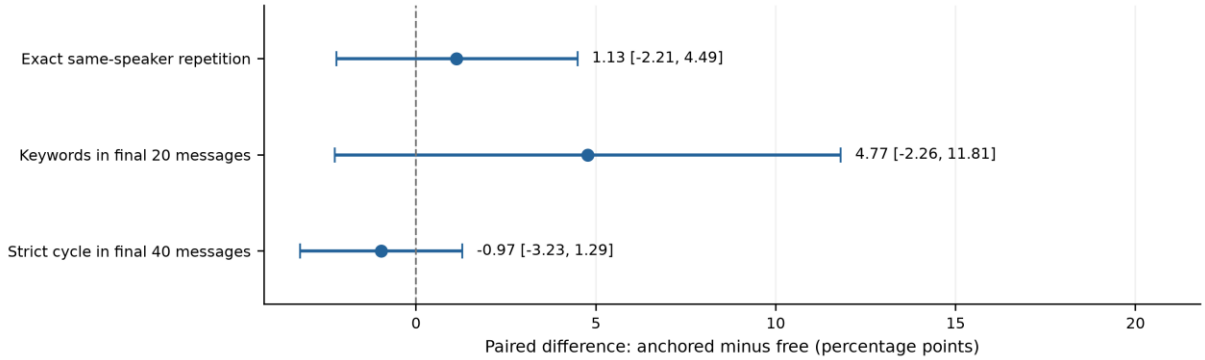


Figure 4. Paired differences between the topic-anchored and unanchored conditions. Points indicate mean differences and horizontal lines indicate 95% intervals from resampling complete replicate blocks. All three proportional measures are expressed in percentage points.

With whole-block sign-flip tests, the smallest BH-adjusted q value among the 14 exploratory prompt comparisons was 0.5997; none reached the 0.05 level. For same-speaker exact repetition, the permutation p value was 0.5152 and q=0.7484. For tail-topic keyword occurrence, p=0.1949 and q=0.7484. These data do not show a stable improvement from this topic-anchoring sentence, but neither do they establish a true effect of exactly zero. In particular, the interval for topic-term coverage still permits some improvement, so non-significance cannot be interpreted as equivalence.

4.4 Timing and differences between openings

A run of 20 consecutive same-speaker near-repetitions occurred within 200 messages in 587 conversations. Among conversations with the event, the median onset was message 42 and the median confirmation was message 61. With unconfirmed conversations coded as 201, the median onset score across the entire sample was 43. Similarity thresholds of 0.8, 0.9, and 0.95 yielded 591, 587, and 586 events, respectively, indicating that the main finding was relatively stable across this threshold range.

Of the conversations with confirmed near-repetition, 542 had at least 100 subsequent observed messages; none showed departure from near-repetition. This denominator makes the available follow-up explicit rather than treating conversations confirmed at different times as having equal follow-up lengths.

Across all 587 confirmed conversations, no subsequent reply had similarity at or below 0.9 through message 200. This describes persistence within the observation window, not irreversibility over an infinite horizon. A single change point fitted to lexical novelty had a median position of message 29 and a Spearman correlation of approximately 0.765 with the near-repetition onset score. Because change points were fitted using complete trajectories, they serve only as retrospective trajectory descriptions.

After averaging the two prompt conditions within each block, bounded onset scores differed among the five fixed openings (Friedman p=0.0050). The mean park-minus-reading-room difference was approximately −19.70 messages, with a block-bootstrap 95% interval of approximately −30.75 to −8.48 messages and q=0.0087 after adjustment across ten pairwise comparisons. Because each topic is represented by only one fixed opening text, this difference may also reflect specific wording or factual configurations; it does not establish that park-related topics are generally more prone to repetition.

4.5 Limits of prediction and clustering

Using features from the first 20 messages to predict near-repetition confirmation at or before message 80, out-of-fold AUCs were 0.720 for logistic regression, 0.697 for the random forest, and 0.688 for gradient boosting. Logistic regression performed best, indicating no consistent advantage from the more complex nonlinear algorithms at this sample size and with these features. The audit added majority-class and probability baselines, block-bootstrap intervals, and label-permutation checks to distinguish useful signal from class proportions and chance fitting.

In the balanced sample of 620 conversations, logistic regression with all 22 features achieved an AUC of 0.7202 (block interval 0.6746 to 0.7651), compared with 0.6977 (0.6555 to 0.7393) for lag-2 similarity alone. The difference of 0.0225 had an interval of −0.0100 to 0.0542, providing no evidence of a stable advantage for the full feature set. At a threshold of 0.5, the full model's balanced accuracy was only 0.5455. Its ordinary accuracy was 75.00%, compared with 74.35% for the majority-class baseline. Across 200 label permutations preserving block structure and refitting the model, the maximum AUC was 0.5889 and the permutation p value was 0.0050. This supports the presence of a ranking signal but does not justify describing the model as a high-accuracy early-warning system.

Across 20 random grouped-fold assignments, the full logistic model had a mean AUC of 0.7213, ranging from 0.7072 to 0.7332. In the later-run classification test, trained on blocks 0 to 44 and tested on the 170 conversations in blocks 45 to 61, the full model's AUC fell to 0.6402, while lag-2 similarity alone achieved 0.6739. This denominator differs from the 171-conversation timing-regression test, which includes the additional conversation. Both are retrospective sensitivity analyses of existing data, rather than independent evaluations on new data collected after freezing the analysis.

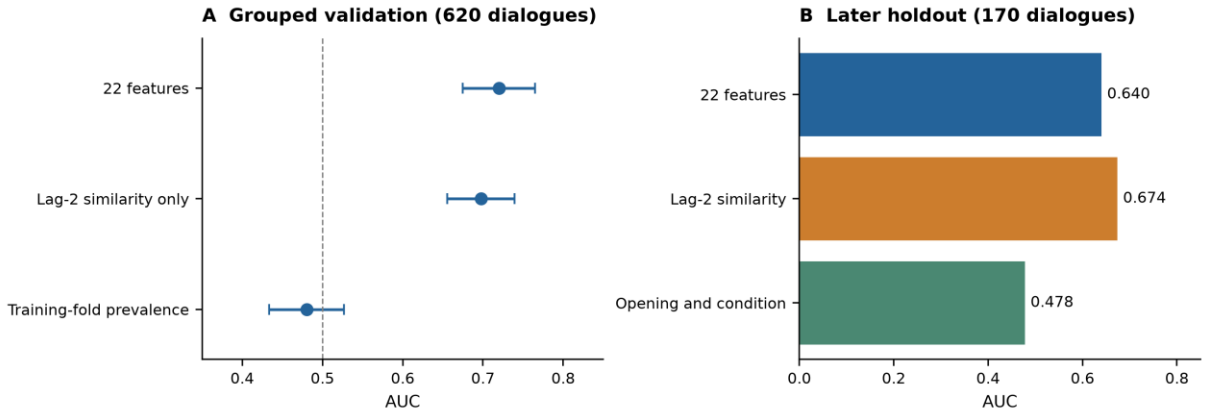


Figure 5. Simple baselines and chronological sensitivity of early warning. Left: grouped out-of-fold AUCs on the same sample, with block-bootstrap 95% intervals. Right: retrospective later-run holdout performance. Both panels use the balanced sample of 620 conversations.

Improvements in confirmation-time prediction were small. Mean absolute errors were 23.90 messages for the training-fold median baseline, 23.45 for ridge regression, 22.94 for the random forest, and 24.62 for gradient boosting. In the 171-conversation test split by execution order, ridge regression had an error of 23.04 messages, exceeding the median baseline's 22.53. The available features therefore offer limited warning signals but do not yet provide stable, precise timing predictions across phases of the run.

Tail clustering did not yield a robust unified categorization. Across k=3 to 10, KMeans achieved a maximum cosine silhouette score of only 0.0995 at k=7; adjusted Rand indices across initializations ranged approximately from 0.468 to 0.803. Agreement between average-linkage hierarchical clustering and KMeans at the same k was close to zero. DBSCAN classified all samples as noise at the three smaller eps values and still labeled 97.9% as noise at eps=0.65. NMF extracted lexical factors associated with terms such as “warmth,” “next time,” and the number 37 (English translations where applicable), but these factors do not constitute validated, stable end-state categories.

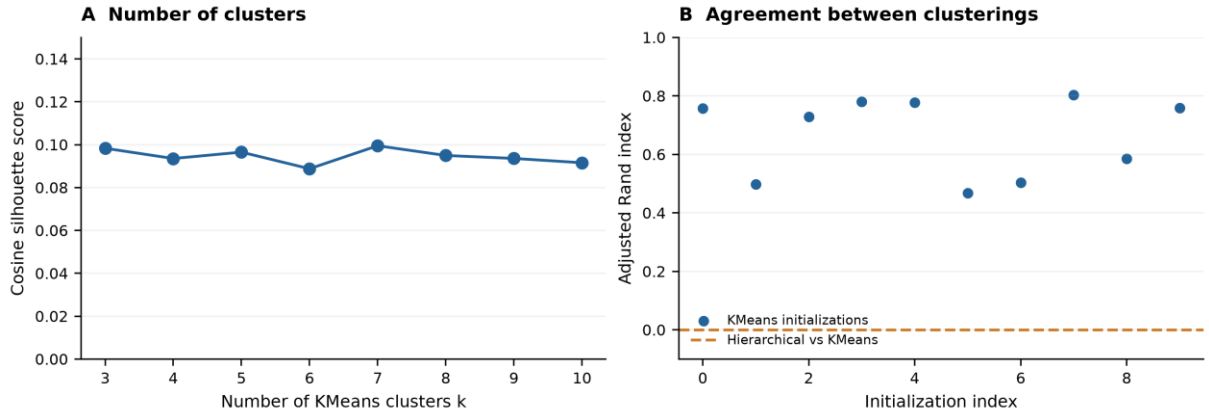


Figure 6. Algorithmic sensitivity of tail-semantic grouping. Low silhouette scores and disagreement between algorithms limit interpretations of a fixed number of categories. DBSCAN findings apply only to the parameter grid tested.

4.6 Retrospective cost evaluation of early stopping

Shorter runs of near-repetition before triggering would retrospectively remove more messages but are more likely to truncate subsequent change. For m=3, 5, and 10, the numbers of conversations with a later non-near-repetitive message were 29, 10, and 1, respectively; for m=20, the number was 0. The 20-message rule triggered in 587 conversations and would have avoided 78,953 messages, or 63.57% of all complete-conversation messages, together with 3,395,705 output tokens. The corresponding recorded generation time was 482,021 seconds, approximately 85.77% of total generation time.

Table 5. Retrospective comparison of stopping rules. Savings are computed from suffixes of already generated trajectories; they do not indicate that equivalent time savings have been observed in a rerun.

Consecutive messages	Conversations triggered	Conversations with later change	Messages avoided	Time avoided
3	588	29	71.96%	89.65%
5	588	10	70.87%	89.13%
10	587	1	68.31%	87.97%
20	587	0	63.57%	85.77%

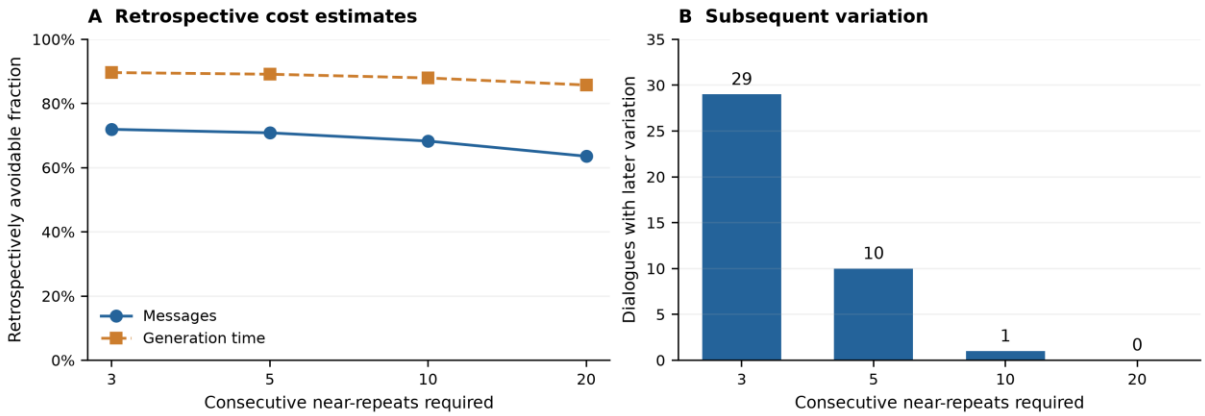


Figure 7. Costs and subsequent change under consecutive-near-repetition stopping rules. Longer confirmation windows reduce the number of instances in which subsequent change would be truncated in this sample, while retaining more computation.

The proportion of avoidable time exceeds the proportion of avoidable messages, consistent with records showing longer contexts and higher inference costs later in the conversations. This rule terminates trajectories that already show little change; it does not improve answer quality. All thresholds were evaluated on the same existing dataset, so retrospective performance cannot substitute for prospective validation on new conversations. Nor can the zero-recovery observation from unintervened trajectories be transferred directly to branches receiving interventions.

4.7 Results of branches from frozen prefixes

All 100 branches from the ten frozen history prefixes were completed, generating 1,200 additional messages with no missing branches. Each condition comprised 20 branches and 240 messages; the two continuation seeds were first averaged within each prefix. The statistical blocks were eight original-seed clusters. The original configuration, repetition penalty of 1.15, short history, and anti-repetition instruction produced no same-speaker changes with character-bigram cosine similarity at or below 0.9. Under the recommended sampling configuration, 18/240 messages met the change criterion, giving a mean fraction of 7.50%. The 4/20 branches with at least one such change came from the park and town openings.

Table 6. Observed continuations from frozen prefixes. Each condition has 20 branches. The change fraction uses all 12 messages; sustained change and strict periodicity use the final 6. Periodicity in this short window tests p≤4 and is distinct from the definition using the final 40 messages in the observational experiment.

Condition	12-message change fraction	At least one change	All final 6 changed	Strict period in final 6	Messages at length cap
Original configuration	0.00%	0/20	0/20	20/20	12/240
Repetition penalty 1.15	0.00%	0/20	0/20	19/20	12/240
Short history	0.00%	0/20	0/20	20/20	12/240
Anti-repetition instruction	0.00%	0/20	0/20	20/20	12/240
Recommended sampling	7.50%	4/20	1/20	15/20	7/240

The primary-endpoint difference between the recommended and original configurations was 7.50 percentage points, with an interval of 1.14 to 17.08 percentage points from 10,000 original-seed-cluster bootstrap resamples. However, the exact sign-flip test over eight clusters yielded p=0.25, and BH adjustment across the four primary comparisons gave q=1.00; the other three comparisons all had

p=q=1.00. The nonparametric bootstrap interval is not an inversion of the exact test. A small number of nonzero differences with the same sign can produce an empirical interval excluding zero while remaining insufficient under a discrete test with few clusters. The difference therefore supports local mutability in the selected prefixes, not statistical evidence of a stable intervention effect. The empirical intervals for the other three interventions collapse to [0,0], reflecting the absence of threshold-crossing change in this sample rather than population-level equivalence or proof that escape is impossible.

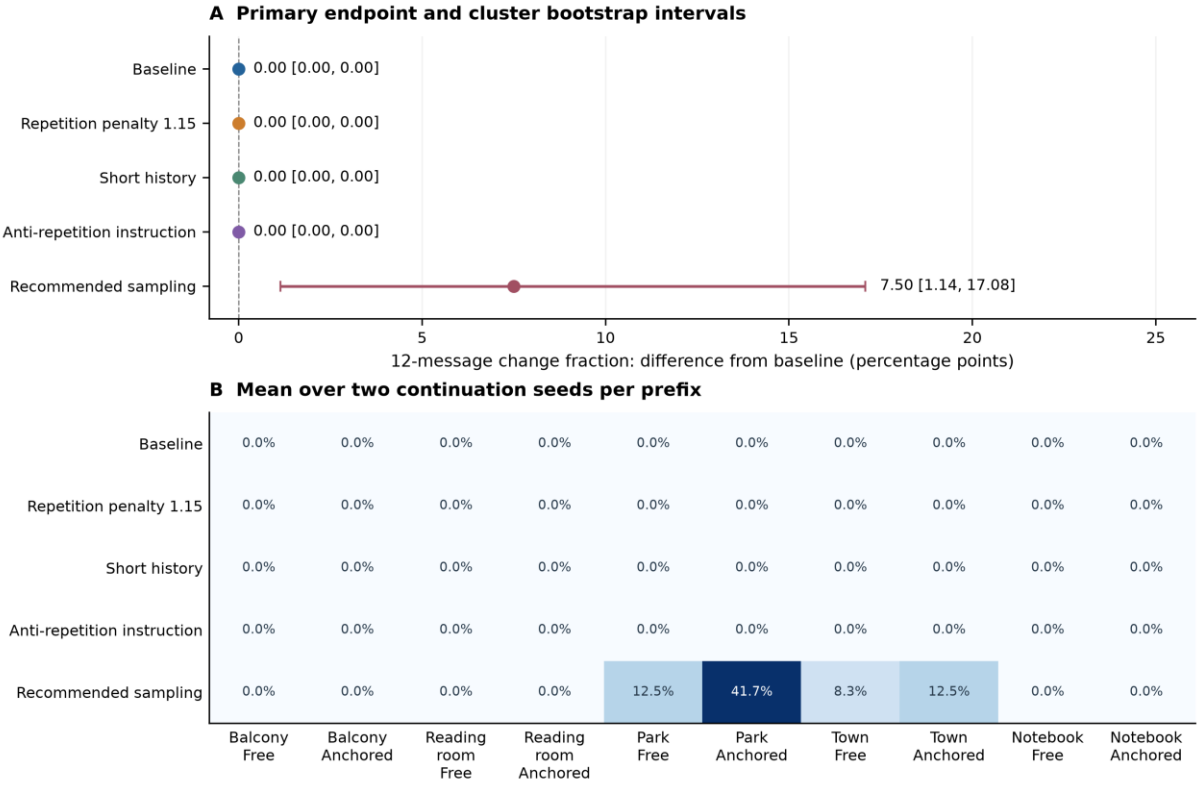


Figure 8. Overall and prefix-level branch-intervention results. Top: differences from the original configuration in the primary endpoint, with 95% empirical bootstrap intervals. Bottom: means across two new seeds within each history prefix. The recommended configuration has an exact cluster-test p=0.25 and primary-comparison BH q=1.00; an interval excluding zero does not replace exact inference with few clusters. The original configuration is the zero-difference reference; degenerate intervals for other conditions do not establish population-level ineffectiveness.

Short-term change did not generally eliminate repetition. Under the recommended configuration, 15/20 branches still exhibited a strict short period in the final six messages, and only 1/20 showed same-speaker change in every one of those final six messages. In the repetition-penalty condition, exact self-repetition fell from 100% under the original configuration to 98.33%, but the changes did not cross the primary similarity threshold, and one new short cycle differed from the original tail. Exact-text difference, departure from near-repetition, and transition into another loop are therefore distinct outcomes; an arbitrary wording change cannot stand in for recovery.

In branch_050, the only branch with sustained change across the final six messages, messages 8, 10, and 12 nevertheless copied the other speaker's messages 7, 9, and 11 verbatim, respectively. A post hoc measure of exact copying between consecutive speakers was 43.75% under the recommended configuration and 50.00% under the original configuration. These descriptions show that greater same-speaker change can coexist with cross-speaker repetition; the experiment did not count this textual change as improved factual accuracy or task progress. Using the original keyword dictionary, hit rates were 60.00% under the original configuration and 62.50% under the recommended configuration; after restricting “37” to a standalone number, the rates were 50.00% and 52.50%, respectively. This again illustrates the distinction between dictionary-based measurement and factual memory.

Original prefixes, role histories, and paired sampling seeds were independently checked for all conditions. Five additional single-message observational replays matched the output tokens of their corresponding branches exactly, and recorded the actual sampling configurations, repetition-penalty processors, and scope of the presence penalty. These execution checks are excluded from the statistical sample of 100 branches. The results concern responses during the twelve messages following already locked histories; they cannot estimate prevention effects from using these conditions starting at the first message.

4.8 Cases without strict periodicity

The 11 conversations failing the primary exact-text periodicity criterion did not all exhibit sustained effective communication. Inspection of samples from their later segments revealed template variations, increasing numbers, and delayed repetition. In episode_00007, for example, A continued proposing different numbers while B repeatedly produced division-style replies. Message 199 raised the relationship between 699 and 34; message 200 stated "699 divided by 34 = 20.617..." (English translation), whereas the actual quotient is approximately 20.5588. Of the 18 decimal-division statements extractable in this case using a fixed regular expression, only 4 were within an error tolerance of 0.005. This is a check of a single post hoc selected case, not a mathematical-reasoning benchmark.

Another conversation repeatedly changed the proposed afternoon meeting time, preserving textual differences without clear task progress. Replacing Arabic numerals with a uniform placeholder increased the number of conversations satisfying tail periodicity from 610 to 612, showing that numerical variation can conceal some template repetition. Conversely, the sets of distinct tail texts differed across all 621 conversations, suggesting that the commonality lies primarily in periodic structure rather than in every conversation eventually producing the same set of sentences.

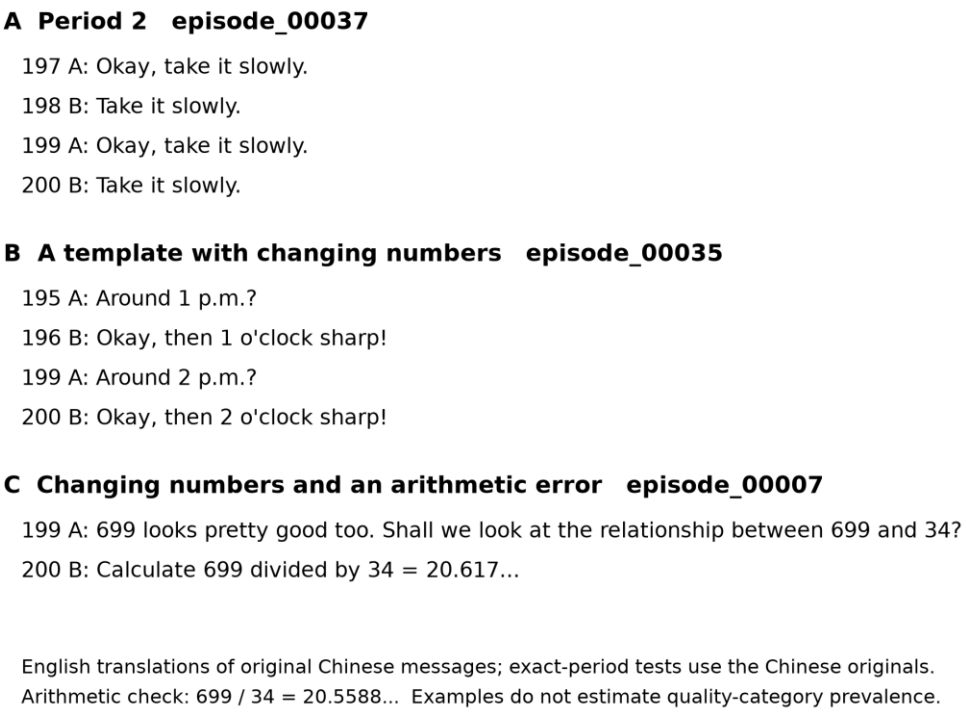


Figure 9. Representative tail excerpts, presented as English translations of the saved Chinese conversations. The examples illustrate raw-string periodicity, minor variation, and numerical templates. Their selection cannot establish the prevalence of unannotated quality categories; all repetition measures were calculated on the original Chinese strings.

5 Discussion

5.1 Separating periodic structure from topic information

The most stable observation is the coexistence of little structural variation and retained topic information. Exact-text periodicity is established directly by string matching, independently of an encoder or clustering algorithm. Tail-topic classification, in turn, shows that different openings leave identifiable features in repetitive text. Together, these findings support treating “what is being discussed” and “whether the exchange continues to add distinguishable expression” as separate evaluation dimensions. Improving topic-relevance scores alone may fail to reveal conversational stagnation, while increasing lexical diversity alone may encourage numerical or template variations without improving content.

Decodability of the tail does not imply complete memory retention. A model may repeatedly mention osmanthus, bookshelves, or the number 37 while omitting or altering initial quantities and relationships. Neither keyword coverage nor embedding cosine similarity resolves this issue. For collaborative systems with explicit goals, externally verifiable task progress and factual correctness should be measured as separate outcomes. The present free-form conversations have no common task answer, so unannotated subjective quality was not interpreted as quantitative improvement.

The granularity of keyword matching also affects the interpretation of retention. For example, the string “37” in the original notebook dictionary also matches “1937.” This indicates only that a character fragment occurs; it does not establish that the model correctly remembers the initial number on the notebook's first page. The original measure retains the established dictionary, and the supplementary material separately reports a post hoc sensitivity analysis imposing numerical boundaries. Topic classification likewise tests only the distinguishability of the fixed opening categories and is not equivalent to complete factual memory.

This distinction is directly observable in episode_00052, which repeatedly states “On December 27, 1937, the British Parliament passed Magna Carta” (English translation). This conflicts with the UK Parliament's account of its first issuance in 1215 [15]. Changing only the notebook dictionary's “37” match to require no adjacent digit reduced the topic-hit rate across the final 20 messages of the 124 notebook conversations from 43.35% to 42.14%, removing 30 hits; it removed 261 hits across their full trajectories. This check did not change the original topic measure or primary tests. Its purpose is to identify measurement error in string-based retention, not to estimate factual-error prevalence across all conversations.

5.2 The roles of decoding and history feedback

The original observations used a fixed temperature below the model-card recommendation and restricted nucleus sampling, while continuously feeding generated text back as the other agent's input. Under this protocol, repeated text enters both histories, creating conditions conducive to further use of similar expressions. Existing studies of repetition degeneration provide context for this self-reinforcing interpretation [1–3], but correlated trajectories alone do not establish a unique mechanism. The branch experiment tests whether output behavior is mutable by holding the starting history fixed while changing sampling or history handling. Its evidence is limited to the selected repetitive prefixes and the short continuation window.

The recommended sampling control is particularly informative. It changes temperature, top_p, and the presence penalty jointly, so any observed local change can be attributed only to the configuration as a whole, not automatically to an individual parameter. Shortening history also reduces both input length and exposure to older text; changes in elapsed time therefore include differences in context-processing cost. The anti-repetition instruction changes the requested task and is not fully comparable to an

intervention that changes only decoding probabilities. These differences are both practical engineering options and design information that must be retained when interpreting causal pathways.

5.3 Comparing simple methods with complex algorithms

String matching and simple linear models provide the clearest evidence in this study. Strict periodicity detection yields directly verifiable counts, logistic regression outperforms the tested tree-based models on the early classification task, and more complex clustering does not produce a reliable unified categorization. Together, these findings show that greater algorithmic complexity does not necessarily increase understanding of interaction states. Evaluation should examine whether an algorithm changes the conclusions, maintains performance across separate validation splits, and produces outputs aligned with a clear research question.

Negative findings are also informative about the study's boundaries. The absence of stable improvement from topic anchoring does not support that sentence as a reliable solution in this setting. Failure of timing prediction to outperform the baseline in the chronological holdout limits deployment claims for early prediction. Disagreement across clustering algorithms indicates that the present data do not support a fixed number of semantic attractor states. These results are retained in the main text or supplementary material to avoid presenting only the best-performing models and conditions.

5.4 Resource control and practical implications

In local batch experiments, repetition detection can support both quality monitoring and management of computing resources. The retrospective evaluation shows that the more conservative rule requiring 20 consecutive near-repetitions identifies many low-variation suffixes that persist to the observation endpoint. This provides a basis for auditable stopping criteria. A practical system could instead trigger a branching intervention, but its subsequent trajectory would then require reassessment; recovery probabilities from unintervened data cannot directly estimate risk in an intervention-enabled system.

Low-variation detection should be interpreted in relation to the application goal. Tolerance for repetition differs between casual conversation, discussions intended to converge on consensus, and research agents expected to continually discover new evidence. The study does not treat all repetition as erroneous or propose a universal threshold for all tasks. Its engineering value lies in specifying what is detected, recording trigger times, quantifying avoidable costs and truncated changes, and making these quantities reproducible.

5.5 Scope and limitations

First, the main dataset covers one 0.8B model, Chinese text, and five fixed fictional openings. The two agents share weights, and A always speaks first. B's first input also contains A's first utterance, so linguistic differences between the agents confound order with input structure and cannot be interpreted as intrinsic personalities or unidirectional causal influence. Second, the original template protocol appends fragments while retaining tokens; cache replay checks do not exclude an effect of this protocol on degeneration. Third, the original sample was stopped at 621 complete conversations, and measures and analyses were not preregistered. Significance tests and model comparisons must therefore be interpreted as exploratory evidence.

Fourth, semantic cosine similarity, keywords, and novel character fragments do not constitute comprehensive quality annotations. Fifth, supervised validation isolates replicate blocks, but the topic texts remain the same five openings across folds; the chronological holdout also comes from the same generation task. Sixth, branch controls begin from a small number of already locked prefixes. Repeating random seeds does not replace additional independent prefixes, and 12 continuation messages assess only short-term change. Finally, observation ends at message 200. Zero observed recovery does not

prove infinite-horizon absorption, and retrospective stopping-rule results do not guarantee the same savings in future runs. These limits do not invalidate period counts for the recorded sample, but they restrict generalization to broader model populations, open-ended topics, and long-term intervention effects.

6 Conclusions

With fixed Qwen3.5-0.8B weights, five fictional Chinese openings, and the generation protocol used here, 98.23% of 621 long two-agent conversations exhibited strict short-period repetition in their final 40 messages. Highly repetitive tails retained topic features that allowed the opening to be identified, indicating that topic retention and interaction change should be measured separately. A single instruction to remain on topic showed no stable improvement, early features provided only limited warning signals, and complex semantic clustering did not support a unified categorization. Across 100 branches from frozen prefixes, the recommended sampling configuration produced only limited, local short-term change, without a stable effect under adjusted inference; some changes also coexisted with repetition between speakers. Periodicity detection, branch verification, and retrospective stopping-rule analysis together provide reproducible methods for auditing interaction states and controlling costs. Reliable evaluation of long agent conversations requires joint attention to periodic structure, topical and factual information, and actual progress under explicitly defined tasks.

Author Contributions and Declarations

Jiexiao Zhang is the sole author and is responsible for conceptualization and study design, software implementation and experimental execution, data curation, formal analysis, interpretation, visualization, and manuscript writing, and takes responsibility for the research content. All model-inference experiments and quantitative analyses were conducted on the author's local equipment. The study uses publicly released model weights and open-source software. Its dialogue data were generated through local inference and were not obtained from external business or personal-conversation datasets.

Funding and support: This study received no external funding, corporate sponsorship, or in-kind support. Competing interests: The author declares no competing interests requiring disclosure. Publication status: At the time of manuscript preparation, this study had not been formally published, posted as a preprint, or placed under review at another journal.

Data and code availability: Message-level experimental records, configurations, sampling seeds, runtime information, file digests, statistical code, and figure data have been preserved. The accompanying supplementary material provides the main result tables, methodological details, and reproduction instructions. Complete reproduction files are maintained by the corresponding author and can be requested at jiexiaozhang1@outlook.com. Public pretrained models are used under their original licenses.

Use of generative artificial intelligence tools: Codex assisted with code implementation, statistical checks, figure preparation, organization of the manuscript, and English translation and language editing. Qwen3.5-0.8B is the local generative model investigated in this study, and BGE is the semantic encoder used for analysis. These tools are not listed as authors; the sole author remains responsible for the research content.

Ethics statement: This study uses only fictional openings and locally generated model text. It involves no recruitment of human participants, personally identifiable data, or animal experiments.

References

[1] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, et al. The Curious Case of Neural Text Degeneration. International Conference on Learning Representations (ICLR), 2020. https://arxiv.org/abs/1904.09751

[2] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, et al. Neural Text Generation with Unlikelihood Training. International Conference on Learning Representations (ICLR), 2020. https://arxiv.org/abs/1908.04319

[3] Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, et al. Learning to Break the Loop: Analyzing and Mitigating Repetitions for Neural Text Generation. Advances in Neural Information Processing Systems, 2022, 35. DOI: 10.52202/068431-0223.
https://proceedings.neurips.cc/paper_files/paper/2022/hash/148c0aeea1c5da82f4fa86a09d4190da-Abstract-Conference.html

[4] Ting-Wen Ko, Jonas Geiping. Attractor States Emerge in Multi-Turn LLM Conversations. arXiv preprint, 2026. DOI: 10.48550/arXiv.2606.30571. https://arxiv.org/abs/2606.30571

[5] Tatsuya Hiraoka, Kentaro Inui. Repetition Neurons: How Do Language Models Produce Repetitions? Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), 2025: 483–495. DOI: 10.18653/v1/2025.naacl-short.41.
https://aclanthology.org/2025.naacl-short.41/

[6] Qwen Team. Qwen3.5-0.8B model card. Hugging Face model repository, 2026.
https://huggingface.co/Qwen/Qwen3.5-0.8B/blob/2fc06364715b967f1860aea9cf38778875588b17/README.md (Accessed: 2026-10-08)

[7] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, et al. C-Pack: Packed Resources For General Chinese Embeddings. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024: 641–649. https://arxiv.org/abs/2309.07597

[8] Yoav Benjamini, Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological), 1995, 57(1): 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x. https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1995.tb02031.x

[9] Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 1979, 7(1): 1–26. DOI: 10.1214/aos/1176344552. https://projecteuclid.org/journals/annals-of-statistics/volume-7/issue-1/Bootstrap-Methods-Another-Look-at-the-Jackknife/10.1214/aos/1176344552.full

[10] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 2017, 40(8): 913–929. DOI: 10.1111/ecog.02881.
https://nsojournals.onlinelibrary.wiley.com/doi/10.1111/ecog.02881

[11] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987, 20: 53–65. DOI: 10.1016/0377-0427(87)90125-7.
https://doi.org/10.1016/0377-0427(87)90125-7

[12] Lawrence Hubert, Phipps Arabie. Comparing partitions. Journal of Classification, 1985, 2: 193–218. DOI: 10.1007/BF01908075. https://doi.org/10.1007/BF01908075

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2011, 12(85): 2825–2830.
https://www.jmlr.org/papers/v12/pedregosa11a.html

[14] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, Jennifer Neville. LLMs Get Lost In Multi-Turn Conversation. International Conference on Learning Representations (ICLR), 2026. https://arxiv.org/abs/2505.06120

[15] UK Parliament. Magna Carta. UK Parliament, n.d. https://www.parliament.uk/magnacarta/ (Accessed: 2026-10-08)